12-1-1993

Second Order Approximation in the Partially Linear Regression Model

Oliver B. Linton

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1065

SECOND ORDER APPROXIMATION IN THE PARTIALLY LINEAR REGRESSION MODEL

Oliver Linton

December 1993

SECOND ORDER APPROXIMATION IN THE PARTIALLY LINEAR REGRESSION MODEL

BY OLIVER LINTON

We examine the second order properties of various quantities of interest in the partially linear regression model. We obtain a stochastic expansion with remainder $o_P(n^{-2\mu})$, where $\mu < 1/2$, for the standardised semiparametric least squares estimator, a standard error estimator, and a studentised statistic. We use the second order expansions to correct the standard error estimates for second order effects, and to define a method of bandwidth choice. A monte carlo experiment provides favourable evidence on our method of bandwidth choice.

KEYWORDS: Semiparametric estimation, partially linear regression, kernel, local polynomial, second order approximations, bandwidth choice, asymptotic expansions.

1. INTRODUCTION

The subject of this paper is the partially linear regression model, considered in Engle, Granger, Rice and Weiss (1986),

$$y_1 = \beta^T Y_2 + \theta(Z) + \epsilon \tag{1}$$

where $\theta(\bullet)$ is an unknown scalar function and ϵ is a zero mean error orthogonal to both Y_2 and $\theta(\bullet)$. This model embodies a compromise between employing a general nonparametric specification $g(Y_2, Z)$, which, if the conditioning variables are high dimensional, would lead to serious loss of precision, and a fully parametric specification which may result in badly biased estimators and inconsistent hypothesis tests. The implicit asymmetry between the effects of Y_2 and Z may be attractive when Y_2 consists of dummy or categorical variables, as in Stock (1989, 1991). This specification arises in various sample selection models, see Ahn and Powell (1993), Newey, Powell, and Walker (1990), and Lee, Rosenzweig and Pitt (1992). It is also the basis of a general specification test for functional form introduced in Delgado and Stengos (1994).

We focus on inference about β. Taking expectations given Z of both sides of (**??**), and subtracting this conditional moment from both sides we obtain the estimating equation

$$y_1 - E[y_1|Z] = \beta^T(Y_2 - E[Y_2|Z]) + \epsilon. \tag{2}$$

Robinson (1988b) used (higher order) Nadaraya-Watson kernel estimates of the regression functions $E[y_1|Z]$ and $E[Y_2|Z]$ to construct a feasible least squares estimate $\widehat{\beta}$ of β. Under regularity conditions restricting the dimensions of Z and imposing some smoothness on the above regression functions, he proved that $\sqrt{n}(\widehat{\beta}-\beta)$ was asymptotically normal. When the errors ϵ are normal, this estimator achieves the semiparametric information bound. See also H.Chen (1988), N.Heckman (1986), and Speckman (1988) for alternative treatments.

There is some evidence that first order asymptotic distributions provide poor approximations in practice to the sampling behavior of semiparametric estimators, at least for the type of sample sizes available with economic data – see the monte carlo evidence presented in Stock (1989) and Stoker (1993) who find less favourable conclusions than Hsieh and Manski (1987). Semiparametric estimators typically require the selection of a smoothing parameter $h(n)$, frequently called the bandwidth, which determines the effective degree of parameterisation taken by the nuisance function for given sample size n, where the number of nuisance parameters must increase with sample size at a certain rate. Intuition gained in analysing the small sample behavior of parametric estimators – see Rothenberg (1984b) – suggests that small sample properties deteriorate with the number of nuisance parameters to be estimated; one therefore expects semiparametric procedures may incur a substantial small sample cost reflecting the ambitious goal of estimating an asymptotically infinite number of nuisance parameters.

A perhaps more serious problem is that the first order approximation does not reflect the choice of $h(n)$. Frequently, the numerical value of point estimates obtained can vary quite considerably with bandwidth. Therefore, one must view reported results with some scepticism. Furthermore, the first order theory does not give any information about how to choose $h(n)$ in practice, and therefore provides an incomplete description of the procedure.

We believe that higher order theory, as expounded in Pfanzagl (1980) and Rothenberg (1984a), can address some of the problems presented by the first order theory.

These methods have a long history of application to econometric problems, Phillips and Park (1988) and Chesher and Spady (1991) being some recent examples, and provide both qualitative and quantitative information about the properties of estimators and test statistics beyond that contained in the first order theory.

We derive a stochastic expansion with remainder $o_P(n^{-2\mu})$, where $0 < \mu < 1/2$, for the standardised semiparametric estimator of β, a standard error estimator, and a studentised statistic. We derive approximations to the moments of the truncated expansions which are then used to compare various alternative implementations, to define second order optimality, and to define an automatic method of bandwidth choice. Our results are qualitatively similar to those developed in Carroll and Härdle (1989) and Härdle, Hart, Marron and Tsybakov (1992) for related semiparametric problems. However, we extend the analysis in several directions. Firstly, we prove that our truncated expansions are equal in distribution to the original statistic to order $n^{-2\mu}$, regularity that is frequently established in the Edgeworth approximation literature − see Rothenberg (1984a) and Robinson (1988a) − and which permits distributional approximation of the Edgeworth type, see Linton (1992) for some results in this direction. We also demonstrate that our bandwidth selection method not only results in a $\sqrt{n}$ consistent estimator of β, but one that is second order optimal.

Section 2 discusses the estimator and test statistics, while section 3 gives the assumptions. Section 4 justifies the expansions, while section 5 gives the second order moment approximations for the quantities of interest. Section 6 uses the second order approximations to define a feasible correction to the standard errors, while in section 7 we develop a second order optimality theory. Section 8 reports the results of a small simulation experiment. Section 9 gives our main conclusions. The Appendix contains the proofs of all theorems.

A word on notation. We use $\Rightarrow$ to denote convergence in distribution, $\stackrel{P}{\rightarrow}$ means convergence in probability, while the symbol $\approx$ denotes asymptotic equivalence in probability, all holding as $n \to \infty$.

2. ESTIMATION

Let the observed data $\{(Y_i^T, Z_i^T)^T\}_{i=1}^n$, where $Y_i = (y_{1i}, y_{2i}, .., y_{D+1i})^T \equiv (y_{1i}, Y_{2i}^T)^T$, be generated by

$$y_{1i} = \beta^T Y_{2i} + \theta(Z_i) + \epsilon_i \ ; \ y_{di} = g_d(Z_i) + \eta_{di}, \ d = 2, 3, .., D+1, \tag{3}$$

where β is a D by 1 vector of unknown parameters and Z_i is a P by 1 vector of observable regressors, while ϵ_i and η_{di} are mean zero conditional on $(Z_i^T, Y_{2i}^T)^T$ and Z_i respectively. Subtracting the conditional (given Z) means from both sides of (**??**), we obtain the regression equation (**??**). Let $\nu_i = y_{1i} - g_1(Z_i)$, with $g_1(Z_i) = E[y_{1i}|Z_i]$, and $\eta_i = (\eta_{2i}, \eta_{3i}, .., \eta_{D+1i})^T$. We examine the properties of the feasible least squares estimator

$$\widehat{\beta} = [n^{-1}\sum_{i=1}^n \widehat{\eta}_i \widehat{\eta}_i^T]^{-1}[n^{-1}\sum_{i=1}^n \widehat{\eta}_i \widehat{\nu}_i] \equiv S_{\widehat{\eta}\widehat{\eta}}^{-1} S_{\widehat{\eta}\widehat{\nu}}, \tag{4}$$

of β, where the residuals $\widehat{\nu}_i = y_{1i} - \widehat{g}_1(Z_i)$ and $\widehat{\eta}_{di} = y_{di} - \widehat{g}_d(Z_i)$, $d = 2, .., D+1$, are derived from nonparametric estimates $\widehat{g}_d$ of g_d. Härdle and Linton (1994) reviews a number of suitable nonparametric regression estimators. We estimate the regression functions by the local polynomial regression method suggested in Stone (1977) and further examined in Fan (1992). It is motivated by the following argument.

Let $\tau_i = (\tau_i^0, \tau_i^{1T}, .., \tau_i^{q-1T})^T$ contain all partial derivatives of the generic regression function g at the point $Z_i = (Z_{i1}, .., Z_{iP})^T$ up to order $q-1$, with, in particular, $\tau_i^0 = g(Z_i)$ and $\tau_i^1 = (\partial g/\partial Z_{i1}, .., \partial g/\partial Z_{iP})^T$. In total there are $m_r = \sum_{j=0}^r \frac{P!}{j!(P-j)!}$ distinct

$r'th$ order partial derivatives so that τ_i is an $m = \sum_{r=0}^{q-1} m_r$ by 1 vector. The function g can be expanded in a q'th order Taylor series so that for Z_j in a neighbourhood $\mathcal{N}(Z_i)$ of Z_i, $g(Z_j) \approx \tau_i^T x_{ij}$, where $x_{ij} = (x_{ij}^0, x_{ij}^{1T}, .., x_{ij}^{q-1T})^T$ contains the corresponding Taylor coefficients with, in particular, $x_{ij}^0 = 1$ and $x_{ij}^1 = (Z_{j1} - Z_{i1}, .., Z_{jP} - Z_{iP})^T$. This suggests estimating τ_i by a least squares regression of y_j on x_{ij}, for $Z_j \in \mathcal{N}(Z_i)$. In fact, we estimate τ_i $-$ down-weighting observations according to their distance from Z_i, and, for convenience, leaving out the i'th contribution (see Robinson (1987) for a discussion of this modification) $-$ by

$$\widehat{\tau}_i = (X_i^T K_i X_i)^{-1} X_i^T K_i y, \tag{5}$$

with $y = (y_1, y_2, .., y_n)^T$ the generic dependent variable and $X_i = (x_{i1}, ., 0, ., x_{in})^T$ the n by m data matrix with x_{ii} replaced by a zero vector, while K_i is the n by n diagonal weighting matrix with i'th element zero and j'th component $k((Z_j - Z_i)/h)$. Here, $k(\bullet)$ is a $P-$dimensional probability density function and $h(n)$ is a scalar bandwidth. Clearly, $\widehat{\tau}_i$ is linear in $\{y_j\}_{j=1}^n$, and in particular

$$\widehat{g}(Z_i) = \sum_{j \neq i} w_{ij} y_j, \;\; i = 1, 2, .., n, \tag{6}$$

where $\{w_{ij}\}_{j=1}^n$ is a sequence of non-stochastic weights, obtained from (**??**). This class of estimators is easy to compute and includes the Nadaraya-Watson estimator $-$ used by Robinson (1988b) $-$ as a special case (when $q = 1$). The role of q is similar to that of kernel order in the higher order kernel method of Bartlett (1963): the pointwise bias of $\widehat{g}(Z_i)$ is of order h^q $-$ see Lemma 1 below. Furthermore, the interior bias does not depend on the design density (it is design adaptive), while there is an automatic correction for boundary bias, see Fan and Gijbels (1992) for a discussion.

3. ASSUMPTIONS

A1. *The fixed regressors* $\{Z_i\}_{i=1}^n$ *have as their support a bounded domain* $\Upsilon \subseteq \mathbb{R}^P$. *Furthermore, there exists a positive density function* f *such that*

$$n^{-1}\sum_{i=1}^{n} v(Z_i) \to \int v(Z)f(Z)dZ,$$

for any bounded continuous function $v(\bullet)$.

The fixed bounded design assumption is made for technical convenience; it allows us to avoid the use of trimming factors (such as in Robinson (1988b)) that are necessary when the denominator of w_{ij} gets arbitrarily small. Our theorems are also true for triangular array designs such as $Z_{ni} = i/n$, in which case $f(Z) = [\int_\Upsilon dZ]^{-1}\mathrm{I}(Z \in \Upsilon)$, where $\mathrm{I}(\bullet)$ is the usual indicator function, but we have not explicitly included this possibility to avoid notational inconvenience. The fixed design assumption is frequently employed in the nonparametric regression literature – see for example Müller (1988) and Härdle (1990). Assumption A1 does not preclude $\{Z_i\}_{i=1}^n$ from having been generated by some random mechanism. For example, suppose that Z_i were iid with density f, then

$$\Pr[n^{-1}\textstyle\sum_{i=1}^{n} v(Z_i) \to \int v(Z)f(Z)dZ] = 1,$$

see Andrews (1991a). In this case, we can interpret our analysis as being conditional on $\{Z_i\}_{i=1}^n$, and our results (Theorems 1-4 below) holding with probability one.

A2. *The sequence of vectors* $u_i = (\epsilon_i, \eta_{2i}, .., \eta_{D+1i})^T$ *are mean zero and independent, while* ϵ_i *is independent of* η_i *for all* i. *Let* $E[u_i u_i^T] = \Sigma(Z_i)$, *where* $\Sigma = \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \Sigma_{\eta\eta} \end{pmatrix}$.

Then $\Sigma(Z)$ is bounded away from zero and infinity on Υ. Finally, assume that for some integers J and $L \geq 5$, $Sup_i E[|\epsilon_i|^J] < \infty$ and $Sup_i E[|\eta_i|^L] < \infty$.

Speckman (1988) conditions on both (Y_2, Z) and considers only scalar design. Robinson (1988b) employs a more general $P-$dimensional random design, where (Y, Z) are iid, both with unbounded support. Our assumption A2 permits conditional heterogeneity with respect to Z which was not allowed for in this paper. However, we shall sometimes impose the following assumption that disallows heteroskedasticity.

A2'. *The sequence of vectors $u_i = (\epsilon_i, \eta_{2i}, .., \eta_{D+1i})^T$ are independent and identically distributed, with mean zero and covariance matrix Σ, while ϵ_i is independent of η_i. Furthermore, for some J and $L > 4$: $E[|\epsilon_i|^J] < \infty$ and $E[|\eta_i|^L] < \infty$.*

Let $\mathcal{G}_t$ be the class of all t-times differentiable functions g defined on Υ whose t'th derivatives satisfy the following Lipschitz condition: there exists $\vartheta > 0$ and $\Gamma < \infty$, such that for all $x, z \in \Upsilon$ and for all $s = (s_1, .., s_P)^T$ with $s_1 + .. + s_P \leq t$, we have

$$|g^{(s)}(z) - g^{(s)}(x)| < \Gamma |z - x|^{\vartheta},$$

where $g^{(s)}(z) = \frac{\partial^{s_1 + .. + s_P}}{\partial^{s_1} z_1 \partial^{s_2} z_2 .. \partial^{s_P} z_P} g(z)$.

A3. *The regression functions $G(Z_i) = (g_1(Z_i), g_2(Z_i), .., g_{D+1}(Z_i))^T \equiv (g_1(Z_i), G_2(Z_i)^T)^T$, are all members of $\mathcal{G}_q$, where $q \geq 2$. Furthermore, f, $\Sigma_{jl} \in \mathcal{G}_1$, for j,l =1,2,.,D+1.*

Finally, we make assumptions about the kernel and the bandwidth sequence.

A4. $nh^{4q} \to 0$ *and* $nh^{2P} \to \infty$.

Our bandwidth restrictions are similar to those in Robinson (1988b), except that no trimming constant is necessary due to our assumption that Υ is bounded.

A5. *The kernel* k *has bounded support and has one continuous partial derivatives in each direction. Furthermore, the matrices*

$$\Omega_{ni} = n^{-1}h^{-P}H^{-1}X_i^T K_i X_i H^{-1}$$

are positive definite with smallest eigenvalues bounded away from zero, where $H = diag\{H_0, H_1, .., H_{q-1}\}$ *is an* m *by* m *diagonal scaling matrix with* H_j *a scalar* m_j *by* m_j *matrix with diagonal element* h^j, $j = 0, 1, .., q-1$.

As $n \to \infty$, $\Omega_{ni} \to f(Z_i)\Omega$, where Ω is a matrix whose elements are the moments of the kernel k. Lindsay (1989) showed, in the special case $P = 1$, that Ω is positive definite if $k(t)$ takes at least $q - 1$ distinct values, which would be satisfied by any continuous kernel.

The properties of the local polynomial estimator, at both interior and boundary points, are described in the following:

LEMMA 1 *(Fan (1992, Theorem 1) and Ruppert and Wand (1992, Theorem 4.2). Let* $\sigma^2(Z_i) = Var(y_i|Z_i)$ *for scalar* y_i*, and suppose A1-A5 are satisfied. Then, as* $n \to \infty$,

$$E[\widehat{\tau}_i] - \tau_i \approx h^q \mathcal{D}_k^q \tau_i \ ; Var[\widehat{\tau}_i] = \sigma^2(Z_i) n^{-1} h^{-P} H^{-1} \Omega_{ni}^{-1} \Psi_{ni} \Omega_{ni}^{-1} H^{-1},$$

where $\Psi_{ni} = n^{-1}h^{-P}H^{-1}X_iK_i^2X_iH^{-1}$ *and* Ω_{ni} *are* $O(1)$*, while*

$$\mathcal{D}_k^q \tau_i = H^{-1}\Omega_{ni}^{-1} \textstyle\sum_{s_1+..+s_P=q} g^{(s)}(Z_i)\omega_{ni}^{(s)}$$

$$\omega_{ni}^{(s)} = n^{-1}h^{-P}H^{-1}X_i^TK_i[\textstyle\prod_{p=1}^P(\frac{Z_{1p}-Z_{ip}}{h})^{s_p}, .., \prod_{p=1}^P(\frac{Z_{np}-Z_{ip}}{h})^{s_p}]^T = O(1).$$

4. ASYMPTOTIC EXPANSIONS

Robinson (1988b) gives very general conditions under which $\widehat{\beta}$ is (first order) asymptotically equivalent to the infeasible least squares estimator $\overline{\beta}$ that satisfies $\sqrt{n}c^T(\overline{\beta} - \beta) \Rightarrow N(0, \sigma^2)$ for any D by 1 vector c, where $\sigma^2 = c^T\overline{\Sigma}_{\eta\eta}^{-1}\overline{\Sigma}_{\eta\eta\epsilon\epsilon}\overline{\Sigma}_{\eta\eta}^{-1}c$ is the usual variance of the least squares estimator under heteroskedasticity, defined through $n^{-1}\sum_{i=1}^n \eta_i\eta_i^T \xrightarrow{P} \overline{\Sigma}_{\eta\eta}$ and $n^{-1}\sum_{i=1}^n \eta_i\eta_i^T\epsilon_i^2 \xrightarrow{P} \overline{\Sigma}_{\eta\eta\epsilon\epsilon}$. We investigate the higher order properties of $\widehat{\beta}$ defined in (**??**) and (**??**) by asymptotic expansions. We also examine the properties of a standard error estimator and of a studentised statistic in the special case of homoskedastic errors. In this case, the asymptotic variance of $\sqrt{n}c^T(\widehat{\beta}-\beta)$ is $\sigma^2 = \sigma_\epsilon^2 c^T \Sigma_{\eta\eta}^{-1} c$, where $\sigma_\epsilon^2 = E[\epsilon_i^2]$ and $\Sigma_{\eta\eta} = E[\eta_i \eta_i^T]$. We estimate $\Sigma_{\eta\eta}$ by $S_{\widehat{\eta}\widehat{\eta}}$ and σ_ϵ^2 by $\widehat{\sigma}_\epsilon^2 = S_{\widetilde{\varepsilon}\widetilde{\varepsilon}} = n^{-1}\sum_{i=1}^n \widetilde{\varepsilon}_i^2$, where $\widetilde{\varepsilon}_i = \widehat{\nu}_i - \widehat{\beta}^T \widehat{\eta}_i$. For convenience, we report results for the scalar standardised quantities $T = \sqrt{n}c^T(\widehat{\beta}-\beta)/\sigma$, $S = \sqrt{n}(\widehat{\sigma}-\sigma)/\sigma$, and $W = \sqrt{n}c^T(\widehat{\beta}-\beta)/\widehat{\sigma}$. An important special case here is where $c = (0,..,0,1,0,..,0)^T$ in which case W can be used to test for the individual significance of the corresponding parameter estimate.

By a Taylor expansion we have

$$T = \sigma^{-1}c^T\overline{\Sigma}_{\eta\eta}^{-1}X_N - \frac{c^T\overline{\Sigma}_{\eta\eta}^{-1}X_D\overline{\Sigma}_{\eta\eta}^{-1}X_N}{\sigma\sqrt{n}} + \frac{c^T\overline{\Sigma}_{\eta\eta}^{-1}X_D\overline{\Sigma}_{\eta\eta}^{-1}X_D\overline{\Sigma}_{\eta\eta}^{-1}X_N}{\sigma n} + \frac{R}{n\sqrt{n}} \equiv T^* + R_T, \tag{7}$$

where $X_N = n^{-1/2}\sum_{i=1}^n \widehat{\eta}_i\widehat{\epsilon}_i$, with $\widehat{\varepsilon}_i = \widehat{\nu}_i - \beta^T\widehat{\eta}_i$, and $X_D = \sqrt{n}(S_{\widehat{\eta}\widehat{\eta}} - \overline{\Sigma}_{\eta\eta})$ are second order weighted U-statistics – see Lee (1990) – whose properties are further discussed in the appendix, while

$$R = -\sigma^{-1}c^T S_{\widehat{\eta}\widehat{\eta}}^{-1} X_D\overline{\Sigma}_{\eta\eta}^{-1}X_D\overline{\Sigma}_{\eta\eta}^{-1}X_D\overline{\Sigma}_{\eta\eta}^{-1}X_N. \tag{8}$$

The moments of the truncated statistic T^* depend on the bandwidth h: in particular, $E[T^*] = O(\sqrt{n}h^{2q})$, while $Var[T^*] = 1 + O(n^{-1}h^{-P})$. Thus, the asymptotic mean squared error of $\widehat{\beta}$,

$$MSE(h) \approx n^{-1}\{1 + O(n^{-1}h^{-P}) + O(nh^{4q})\},$$

is minimized by choosing a bandwidth $h(n) = O(n^{-\pi})$, where $\pi = 2/(4q+P)$, in which case a correction to the second moment of T^* of order $n^{-2\mu}$ results, where $\mu = (4q-P)/2(4q+P)$. This bandwidth balances the order of magnitude of variance and squared bias corrections so that $O(nh^{4q}) = O(n^{-1}h^{-P}) = O(n^{-2\mu})$ (when $P=1$ and $q=2$, $\pi = 2/9$ and $\mu = 7/18$). Furthermore, $R_T = o_P(n^{-2\mu})$.

In the next section we calculate approximations to the moments of T^*, which we shall interpret as approximations to the 'moments' of T. This methodology has a long tradition of application in econometric problems following Nagar (1959) – see the references in Rothenberg (1984a). When $Sup_n E[T^2] < \infty$, we expect that $E[T^2] = E[T^{*2}] + o(n^{-2\mu})$, provided $R_T = o_P(n^{-2\mu})$, but see Srinavasan (1970) for a cautionary tale in this regard. These conditions are satisfied in the problems examined in Carroll and Härdle (1989) and Härdle, Hart, Marron and Tsybakov (1992), but not here – T, S, and W do not necessarily have uniformly bounded moments. In this case, some additional justification for examining the moments of the truncated statistic must be given. We establish the stronger regularity that T and T^* have the same distribution to order $n^{-2\mu}$ (similar versions of this result can also be obtained for S and W). Therefore, our moment approximations can be interpreted as the moments of the approximating distribution.

4.1 Second Order Approximation for T

In the appendix we obtain

THEOREM 1. *Assume that A1-A5 hold, with J and $L > M_1 \equiv Max\{12, 32\mu/(3-4\mu)\}$, where $\mu = (4q-P)/2(4q+P)$, and let $h(n) = O(n^{-\pi})$, where $\pi = 2/(4q+P)$. Then,* (1.1) T *and* T^* *have the same distribution to order* $n^{-2\mu}$, (1.2) $Sup_n E[T^{*2}] < \infty$, *and*

$$(1.3)\ E[T^*] = \sqrt{n}h^{2q}\mathcal{B} + o(n^{-\mu})\ ;\ Var[T^*] = 1 + n^{-1}h^{-P}\mathcal{V} + o(n^{-2\mu}),$$

where the $O(1)$ quantities $\mathcal{B}$ and $\mathcal{V}$ are

$$\mathcal{B} = \sigma^{-1}c^T\overline{\Sigma}_{\eta\eta}^{-1}\zeta_q^{\eta\theta}\ ;\ \mathcal{V} = \sigma^{-2}c^T\overline{\Sigma}_{\eta\eta}^{-1}\{\sum_{i=1}^{n}\sigma_\epsilon^2(Z_i)\Sigma_{\eta\eta}(Z_i)\sum_{j\neq i}nh^P\rho_{ij}^2\}\overline{\Sigma}_{\eta\eta}^{-1}c, \tag{9}$$

with $\rho_{ij} = n^{-1/2}[\sum_{k\neq j,k\neq i} w_{ki}w_{kj} - 2w_{ij}]$ *and* $\zeta_q^{\eta\theta} = n^{-1}\sum_{i=1}^{n}\widetilde{B}_i(\theta)\widetilde{B}_i(G_2)$, *where for any function* $g \in \mathcal{G}_q$,

$$\widetilde{B}_i(g) \equiv h^{-q}[\sum_{j\neq j} w_{ij}g(Z_j) - g(Z_i)] = O(1).$$

REMARK: When $\mu < 36/80$, $M_1 = 12$, otherwise higher moments are required.

REMARK: The nonparametric residuals $\widehat{\eta}_i$ and $\widehat{\nu}_i$ are not zero mean in general. This suggests that one may wish to use recentred residuals in the final stage regression (**??**). Similar second order moment approximations hold for the intercept OLS estimator. In this case, the variance is the same as in Theorem 1, but the bias involves the 'demeaned' curvature measures – i.e. we use the formulae of Theorem 1 but replace $\zeta_q^{\eta\theta}$ by $\zeta_q^{\eta\theta} - \zeta_q^{\theta}\zeta_q^{\eta}$, where $\zeta_q^{\eta} = n^{-1}\sum_{i=1}^{n}\widetilde{B}_i(G_2)$ and $\zeta_q^{\theta} = n^{-1}\sum_{i=1}^{n}\widetilde{B}_i(\theta)$. In many cases, including an intercept will result in an improvement of the bias, see (**??**) below.

By Lemma 1, we can further approximate $\widetilde{B}_i(g)$ by $\mathcal{D}_k^q g(Z_i)$, (the first element of $\mathcal{D}_k^q \tau_i$), which is a linear combination of the $q'th$ order partial derivatives of g weighted by functionals of the kernel. Consider the special case where u_i are homoskedastic and the local linear procedure is used in estimating G. In this case, $\mathcal{D}_k^q g(Z_i) = \alpha_1^2(k)g''(Z_i)/4$, and, when $c = 1$ and $\Upsilon = [0, 1]$,

$$E[T^*] \approx \sqrt{n}h^4\alpha_1^2(k)\frac{\int \theta''(Z)g_2''(Z)f(Z)dZ}{4\sigma_\epsilon\sigma_\eta}\ ;\ Var[T^*] \approx 1 + n^{-1}h^{-1}\alpha_2(k_*), \quad (10)$$

where $k_*(t) = k * k(t) - 2k(t)$ is the 'twiced' kernel in which $k * k(t) = \int k(t-s)k(s)ds$ is the convolution of k with itself, while the kernel constants $\alpha_1(k) = \int t^2k(t)dt$

and $\alpha_2(k) = \int k^2(t)dt$ are both positive. These constants can be calculated quite readily for most frequently used kernels k, see Linton (1991): for the quadratic kernel $k(t) = \frac{3}{4}(1-t^2)\mathrm{I}(|t|<1)$, $\alpha_1 = 0.2$ and $\alpha_2 = 0.835$. Now suppose that the signal to noise ratios $\theta''(Z)/\sigma_\epsilon$ and $g_2''(Z)/\sigma_\eta$ are constant and equal to δ_1 and δ_2 respectively, which is consistent with regression functions of the form $g(Z) = \delta_i Z^2/2$ and Σ being scalar. We computed the MSE at a range of bandwidths for sample size $n = 100$ using both (**??**), with $\widetilde{B}_i(g;h)$ and $\mathcal{V}(h) = \sum_{i=1}^n \sum_{j\neq i} nh\rho_{ij}^2$ which we call "exact", and the asymptotic approximation (**??**). Figures 1-4 display the MSE plotted against $\ln(h)$ – see Marron and Wand (1992) for a discussion of why the log-scale is more appropriate – for four different configurations of δ_1 and δ_2. There is a difference between the two approximations for large bandwidths; however, both indicate modest small sample effects for bandwidths near the optimum, although if bandwidth is chosen poorly the cost can be excessive.

$$*** \text{ Figs } 1-4 \text{ here } ***$$

A value of $\delta_1 = 20$ seems quite extreme to us, and would certainly be very pronounced in the data. We therefore feel that at least for scalar Z, the small sample properties of $\widehat{\beta}$ should be acceptable unless an extreme bandwidth is chosen.

4.2. Second Order Approximation for S and W

We now consider the properties of the standard error estimate and test statistic in the special case where the errors are homoskedastic. Standard error estimates can be very poorly behaved in finite samples even in situations where local smoothing is not employed, as Chesher and Jewitt (1987) have quantified. We proceed as for T to make a stochastic expansion for S; but in this case, since we only establish order in probability properties of the remainder, we only collect terms that are larger than $O_P(n^{-1})$. This results in

$$S = \frac{\sqrt{n}(\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2)}{2\sigma_\epsilon^2} + \frac{c^T\sqrt{n}(S_{\hat\eta\hat\eta}^{-1} - \Sigma_{\eta\eta}^{-1})c}{2c^T\Sigma_{\eta\eta}^{-1}c} - \frac{[\sqrt{n}(\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2)]^2}{8\sigma_\epsilon^4\sqrt{n}} - \frac{[c^T\sqrt{n}(S_{\hat\eta\hat\eta}^{-1} - \Sigma_{\eta\eta}^{-1})c]^2}{8[c^T\Sigma_{\eta\eta}^{-1}c]^2\sqrt{n}} + \frac{[c^T\sqrt{n}(S_{\hat\eta\hat\eta}^{-1} - \Sigma_{\eta\eta}^{-1})c][\sqrt{n}(\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2)]}{4\sigma^2\sqrt{n}} + O_P(n^{-1}), \quad (11)$$

Now let S^* be the truncated statistic obtained after using the further approximations:

$$\sqrt{n}(S_{\hat\eta\hat\eta}^{-1} - \Sigma_{\eta\eta}^{-1}) = -\Sigma_{\eta\eta}^{-1}X_D\Sigma_{\eta\eta}^{-1} + \frac{\Sigma_{\eta\eta}^{-1}X_D\Sigma_{\eta\eta}^{-1}X_D\Sigma_{\eta\eta}^{-1}}{\sqrt{n}} + O_P(n^{-1})$$

$$\sqrt{n}(\widehat{\sigma}_\epsilon^2 - \sigma_\epsilon^2) = X_V - \frac{2\sqrt{n}(\widehat{\beta}-\beta)^T X_N}{\sqrt{n}} + \frac{\sqrt{n}(\widehat{\beta}-\beta)^T\Sigma_{\eta\eta}\sqrt{n}(\widehat{\beta}-\beta)}{\sqrt{n}} + O_P(n^{-1}),$$

where $X_V = \sqrt{n}(S_{\widehat{\varepsilon}\widehat{\varepsilon}} - \sigma_\varepsilon^2)$ with $\widehat{\varepsilon}_i = \widehat{\upsilon}_i - \beta^T\widehat{\eta}_i = \widetilde{\varepsilon}_i + \frac{1}{\sqrt{n}}\sqrt{n}(\widehat{\beta}-\beta)^T\widehat{\eta}_i$.

The following theorem provides some information about the small sample properties of our semiparametric standard errors:

THEOREM 2. *Assume the same conditions as in Theorem 1, except that A2' replaces A2. Then,*

$$E[S^*] = \sqrt{n}h^{2q}[\frac{\zeta_q^{\theta\theta}}{2\sigma_\epsilon^2} - \frac{c^T\Sigma_{\eta\eta}^{-1}\zeta_q^{\eta\eta}\Sigma_{\eta\eta}^{-1}c}{2c^T\Sigma_{\eta\eta}^{-1}c}] + o(n^{-\mu}),$$

where $\zeta_q^{\theta\theta} = n^{-1}\sum_{i=1}^n[\widetilde{B}_i(\theta)]^2$ *and* $\zeta_q^{\eta\eta} = n^{-1}\sum_{i=1}^n \widetilde{B}_i(G_2)\widetilde{B}_i(G_2)^T$. *If, in addition,* u_i *are symmetrically distributed about zero, then*

$$Var[S^*] = s^2[1 + n^{-1}h^{-P}\{\sum_{i=1}^n\sum_{j\neq i} nh^P\rho_{ij}^2\}] + o(n^{-2\mu}),$$

where $s^2 = 1 + (\kappa_{4\epsilon} + \kappa_{4\eta^*})/4$, *with* $\kappa_{4\epsilon}$ *and* $\kappa_{4\eta^*}$ *being the standardised fourth cumulants of* ϵ_i *and* $\eta_i^* \equiv c^T\Sigma_{\eta\eta}^{-1}\eta_i$.

Interestingly, the asymptotic bias of $\widehat{\sigma}$ is of the same order of magnitude as that of $\widehat{\beta}$, i.e. of order $n^{-(\mu+1/2)}$, while its direction could be upward or downward according to the relative magnitudes of $\zeta_q^{\theta\theta}$ and $\zeta_q^{\eta\eta}$.

Andrews (1989b), Robinson (1989), Stoker (1989) and Delgado and Stengos (1994) establish the limiting behavior of various semiparametric testing procedures. Their small sample properties have been investigated by monte carlo experimentation. However, no general conclusion can be drawn from this analysis: Delgado and Stengos (1994) find a tendency towards under-rejection under the null, while Robinson (1989) finds both under and over-rejection for different configurations. In both cases, the power appeared to vary considerably with the model parameters as well as with bandwidth.

We provide analytical results that bear on the small sample behavior of the studentised statistic described in section 2, under the null hypothesis that β is the true value. The studentised statistic can be written as $W = T\{1 + n^{-1/2}S\}^{-1}$, and, expanding out the denominator and replacing T by T^*, we get

$$W = T^* - n^{-1/2}\sigma^{-1}c^T\overline{\Sigma}_{\eta\eta}^{-1}X_N S^* + O_P(n^{-1}) \equiv W^* + O_P(n^{-1}).$$

Then, since $Cov[X_N, n^{-1/2}X_N S^*] = o(n^{-2\mu})$, we have

THEOREM 3. *Assume the same conditions as in Theorem 1, except that A2' replaces A2. Then,* $E[W^*] = E[T^*] + o(n^{-\mu})$ *and* $Var[W^*] = Var[T^*] + o(n^{-2\mu})$.

If the third cumulants of u_i are zero, then, provided the moments exist, the skewness of both T^* and W^* (as well as higher cumulants) are $o(n^{-2\mu})$. In this case, the size of the hypothesis test based on W will, to second order, be largely affected by bandwidth through the location affect (the bias of $\widehat{\beta}$). Departures from nominal size could take either direction, and one may find over or under-rejections with this procedure. In situations where the location effect is small (i.e. when $\theta(\bullet)$ is close to linear), the scale effect should dominate and the test should over-reject.

4.3. Corrected Standard Errors

The formulae of Theorem 1 can be used to correct the standard errors for the small sample effect of bandwidth. In the special case of homoskedastic errors, the asymptotic variance of $\sqrt{n}c^T(\widehat{\beta}-\beta)$ is

$$\widetilde{\sigma}_n^2 = \sigma^2[1 + n^{-1}h^{-P}\{\sum\sum_{j\neq i} nh^P \rho_{ij}^2\}] + o(n^{-2\mu}).$$

In this case, we recommend using the modified quantity

$$\widehat{\sigma}^* = \widehat{\sigma}[1 + n^{-1}h^{-P}\ \{\sum\sum_{j\neq i} nh^P \rho_{ij}^2\}/2]$$

as standard errors for $\widehat{\beta}$. From Theorem 2, $\sqrt{n}(\widehat{\sigma}-\sigma) = \mathcal{X} + O_P(n^{-\mu})$, where $\mathcal{X} = O_P(1)$, and therefore $\widehat{\sigma} - \widetilde{\sigma}_n - n^{-1/2}\mathcal{X} = O_P(n^{-2\mu})$ and $\widehat{\sigma}^* - \widetilde{\sigma}_n - n^{-1/2}\mathcal{X} = o_P(n^{-2\mu})$. In this sense, the modified standard error provides better confidence intervals. The correction amounts to doing a degrees of freedom adjustment, similar in spirit to those discussed in Andrews (1991a).

The asymptotic bias of $\widehat{\beta}$ can also be estimated from the formula given in Theorem 1. This information can be used to adjust confidence statements about $\widehat{\beta}$ and to size adjust the test statistic, through Edgeworth type approximations as in Rothenberg (1984c) – see Cavanagh (1989) and Linton (1992). However, measuring the location effect requires estimation of the higher derivatives of the regression functions. In practice, it may be desirable to pursue a strategy that minimises the effect of bias on inference. If $\widehat{G}$ were sufficiently under smoothed, i.e. $hn^{\pi} \to 0$, the asymptotic bias of $\widehat{\beta}$ may be treated as negligible. In this case, $\widehat{\sigma}^*$ should provide well centred confidence intervals. This strategy is employed in many applications of nonparametric methods – see for example Bierens and Pott-Buter (1990) and Banks, Blundell and Lewbel (1992).

5. SECOND ORDER OPTIMALITY AND BANDWIDTH CHOICE

In this section, we develop an optimality theory for estimators of β using our second order moment approximations. We also suggest a feasible method for attaining the optimal performance.

Let $\Psi(q,P)$ denote the class of semiparametric least squares estimators of β based on the local polynomial regression scheme of order q with a bandwidth sequence of the form $h(n) = \gamma n^{-\pi}$, for any $\gamma > 0$. The optimal estimator within Ψ employs $h_0 = \gamma_0 n^{-\pi}$, where

$$\gamma_0 = (P\mathcal{V}/4q\mathcal{B}^2)^{1/(4q+P)}, \tag{12}$$

in which case, to a second order approximation,

$$MSE(h) \approx n^{-1}\left\{1 + n^{-2\mu}[(4q/P)^{-4q/(4q+P)} + (4q/P)^{P/(4q+P)}]\mathcal{B}^{2P/(4q+P)}\mathcal{V}^{4q/(4q+P)}\right\}. \tag{13}$$

This is a lower bound for all estimators within Ψ.

We examine (**??**) and (**??**) in the special case described in Section 5, and for simplicity take $\delta_1 = \delta_2 = \delta$. In Figs 5 and 6 below we graph h_0 and (**??**) as functions of δ for several sample sizes:

$* * *$ Figs $5 - 6$ here $* * *$

Even for very large δ, the optimal MSE is not much above its asymptotic value. This suggests there may be considerable worth in getting in a neighbourhood of the optimal bandwidth when δ is large.

The optimal bandwidth is not feasible, since γ_0 depends on the unknown quantities G and Σ. There are numerous techniques for automatic bandwidth choice in the nonparametric literature − see Härdle (1990) and Jones, Marron, and Sheather (1992) for reviews in the context of nonparametric regression and density estimation respectively. Many applied studies use a form of cross-validation, in which h is chosen to minimise a suitable criterion function such as a (leave-one-out) least squares criterion (Stock (1991) and Engle et al (1986)) or a pseudo-likelihood (Robinson (1991a). In nonparametric regression and density estimation this method gives the optimal bandwidth when judged by an asymptotic (integrated) MSE error criterion − see Hall (1983).

For semiparametric problems, few such optimality results exist − Robinson (1991a) shows only that using a cross-validated bandwidth does not affect the first order limiting distribution. In our case, the optimal bandwidths for estimating β and for estimating G are of different orders of magnitude; therefore, the bandwidth selected by cross-validation will be the wrong order of magnitude. We consider an alternative method − the plug-in. This consists of two stages: firstly, γ_0 is estimated consistently by $\widehat{\gamma}_0$ say, and then $\widehat{\gamma}_0 n^{-\pi}$ defines the window used to estimate β in a final stage. This method was employed by Härdle, Hart, Marron, and Tsybakov (1992) for $\sqrt{n}$ consistent semiparametric estimators of average derivatives.

We now demonstrate that such a procedure can attain the second order bound (**??**). Let $\widehat{\gamma}_0$ be any n^{φ} consistent estimator of γ_0, where $\mu < \varphi \leq 1/2$. For technical reasons − see Bickel (1982) for a similar application of this device − we shall employ a discretised and Winsorized version $\widetilde{\gamma}_0$ that is defined as the closest point to $\widehat{\gamma}_0$ in the grid

$$\left\{ in^{-\varphi} : \ 0 < \gamma_L \leq in^{-\varphi} \leq \gamma_U < \infty, \ i \ integer \right\},$$

where γ_L and γ_U are selected in advance. Provided $\gamma_0 \in [\gamma_L, \gamma_U]$, $\widetilde{\gamma}_0$ is also n^{φ} consistent. We use $\widetilde{h}_0 = \widetilde{\gamma}_0 n^{-\pi}$ to define our final estimator of β, $\widehat{\beta}(\widetilde{h}_0)$.

Firstly, note that

$$T(\widetilde{h}_0) = T^*(\widetilde{h}_0) + R_T(\widetilde{h}_0),$$

where $T^*(\widetilde{h}_0)$ and $R_T(\widetilde{h}_0)$ are the truncated and remainder statistics defined in (**??**) and (**??**) evaluated at the bandwidth $\widetilde{h}_0$. By the mean value theorem,

$$T^*(\widetilde{h}_0) = T^*(h_0) + (\widetilde{h}_0 - h_0)\frac{\partial T^*}{\partial h}(\widetilde{h}_0^*),$$

where $\widetilde{h}_0^*$ lies between $\widetilde{h}_0$ and h_0. The following theorem establishes that both $(\widetilde{h}_0 - h_0)\frac{\partial T^*}{\partial h}(\widetilde{h}_0^*)$ and $R_T(\widetilde{h}_0)$ are small in distribution.

THEOREM 4. *Let $\phi < \varphi$. Assume A1, A2', and A3-A5 hold, with $J, L \geq M_2$, where $M_2 \equiv Max\{\frac{16\mu+8(\varphi-\phi)}{\Xi+1}, \frac{8\mu+4(\varphi-\phi)}{\Xi+1/2}, \frac{4\mu+2(\varphi-\phi)}{\varphi-\mu}, \frac{32\mu+16(\varphi-\phi)}{3-4\mu}\}$, with $\Xi = \varphi - \mu - [\mu - \mu^*]$ and $\mu^* = (4q - 3P)/2(4q + P)$, and that $\gamma_0 \in [\gamma_L, \gamma_U]$. Suppose also that for some $\lambda > 0$,*

$$\text{A6.}\quad \Pr[n^{\phi}|\widetilde{\gamma}_0 - \gamma_0| > \lambda] = o(n^{-2\mu}).$$

Then the distributions of $T(\widetilde{h}_0)$ and $T^*(h_0)$ differ by $o(n^{-2\mu})$.

Therefore, the plug-in estimator is asymptotically equivalent to order $n^{-2\mu}$ in distribution to the second order optimal estimator.

We now discuss how γ_0 is to be estimated. Clearly, Σ can be replaced by $\widehat{\Sigma}$; the question is how to estimate:

$$\widetilde{B}_i(g; h_0) = h_0^{-q}[\sum_{j\neq i} w_{ij}(h_0)g(Z_j) - g(Z_i)]\ ;\ \mathcal{V}(h_0) = \sum\sum_{j\neq i} n h_0^P \rho_{ij}^2(h_0). \qquad (14)$$

Note that $\mathcal{V}(h_0) \approx \mathcal{V}(h)$ for any bandwidth sequence h satisfying A4. Therefore, $\mathcal{V}(h_0)$ can be estimated by $\mathcal{V}(h)$. We discuss two alternative methods of estimating $\widetilde{B}_i(g; h_0)$. Firstly, take $\widetilde{B}_i(\widehat{g}_{h^*}; h^*)$, where $\widehat{g}_{h^*}$ is a preliminary estimate of g based on a bandwidth h^*. This method is convenient to use, and performed well in the simulations reported below. The second method uses the asymptotic approximation $\widetilde{B}_i(g; h_0) \approx \mathcal{D}_k^q g(Z_i)$ and replaces $\mathcal{D}_k^q g(Z_i)$ by an estimate $\widehat{\mathcal{D}_k^q g}(Z_i)$. The $q'th$ partial derivatives of g can be estimated by a local polynomial regression estimator of order $q^* > q$ with bandwidth h^*, but also by the series method of Andrews (1991a).

We now prove that this latter method can satisfy (A6). Let ψ be the vector containing $\zeta_q^{\eta\theta}$ and the unique elements of Σ. Then $\gamma_0 = \gamma(\psi_0)$, where $\gamma(\bullet)$ is continuously differentiable at ψ_0. By the mean value theorem

$$\widehat{\gamma}_0 - \gamma_0 = \gamma'(\psi_0^*)^T(\widehat{\psi}_0 - \psi_0),$$

where γ' is the first derivative vector of γ, while ψ_0^* is intermediate between $\widehat{\psi}_0$ and ψ_0. Properties of average derivative estimators similar to $\widehat{\zeta_q^{\eta\theta}} = n^{-1}\sum_{i=1}^n \widehat{\mathcal{D}_k^q\theta}(Z_i)\widehat{\mathcal{D}_k^q G}_2(Z_i)$ are considered in Andrews (1991a), Härdle and Stoker (1989), and Powell, Stock and Stoker (1989). When G is sufficiently smooth, $\widehat{\zeta_q^{\eta\theta}}$ (and hence $\widehat{\gamma}_0$) should be consistent, and even $\sqrt{n}$ consistent under moment and smoothness conditions and appropriate restrictions on q^* and h^*. If $Sup_n E[|n^{\varphi}(\widehat{\psi}_0 - \psi_0)|^F] < \infty$, where $F(\varphi - \phi) > 2\mu$, then

$$\Pr[|n^{\phi}(\widehat{\psi}_0 - \psi_0)| > \lambda] \leq \frac{E[|n^{\varphi}(\widehat{\psi}_0 - \psi_0)|^F]}{\lambda^F n^{(\varphi-\phi)F}} = o(n^{-2\mu})$$

by Markov's inequality. That the required moments exist can be verified for local polynomial estimators of average derivatives under smoothness and moment conditions. To handle $\gamma'(\psi_0^*)$ a truncation argument is used.

COROLLARY: *Assume that A1-A5 hold, with J and $L> M_3 \equiv 4\mu/[(\varphi-\phi)]$, and in addition $G \in \mathcal{G}_{q^*}$. Further suppose that $nh^{*2(q^*-q)} \to 0$ and $nh^{*(P+q)} \to \infty$. Then* A6 *is satisfied.*

Whatever method is employed to estimate γ_0, an additional smoothing parameter has to be selected, and in this sense the plug-in method is not fully automatic. However, evidence presented in Park and Marron (1990) and Sheather and Jones (1991) suggests that the final estimate may not be so sensitive to the choice of the secondary smoothing parameter, and one can employ some arbitrary rule for estimation of the bandwidth constants with little cost. The so-called rule of thumb approach, see Silverman (1986) and Andrews (1991b), offers an alternative plug-in implementation that does not require an explicit preliminary bandwidth to be chosen. In this approach one specifies, for the purposes of bandwidth choice only, a parametric model for $G(\bullet)$ such as making it a polynomial function of Z. Parametric procedures are then used to get a preliminary fit $\widehat{G^*}$ and derivatives thereof which are then plugged into the optimal bandwidth formula. This method achieves the more modest objective of being second order optimal for the particular model chosen for G, although the correct order of magnitude for $\widehat{h}$ is guaranteed for all G.

6. SIMULATIONS

This section contains the results of a simulation experiment designed to evaluate our bandwidth selection procedure.

We generated $20,000$ samples of size $n = 100$ from:

$$y_{1i} = 1 + y_{2i} + 0.5\delta_1 Z_i^2 + \epsilon_i, \;\; y_{2i} = \frac{1}{\ln(\delta_2+1)} \exp(\ln(\delta_2+1)Z_i) + \eta_i,$$

where (ϵ_i, η_i) were iid Gaussian with unit covariance matrix, while $Z_i = i/n, i = 1, 2, .., n$. Note that $n^{-1}\sum_{i=1}^{n} \theta''(Z_i) g_2''(Z_i) = \delta_1\delta_2$. We examined the following parameter values: I $(\delta_1 = -20, \delta_2 = 4)$, II $(\delta_1 = -50, \delta_2 = 10)$, III $(\delta_1 = -100, \delta_2 = 20)$, and IV $(\delta_1 = -150, \delta_2 = 30)$. All regression functions were estimated by local linear regression with quadratic kernel. Equally spaced designs are not encountered frequently in econometric applications. However, our primary purpose here is to evaluate the quality of our asymptotic approximations and for this reason we adopt the most convenient sampling scheme that allows us to vary what we think are the key parameters of our formulae.

We first evaluated the performance of $\widehat{\beta}$ for each of models I-IV at a grid of 20 different bandwidths. Figures $7 - 10$ show the simulation MSE (scaled by 100) for models I-IV.

Figures 7-10 here

These results confirm the validity of our second order approximations; witness Figure 11 which superimposes the simulation MSE and second order approximation (calculated from (**??**)) for model IV.

Figure 11 here

The approximations are remarkably close to the actual performance of the estimator for bandwidths in the intermediate range close to the optimum.

Finally, we evaluated our plug-in procedure. The initial bandwidth $h^* = 0.25$ was used to estimate the bias and variance constants by

$$\widetilde{B}_i(\widehat{g}_{h^*}; h^*) = h^{*-2}\left[\sum_{j\neq i} w_{ij}(h^*)\widehat{g}_{h^*}(Z_j) - \widehat{g}_{h^*}(Z_i)\right] \; ; \; \mathcal{V}(h^*) = \sum_{j\neq i}\sum n h^* \rho_{ij}^2(h^*).$$

We did not Winsorize nor discretize the resulting $\widehat{h}_0$. Figure 12 shows the density of $\ln(\widehat{h}_0)$ for models I-IV.

Figure 12 here

Finally, the performance of the resulting estimator $\widehat{\beta}(\widehat{h}_0)$ is given below:

TABLE 1. PERFORMANCE OF $\widehat{\beta}(\widehat{h}_0)$

	Moments		**Quantiles**				
Model	$\overline{x}$	s	**1%**	**25%**	**50%**	**75%**	**99%**
I	0.9971	0.1045	0.7511	0.9269	0.9966	1.0663	1.2451
II	0.9973	0.1058	0.7498	0.9270	0.9968	1.0683	1.2472
III	0.9971	0.1074	0.7462	0.9248	0.9976	1.0689	1.2521
IV	0.9969	0.1087	0.7415	0.9239	0.9975	1.0690	1.2542

where $\overline{x}$ and s denote simulation mean and standard deviation respectively.

7. CONCLUSIONS

Our work suggests several qualitative predictions about the higher order properties of the various semiparametric statistics we considered. When using such procedures the practitioner has considerable latitude in choosing the number of nuisance parameters to estimate. While this choice is not reflected in the first order theory, it can have considerable impact on the actual performance of estimators and test statistics.

In particular, we found a variance inflation proportional to $n^{-1}h^{-P}$ which for small bandwidths can be of comparable magnitude to the asymptotic variance. The bias of the estimator is more difficult to measure since it depends on the unknown regression functions, nevertheless the direction of the bias may be inferred from information about their concavity.

Accounting for these higher order properties is essential if reliable inference is to be carried out; we suggested ways of accomplishing this. In particular, the degrees of freedom adjustment considered in section 6 is relatively easy to implement and at least reflects the bandwidth in a way which guards against its inappropriate use.

We also constructed an estimator of β that was second order optimal. Although our standard of optimality compares estimators only within a much more restricted class than the first order optimality theory of Bickel et al (1993) and Newey (1990a), it does provide a benchmark against which to compare alternative first order equivalent procedures. In practice, our procedure seems to work quite well even for samples as small as 100, provided there is not too much nonlinearity.

Our approximations are based on computing the first two moments of the truncated Taylor series approximation to the standardised statistics. These calculations can be carried out for a much wider class of semiparametric models possessing smoothness properties. Some preliminary work dealing with such a general situation is presented in Linton (1992).

ACKNOWLEDGEMENTS

I would like to thank Tom Rothenberg, Chris Cavanagh, Don Andrews and Wolfgang Härdle for helpful comments. The co-editor and two referees deserve considerable credit for their help in preparing this manuscript. Remaining errors are mine.

Cowles Foundation for Research in Economics, Yale University.

APPENDIX

In the first section we give some preliminary lemmas that describe the properties of the local polynomial estimator and various averages derived from it, while in section B we give proofs of the theorems.

In the sequel, let $|a| = (a'a)^{1/2}$ be the Euclidean norm of the P-vector a, and let $\|A\| = \lambda_{\max}(A) \equiv Max_a \frac{a'Aa}{a'a}$ be the matrix norm of any real symmetric matrix A, where $\lambda_{\max}$ denotes largest eigenvalue. We make use of the following argument. Let a and b be vectors and let A and B be matrices. Then $|a'ABb| \leq (a'A^2a)^{1/2}(b'B^2b)^{1/2}$ by the Cauchy-Schwarz inequality, while for positive definite A, $a'Aa = a'a\frac{a'Aa}{a'a} \leq a'a\|A\|$. Therefore,

$$|a'ABb| \leq |a||b|\,\|A\|\,\|B\|\,. \tag{15}$$

Finally, we will use χ_1, χ_2, .. to denote positive finite constants.

A. PRELIMINARY LEMMAS

A1. *Properties of Local Polynomial Weights*

We now derive various properties of the weighting sequence $\{w_{ij}\}_{i,j=1}^n$ defined in (6) with $h = O(n^{-\pi})$.

LEMMA 2: $\sum_{j\neq i} w_{ij} = 1$; $\sum_{i\neq j} w_{ij} = 1 + o(1)$.

LEMMA 3: *Let* $\rho_{ij} = n^{-1/2}[\sum_{k\neq j,k\neq i} w_{ki}w_{kj} - 2w_{ij}]$ *and let* $\rho_{ii} = n^{-1/2}\sum_{k\neq i} w_{ki}^2$. *Then* $\sum\sum_{i\neq j} n^{-1}w_{ij}^2$, $\sum\sum_{i\neq j} \rho_{ij}^2$, *and* $n^{-1/2}\sum_{i=1}^n \rho_{ii}$ *are* $O(n^{-1}h^{-P})$, *while* $\sum_{i=1}^n \rho_{ii}^2 = O(n^{-2}h^{-2P})$.

PROOF*:* By the triangle inequality, we have

$$\sum\sum_{i\neq j} \rho_{ij}^2 \leq 8\{\sum\sum_{i\neq j} [n^{-1/2}(\sum_{k\neq j,k\neq i} w_{ki}w_{kj})]^2 \ + n^{-1}\sum\sum_{i\neq j} w_{ij}^2\},$$

and by Cauchy-Schwarz

$$[n^{-1/2}\sum_{k\neq j,k\neq i} w_{ki}w_{kj}]^2 \leq [n^{-1/2}\sum_{k\neq j,k\neq i} w_{ki}^2][n^{-1/2}\sum_{k\neq j,k\neq i} w_{kj}^2].$$

By assumptions A1 and A5:

C1 $\#\{j:\ w_{ij}\neq 0\} = O(nh^P)$

C2 $|w_{ij}| \leq \chi_1 n^{-1}h^{-P}$,

where #A denotes the cardinality of a set A, and **C1** holds uniformly in i. Therefore, $\sum_{k\neq j,k\neq i} w_{ki}^2 \leq \chi_2 n^{-1}h^{-P}$ uniformly in i, and the first two results follow. Also, by interchanging summations, $n^{-1/2}\sum_{i=1}^n \rho_{ii} = n^{-1}\sum_{k\neq i}(\sum_{i=1}^n w_{ki}^2)$, which is $O(n^{-1}h^{-P})$ by the same arguments as above. Similarly for $\sum_{i=1}^n \rho_{ii}^2$.

LEMMA 4*:* $\sum\sum_{i\neq j} [\frac{\partial\rho_{ij}}{\partial h}]^2 = O(n^{-1}h^{-(P+2)})$.

PROOF. Applying the triangle inequality and Cauchy-Schwarz, it suffices to establish the order of magnitude of $\sum\sum_{i\neq j} [\frac{\partial w_{ij}}{\partial h}]^2$. Let $\widetilde{K}_i$ be the diagonal matrix with typical element

$$\frac{\partial}{\partial h}k(\frac{Z_j - Z_i}{h}) = -h^{-1}\sum_{p=1}^{P} k_p(\frac{Z_j - Z_i}{h})(\frac{Z_{jp} - Z_{ip}}{h}),$$

where $k_p(t) = \frac{\partial}{\partial t_p}k(t)$. Then, because each $k_p(t)$ is bounded and of bounded support, there exists a χ_3 such that

$$\underset{i,j\leq n}{Max}\,|\frac{\partial}{\partial h}k(\frac{Z_j - Z_i}{h})| \leq \chi_3 h^{-1}.$$

Therefore, since $\frac{\partial w_{ij}}{\partial h}$ is the j'th element of the first row of

$$-(X_i^T K_i X_i)^{-1}(X_i^T \widetilde{K}_i X_i)(X_i^T K_i X_i)^{-1} X_i^T K_i + (X_i^T K_i X_i)^{-1} X_i^T \widetilde{K}_i,$$

there exists χ_4 such that $|\frac{\partial w_{ij}}{\partial h}| \leq \chi_4 n^{-1} h^{-(P+1)}$. Furthermore, $\#\{(i,j) : \frac{\partial w_{ij}}{\partial h} \neq 0\} = O(n^2 h^P)$, by the bounded support of k. Therefore, $n^{-1}h^{-(P-1)}H^{-1}X_i^T \widetilde{K}_i X_i H^{-1}$, $n^{-1}h^{-(P-1)}H^{-1}X_i^T \widetilde{K}_i K_i X_i H^{-1}$ and $n^{-1}h^{-(P-2)}H^{-1}X_i^T \widetilde{K}_i^2 X_i H^{-1}$ are $O(1)$, uniformly in i, and the result follows.

A2. *Properties of Standardised Averages of Local Polynomial Estimators*

The truncated statistics T^*, S^*, and W^* depend on the weighted U-statistics X_D, X_N, and X_V, where:

$$\begin{aligned}
X_N &\equiv \sqrt{n} S_{\hat{\eta}\hat{\epsilon}} &&= n^{-1/2} \textstyle\sum_{i=1}^n [\eta_i - (B_{\eta i} + V_{\eta i})][\epsilon_i - (B_{\theta i} + V_{\epsilon i})] \\
X_D &\equiv \sqrt{n}(S_{\hat{\eta}\hat{\eta}} - \overline{\Sigma}_{\eta\eta}) &&= n^{-1/2} \textstyle\sum_{i=1}^n \left\{ [\eta_i - (B_{\eta i} + V_{\eta i})][\eta_i - (B_{\eta i} + V_{\eta i})]^T - \overline{\Sigma}_{\eta\eta} \right\} \\
X_V &\equiv \sqrt{n}(S_{\widehat{\varepsilon}\widehat{\varepsilon}} - \overline{\sigma}_\varepsilon^2) &&= n^{-1/2} \textstyle\sum_{i=1}^n \left\{ [\varepsilon_i - (B_{\theta i} + V_{\varepsilon i})]^2 - \overline{\sigma}_\varepsilon^2 \right\},
\end{aligned}$$

where $S_{\widehat{\varepsilon}\widehat{\varepsilon}} = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i^2$, with $\widehat{\varepsilon}_i = \widehat{\upsilon}_i - \beta^T \widehat{\eta}_i$, while $n^{-1}\sum_{i=1}^n \varepsilon_i^2 \xrightarrow{P} \overline{\sigma}_\varepsilon^2$, and $B_{\eta i} = (B_{2i}, .., B_{D+1,i})^T$, $V_{\eta i} = (V_{2i}, .., V_{D+1,i})^T$, with for $d = 1, 2, .., D+1$:

$$\begin{aligned}
&V_{di} = \textstyle\sum_{j \neq i} w_{ij}[y_{dj} - g_d(Z_j)] \ ; \ B_{di} = \sum_{j \neq i} w_{ij}[g_d(Z_j) - g_d(Z_i)], \\
&B_{\theta i} = B_{1i} - \beta^T B_{\eta i} = \textstyle\sum_{j \neq i} w_{ij}[\theta(Z_j) - \theta(Z_i)] \ ; \ V_{\epsilon i} = V_{1i} - \beta^T V_{\eta i} = \sum_{j \neq i} w_{ij}\epsilon_j.
\end{aligned}$$

Note that for example $B_{1i} = h^2 \widetilde{B}_i(g_1)$. By expanding out the brackets and collecting terms we obtain

$$X_N = \widetilde{X}_N + b_N + L_N + Q_N, \ X_D = \widetilde{X}_D + b_D + L_D + Q_D, \ X_V = \widetilde{X}_V + b_V + L_V + Q_V,$$

where the $O_P(1)$ leading terms are:

$$\widetilde{X}_V = n^{-1/2}\sum_{i=1}^{n}(\epsilon_i^2 - \overline{\sigma}_\epsilon^2)\ ;\ \widetilde{X}_N = n^{-1/2}\sum_{i=1}^{n}\eta_i\epsilon_i\ ;\ \widetilde{X}_D = n^{-1/2}\sum_{i=1}^{n}(\eta_i\eta_i^T - \overline{\Sigma}_{\eta\eta}),$$

while the biases terms are:

$$b_N = n^{-1/2}\sum_{i=1}^n B_{\eta i}B_{\theta i}$$
$$b_D = n^{-1/2}\sum_{i=1}^n B_{\eta i}B_{\eta i}^T + n^{-1/2}\sum_{i=1}^n E[V_{\eta i}V_{\eta i}^T] \equiv b_{D1} + b_{D2}$$
$$b_V = n^{-1/2}\sum_{i=1}^n B_{\theta i}^2 + n^{-1/2}\sum_{i=1}^n E[V_{\epsilon i}^2] \equiv b_{V1} + b_{V2}.$$

The linear and quadratic statistics are

$$L_N = n^{-1/2}\sum_{i=1}^n(V_{\eta i} - \eta_i)B_{\theta i} + n^{-1/2}\sum_{i=1}^n B_{\eta i}(V_{\epsilon i} - \epsilon_i) + \sum_{i=1}^n \rho_{ii}\epsilon_i\eta_i \equiv \sum_{j=1}^3 L_{Nj}$$
$$L_D = n^{-1/2}\sum_{i=1}^n(V_{\eta i} - \eta_i)B_{\eta i}^T + n^{-1/2}\sum_{i=1}^n B_{\eta i}(V_{\eta i} - \eta_i)^T + \sum_{i=1}^n \rho_{ii}(\eta_i\eta_i^T - \overline{\Sigma}_{\eta\eta}) \equiv \sum_{j=1}^3 L_{Dj}$$
$$L_V = 2n^{-1/2}\sum_{i=1}^n(V_{\epsilon i} - \epsilon_i)B_{\theta i} + \sum_{i=1}^n \rho_{ii}(\epsilon_i^2 - \overline{\sigma}_\epsilon^2) \equiv L_{V1} + L_{V2}$$
$$Q_N = \sum\sum_{i\neq j} \rho_{ij}\epsilon_i\eta_j\ ;\ Q_D = \sum\sum_{i\neq j} \rho_{ij}\eta_i\eta_j^T\ ;\ Q_V = \sum\sum_{j\neq i} \rho_{ij}\epsilon_i\epsilon_j,$$

where $\rho_{ii} = n^{-1/2}\sum_{k\neq i} w_{ki}^2$ and $\rho_{ij} = n^{-1/2}[\sum_{k\neq j,k\neq i} w_{ki}w_{kj} - 2w_{ij}],\ j \neq i.$

The properties of X_N, X_D, and X_V can be determined from the lemmas below.

LEMMA 5*: As* $n \to \infty$,

$$b_N \approx \sqrt{n}h^{2q}\{n^{-1}\sum_{i=1}^n \mathcal{D}_k^q\theta\mathcal{D}_k^q G_2(Z_i)\}$$
$$b_{D1} \approx \sqrt{n}h^{2q}\{n^{-1}\sum_{i=1}^n \mathcal{D}_k^q G_2\mathcal{D}_k^q G_2^T(Z_i)\}$$
$$b_{V1} \approx \sqrt{n}h^{2q}\{n^{-1}\sum_{i=1}^n [\mathcal{D}_k^q\theta(Z_i)]^2\}.$$

PROOF. Lemma 1 can be applied directly.

LEMMA 6*:* $b_{D2}, b_{V2} = \sum\sum_{i\neq j} n^{-1/2}w_{ij}^2\sigma(Z_j) = O(n^{-1/2}h^{-P})$, where σ^2 is either the scalar σ_ε^2 or the matrix $\Sigma_{\eta\eta}$. Follows from **C1** and **C2**.

LEMMA 7*: $L_{D1}, L_{D2}, L_{V1}, L_{N1}, L_{N2} = o_P(h^q)$, but $L_{D3}, L_{N3}, L_{V2} = O_P(n^{-1}h^{-P})$.*

PROOF. Consider $L_{N1} = n^{-1/2}\sum_{i=1}^{n}(V_{\eta i} - \eta_i)B_{\theta i}$. By interchanging summations, we obtain $L_{N1} = n^{-1/2}\sum_{i=1}^{n}\xi_i\eta_i$, where $\xi_i = [\sum_{j\neq i} w_{ji}B_{\theta j} - B_{\theta i}]$. Therefore,

$$\|Var[L_{N1}]\| = \left\|n^{-1}\sum_{i=1}^{n}\xi_i^2\Sigma_{\eta\eta}(Z_i)\right\| \leq \chi_5 n^{-1}\sum_{i=1}^{n}\xi_i^2.$$

Furthermore, by the triangle inequality and Cauchy-Schwarz,

$$\xi_i^2 \leq 2[(\textstyle\sum_{j\neq i} w_{ji} - 1)B_{\theta i}]^2 + 2\{\underset{j\leq n}{Max}\,(B_{\theta j} - B_{\theta i})^2\}\sum_{j\neq i} w_{ji}^2] = o(h^{2q}).$$

uniformly in i, because $\sum_{j\neq i} w_{ji}^2 = O(n^{-1}h^{-P})$ by **C1** and **C2** and $\sum_{j\neq i} w_{ji} - 1 = o(1)$ by Lemma 2.

For L_{N3} we have

$$\|Var[L_{N3}]\| = \left\|\sum_{i=1}^{n}\rho_{ii}^2\sigma_\epsilon^2(Z_i)\Sigma_{\eta\eta}(Z_i)\right\| \leq \chi_6|\sum_{i=1}^{n}\rho_{ii}^2| = O(n^{-2}h^{-2P}),$$

by the boundedness of $\sigma_\epsilon^2(Z_i)\Sigma_{\eta\eta}(Z_i)$ and Lemma 3**.**

LEMMA 8*: $Q_D, Q_N, Q_V = O_P(n^{-1/2}h^{-P/2})$. Furthermore,*

$$\begin{aligned} Var[Q_N] &= \sum\sum_{i\neq j}\rho_{ij}^2\sigma_\epsilon^2(Z_i)\Sigma_{\eta\eta}(Z_j) &\approx& \textstyle\sum_i \sigma_\epsilon^2(Z_i)\Sigma_{\eta\eta}(Z_i)\sum_{j\neq i}\rho_{ij}^2 \\ Var[Q_V] &= \sum\sum_{i\neq j}\rho_{ij}^2\sigma_\epsilon^2(Z_i)\sigma_\epsilon^2(Z_j) &\approx& \textstyle\sum_i \sigma_\epsilon^4(Z_i)\sum_{j\neq i}\rho_{ij}^2 \end{aligned}$$

PROOF. The order of magnitude follows from Lemma 3, because $\Sigma(Z_j)$ is uniformly bounded. The approximation is valid because $\Sigma \in \mathcal{G}_1$.

LEMMA 9*: Assume that A1-A5 hold, where J,L$>2M$ for some integer M. Then $Sup_nE[|X_N|^M] < \infty$ and $Sup_nE[\|X_D\|^M] < \infty$.*

PROOF. First, using a standard inequality

$$E[|X_N|^M] \le \chi_7(E[|\widetilde{X}_N|^M] + |b_N|^M + E[|L_N|^M] + E[|Q_N|^M]).$$

By standard results for independent random variables, $2M$ uniformly bounded moments are sufficient to guarantee that $Sup_n E[|n^{-1/2}\sum_{i=1}^{n}\eta_i\epsilon_i|^M] < \infty$. Similarly $Sup_n E[|L_N|^M] < \infty$, while $Sup_n |b_N|^M < \infty$ because of the boundedness of the q'th derivatives of G.

We now establish that $Sup_n E[|Q_N|^M] < \infty$, where

$$Q_N \equiv \sum\sum_{j>i} \rho_{ij}\epsilon_i\eta_j + \sum\sum_{i>j} \rho_{ij}\epsilon_i\eta_j.$$

Without loss of generality restrict attention to $\sum\sum_{i<j} \rho_{ij}\epsilon_i\eta_j$. Let $W_i = \sum_{j>i}\rho_{ij}\epsilon_i\eta_j$ for each i, then $\sum\sum_{i<j} \rho_{ij}\epsilon_i\eta_j = \sum_i W_i$, where $\{W_i\}$ forms a martingale difference sequence with respect to the sigma field generated by $\{u_{i-1}, .., u_1\}$. By the standard martingale techniques of Hall and Heyde (1980),

$$E[|\sum_i W_i|^M] \le \chi_8(\sum\sum_{j>i} \rho_{ij}^2)^{M/2},$$

provided $Sup_i E[|\epsilon_i|^M] < \infty$ and $Sup_i E[|\eta_i|^M] < \infty$, see Mikosch (1991), Lemma 1.3. Therefore, applying Lemma 3, the result is established.

The same arguments apply to X_D.

LEMMA 10: *Assume that A1-A5 hold, where J,L> $2M$ for some integer M. Then*

$$\mathbf{L1}.\ Sup_n E[|n^{(\mu-\pi)}\frac{\partial X_N}{\partial h}|^M] < \infty\ ;\ \mathbf{L2}.\ Sup_n E[\left\| n^{(\mu^*-\pi)}\frac{\partial X_D}{\partial h}\right\|^M] < \infty,$$

where $\mu^* = (4q-3P)/2(4q+P)$.

PROOF OF **L1**. By A3, $\frac{\partial b_N}{\partial h} = O(\sqrt{n}h^{2q-1})$, where $h = O(n^{-\pi})$, so that $n^{(\mu-\pi)}\frac{\partial b_N}{\partial h}(h) = O(1)$. We consider $\frac{\partial L_{N1}}{\partial h}(h) = n^{-1/2}\sum_{i=1}^{n}\frac{\partial \xi_i}{\partial h}\eta_i$, where

$$\frac{\partial \xi_i}{\partial h} = (\sum_{j \neq i} [\frac{\partial w_{ji}}{\partial h} B_{\theta j} + w_{ji} \frac{\partial B_{\theta j}}{\partial h}] - \frac{\partial B_{\theta i}}{\partial h}).$$

Since $\sum_{j \neq i} [\frac{\partial w_{ji}}{\partial h}]^2 = O(n^{-1} h^{-(P+2)})$ and $\frac{\partial B_{\theta i}}{\partial h} = O(h^{q-1})$, we have $Var[\frac{\partial L_{N1}}{\partial h}(h)] = O(h^{2q-2})$. Therefore, since the required moments exist, $E[|n^{(\mu-\pi)} \frac{\partial L_{N1}}{\partial h}(h)|^M] = O(1)$. Similarly for $\frac{\partial L_{N2}}{\partial h}$. Writing $\frac{\partial Q_N}{\partial h} = \sum\sum_{i \neq j} \frac{\partial \rho_{ij}}{\partial h} \epsilon_i \eta_j$, we have

$$E[|n^{(\mu-\pi)} \frac{\partial Q_N}{\partial h}(h)|^M] \leq \chi_9 \{\sum\sum_{i \neq j} n^{2(\mu-\pi)} [\frac{\partial \rho_{ij}}{\partial h}]^2\}^{M/2},$$

where $\frac{\partial \rho_{ij}}{\partial h} = \frac{\partial}{\partial h}[n^{-1/2}(\sum_{k \neq j, k \neq i} w_{ki} w_{kj} - 2w_{ij})]$, provided $Sup_i E[|\epsilon_i|^M] < \infty$ and $Sup_i E[|\eta_i|^M] < \infty$. By Lemma 4, this is bounded. Therefore, $E[|n^{(\mu-\pi)} \frac{\partial Q_N}{\partial h}(h)|^M] = O(1)$. By the triangle inequality, $Sup_n E[|n^{(\mu-\pi)} \frac{\partial X_N}{\partial h}|^M] < \infty$ as required.

Similar arguments apply to **L2**, although in this case $\frac{\partial b_{D2}}{\partial h} = O(n^{-1/2} h^{-(P+1)})$, which accounts for the different value of the norming constant μ^*.

B. PROOF OF THEOREMS

PROOF OF THEOREM 1

(1.1) To establish this we use the following result of Sargan and Mikhail (1971): for all x and ζ,

$$|\Pr[T \leq x] - \Pr[T^* \leq x]| \leq \Pr[|R_T| > \zeta] + \Pr[|T^* - x| < \zeta], \tag{16}$$

see Rothenberg (1984a). Provided T^* has a bounded density, the last term is $O(\zeta)$ as $\zeta \to 0$. We therefore choose $\zeta = O([n^{2\mu} \log n]^{-1})$ and show that for some positive constant χ_{10},

$$\Pr[|R_T| > \frac{\chi_{10} n\sqrt{n}}{n^{2\mu} \log n}] = o(n^{-2\mu}).$$

Let $A = \left\{|R_T| > \frac{\chi_{10} n\sqrt{n}}{n^{2\mu} \log n}\right\}$ and $B = \{\|S_{\hat{\eta}\hat{\eta}}\| > \chi_{11}\}$. We have

$$\Pr[A] \leq \Pr[A \cap B] + \Pr[B^c], \tag{17}$$

where

$$\Pr(B^c) \leq \Pr(\|X_D\| > \chi_{12}\sqrt{n}) \leq \chi_{13} \frac{E[\|X_D\|^2]}{n}$$

by Markov's inequality. Therefore, provided $Sup_n E[\|X_D\|^2] < \infty$, $\Pr[B^c] = o(n^{-2\mu})$.

By (A.**??**): when $\|S_{\hat{\eta}\hat{\eta}}\| > \chi_{11}$, $|R_T| \leq \chi_{14} |X_N| \, \|X_D\|^3$. Therefore $A \cap B \subset \overline{A}$, where $\overline{A} = \left\{|X_N| \, \|X_D\|^3 > \frac{\chi_{15} n\sqrt{n}}{n^{2\mu} \log n}\right\}$, and

$$\Pr[A \cap B] \leq \Pr[\overline{A}] \leq \chi_{16} [\log n]^\rho \frac{E[|X_N|^\rho \, \|X_D\|^{3\rho}]}{n^{(\frac{3}{2} - 2\mu)\rho}}$$

for any ρ. Provided the relevant moments are uniformly bounded and $(\frac{3}{2} - 2\mu)\rho > 2\mu$, the result follows.

Since J,$L > M_1$, the relevant moments exist, by the following argument. By Hölder's inequality

$$E[|X_N|^{\rho}\,\|X_D\|^{3\rho}] \leq \{E[|X_N|^{4\rho}]\}^{1/4}\{E[\|X_D\|^{4\rho}]\}^{3/4},$$

while $Sup_n E[\|X_D\|^{4\rho}] < \infty$ and $Sup_n E[|X_N|^{4\rho}] < \infty$, provided 8ρ moments exist, by Lemma 9.

(1.2) By (A.**??**) and Hölder's inequality,

$$E|T^*|^2 \leq \chi_{17} E[|X_N|^2\,\|X_D\|^4] \leq \chi_{17}\{E[|X_N|^6]\}^{1/3}\{E[\|X_D\|^6]\}^{2/3},$$

while $Sup_n E[\|X_D\|^6] < \infty$ and $Sup_n E[|X_N|^6] < \infty$ by Lemma 9, since $J, L \geq 12$.

(1.3) Since $X_D = O_P(1)$, we can drop the last term on the right of (**??**). Furthermore, some of the cross terms in $\frac{c^T\overline{\Sigma}_{\eta\eta}^{-1}X_D\overline{\Sigma}_{\eta\eta}^{-1}X_N}{\sigma\sqrt{n}}$ can be dropped because $X_D = \widetilde{X}_D + b_{D2} + O_P(n^{-\mu})$ and $X_N = \widetilde{X}_N + O_P(n^{-\mu})$. Therefore,

$$\begin{aligned} T^* &= \sigma^{-1}c^T\overline{\Sigma}_{\eta\eta}^{-1}\left\{\widetilde{X}_N + b_N + Q_N + L_{N1} + L_{N2} + [L_{N3} - \frac{b_{D2}^T\overline{\Sigma}_{\eta\eta}^{-1}\widetilde{X}_N}{\sqrt{n}}] - \frac{\widetilde{X}_D^T\overline{\Sigma}_{\eta\eta}^{-1}\widetilde{X}_N}{\sqrt{n}}\right\} + o_P(n^{-2\mu}) \\ &\equiv T^{**} + o_P(n^{-2\mu}), \end{aligned} \tag{18}$$

From Lemma 7, $L_{N1}, L_{N2} = o_P(h^q)$, while both are uncorrelated with $\widetilde{X}_N$, since $E[\epsilon_i\eta_i\eta_i^T] = 0 = E[\epsilon_i^2\eta_i]$, by A2. Therefore, since $h = O(n^{-2/(4q+P)})$, neither L_{N1} nor L_{N2} contribute to the second moments of T^* to $O(n^{-2\mu})$. From Lemmas 6 and 7, $b_{D2} = O(n^{-1/2}h^{-P})$ and $L_{N3} = O_P(n^{-1}h^{-P})$. However, although L_{N3} and $b_{D2}^T\overline{\Sigma}_{\eta\eta}^{-1}\widetilde{X}_N$ are both individually correlated with $\widetilde{X}_N$, the linear combination $L_{N3} - \frac{b_{D2}^T\overline{\Sigma}_{\eta\eta}^{-1}\widetilde{X}_N}{\sqrt{n}}$ is not. Furthermore, the covariance between $\frac{\widetilde{X}_D^T\overline{\Sigma}_{\eta\eta}^{-1}\widetilde{X}_N}{\sqrt{n}}$ and $\widetilde{X}_N$ is $O(n^{-1})$. Therefore,

$$E[T^*] = \sigma^{-1}c^T\overline{\Sigma}_{\eta\eta}^{-1}b_N + o(n^{-\mu})\ ;\ Var[T^*] = Var[\widetilde{X}_N] + Var[\sigma^{-1}c^T\overline{\Sigma}_{\eta\eta}^{-1}Q_N] + o(n^{-2\mu}),$$

where b_N is approximated in Lemma 5, while $Var[Q_N] = O(n^{-1}h^{-P})$ is given in Lemma 8.

PROOF OF THEOREM 2

By using that $X_N = \widetilde{X}_N + O_P(n^{-\mu})$, $X_D = \widetilde{X}_D + b_{D2} + O_P(n^{-\mu})$, $X_V = \widetilde{X}_V + b_{V2} + O_P(n^{-\mu})$, and $\sqrt{n}(\widehat{\beta} - \beta) = \Sigma_{\eta\eta}^{-1}\widetilde{X}_N + O_P(n^{-\mu})$, we can drop many cross terms, and find

$$S^* = \frac{[X_V - \frac{1}{\sqrt{n}}\widetilde{X}_N^T\Sigma_{\eta\eta}^{-}\widetilde{X}_N]}{2\sigma_\epsilon^2} - \frac{c^T\Sigma_{\eta\eta}^{-1}X_D\Sigma_{\eta\eta}^{-1}c}{2c^T\Sigma_{\eta\eta}^{-1}c} + \frac{c^T\Sigma_{\eta\eta}^{-1}\{\widetilde{X}_D + b_{D2}\}\Sigma_{\eta\eta}^{-1}\{\widetilde{X}_D + b_{D2}\}\Sigma_{\eta\eta}^{-1}c}{2c^T\Sigma_{\eta\eta}^{-1}c\sqrt{n}} - \frac{[\widetilde{X}_V + b_{V2}]^2}{8\sigma_\epsilon^4\sqrt{n}} - \frac{[c^T\Sigma_{\eta\eta}^{-1}\{\widetilde{X}_D + b_{D2}\}\Sigma_{\eta\eta}^{-1}c]}{8[c^T\Sigma_{\eta\eta}^{-1}c]^2\sqrt{n}} - \frac{c^T\Sigma_{\eta\eta}^{-1}\{\widetilde{X}_D + b_{D2}\}\Sigma_{\eta\eta}^{-1}c[\widetilde{X}_V + b_{V2}]}{4\sigma^2\sqrt{n}} + o_P(n^{-2\mu}), \qquad (19)$$

Furthermore, terms involving $n^{-1/2}b_{D2}^2$, $n^{-1/2}b_{V2}^2$, or $n^{-1/2}b_{D2}b_{V2}$ are $o_P(n^{-2\mu})$, since $\sqrt{n}h^P \to \infty$. The covariances between X_V, X_D and: $n^{-1/2}\widetilde{X}_N^T\Sigma_{\eta\eta}^{-1}\widetilde{X}_N$, $n^{-1/2}c^T\Sigma_{\eta\eta}^{-1}\widetilde{X}_D\Sigma_{\eta\eta}^{-1}\widetilde{X}_D\Sigma_{\eta\eta}^{-1}c$, $n^{-1/2}[c^T\Sigma_{\eta\eta}^{-1}\widetilde{X}_D\Sigma_{\eta\eta}^{-1}c]^2$, $n^{-1/2}c^T\Sigma_{\eta\eta}^{-1}\widetilde{X}_D\Sigma_{\eta\eta}^{-1}c\widetilde{X}_V$, and $n^{-1/2}\widetilde{X}_V^2$, are $o(n^{-2\mu})$. Finally, we substitute $b_{V2} = \sqrt{n}\pi_n\sigma_\epsilon^2$ and $b_{D2} = \sqrt{n}\pi_n\Sigma_{\eta\eta}$ into (A.**??**), where $\pi_n = \sum\sum_{j \neq i} n^{-1}w_{ij}^2 = O(n^{-1}h^{-P})$ by Lemma 3, and obtain several cancellations. In conclusion, we have to calculate the second moments of

$$S^{**} = \frac{X_V}{2\sigma_\epsilon^2} - \frac{c^T\Sigma_{\eta\eta}^{-1}X_D\Sigma_{\eta\eta}^{-1}c}{2c^T\Sigma_{\eta\eta}^{-1}c} + \pi_n\left\{\frac{c^T\Sigma_{\eta\eta}^{-1}\widetilde{X}_D\Sigma_{\eta\eta}^{-1}c}{c^T\Sigma_{\eta\eta}^{-1}c} - \frac{\widetilde{X}_V}{4\sigma_\epsilon^2}\right\},$$

We first calculate the mean of S^{**}. This is

$$\frac{b_V}{2\sigma_\epsilon^2} - \frac{c^T\Sigma_{\eta\eta}^{-1}b_D\Sigma_{\eta\eta}^{-1}c_1}{2c^T\Sigma_{\eta\eta}^{-1}c} + o(n^{-\mu}) = \frac{b_{V1}}{2\sigma_\epsilon^2} - \frac{c^T\Sigma_{\eta\eta}^{-1}b_{D1}\Sigma_{\eta\eta}^{-1}c}{2c^T\Sigma_{\eta\eta}^{-1}c} + o(n^{-\mu}),$$

since $\frac{b_{V2}}{2\sigma_\epsilon^2} - \frac{c^T\Sigma_{\eta\eta}^{-1}b_{D2}\Sigma_{\eta\eta}^{-1}c}{2c^T\Sigma_{\eta\eta}^{-1}c} = 0$. Therefore, the bias is as stated.

We now calculate the variance. Firstly, we rearrange S^{**} so that

$$S^{**} = \frac{X_V - \pi_n\widetilde{X}_V}{2\sigma_\epsilon^2} - \frac{c^T\Sigma_{\eta\eta}^{-1}X_D\Sigma_{\eta\eta}^{-1}c - \pi_n c^T\Sigma_{\eta\eta}^{-1}\widetilde{X}_D\Sigma_{\eta\eta}^{-1}c}{2c^T\Sigma_{\eta\eta}^{-1}c}.$$

Further simplification results because: $Cov[L_{V2} - \pi_n \widetilde{X}_V, \widetilde{X}_V] = 0$ and $Cov[L_{D2} - \pi_n \widetilde{X}_D, \widetilde{X}_D] = 0$ while if the third moment array of u_i is zero, $Cov[L_{V1}, \widetilde{X}_V] = 0 = Cov[L_{D1}, \widetilde{X}_D]$. Therefore, the variance of S^{**} is

$$\frac{Var[\widetilde{X}_V] + Var[Q_V]}{4\sigma_\epsilon^4} + \frac{Var[c^T \Sigma_{\eta\eta}^{-1} \widetilde{X}_D \Sigma_{\eta\eta}^{-1} c] + Var[c^T \Sigma_{\eta\eta}^{-1} Q_D \Sigma_{\eta\eta}^{-1} c]}{4[c^T \Sigma_{\eta\eta}^{-1} c]^2},$$

and the result follows by Lemma 8.

PROOF OF THEOREM 3

From (A.**??**), $S = \frac{\widetilde{X}_V}{2\sigma_\epsilon^2} - \frac{c^T \Sigma_{\eta\eta}^{-1} \widetilde{X}_D \Sigma_{\eta\eta}^{-1} c}{2c^T \Sigma_{\eta\eta}^{-1} c} + O_P(n^{-\mu})$. Therefore,

$$W^* = T^{**} - \frac{c^T \overline{\Sigma}_{\eta\eta}^{-1} \widetilde{X}_N}{\sigma\sqrt{n}} \left\{ \frac{\widetilde{X}_V}{2\sigma_\epsilon^2} - \frac{c^T \Sigma_{\eta\eta}^{-1} \widetilde{X}_D \Sigma_{\eta\eta}^{-1} c}{2c^T \Sigma_{\eta\eta}^{-1} c} \right\} + o_P(n^{-2\mu}).$$

The covariance of $n^{-1/2} \widetilde{X}_N \left\{ \frac{\widetilde{X}_V}{2\sigma_\epsilon^2} - \frac{c^T \Sigma_{\eta\eta}^{-1} \widetilde{X}_D \Sigma_{\eta\eta}^{-1} c}{2c^T \Sigma_{\eta\eta}^{-1} c} \right\}$ with $\widetilde{X}_N$ is $O(n^{-1})$, so that the second moments of W^* agree with those of T^{**} to the required order.

PROOF OF THEOREM 4

We have to show that

4.1 $\Pr[|R_T(\widetilde{h}_0)| > \frac{\chi_{18} n\sqrt{n}}{n^{2\mu} \log n}] = o(n^{-2\mu})$

4.2 $\Pr[|\frac{\widetilde{h}_0 - h_0}{h_0} h_0 \frac{\partial T^*}{\partial h}(\widetilde{h}_0^*)| > \frac{\chi_{18}}{n^{2\mu} \log n}] = o(n^{-2\mu})$.

We first establish **4.1**. Let $\mathcal{N}_{\gamma_0} = [\gamma_0 - \lambda n^{-\phi}, \gamma_0 + \lambda n^{-\phi}]$, and let $\mathcal{C}_1, .., \mathcal{C}_Q$ be sets of diameter $O(n^{-\varphi})$ that cover $\mathcal{N}_{\gamma_0}$. Each $\mathcal{C}_j$ contains only a finite number of points $d_j \leq d$ for some $d < \infty$, while $Q = O(n^{(\varphi - \phi)})$. Let $A = \left\{ |R_T(\widetilde{h}_0)| > \frac{\chi_{18} n\sqrt{n}}{n^{2\mu} \log n} \right\}$ and

$B = \{\widetilde{\gamma}_0 \in \mathcal{N}_{\gamma_0}\}$. Then $\Pr[B^c] = o(n^{-2\mu})$ by assumption. Letting $B_j = \{\widetilde{\gamma}_0 \in \mathcal{C}_j\}$, we have $\Pr[A \cap B] \leq \sum_{j=1}^{Q} \Pr[A \cap B_j]$, while

$$A \cap B_j \subset \overline{A}_j = \left\{ \underset{\mathcal{C}_j}{Max}\ |X_N| \left\| X_D \right\|^3 > \frac{\chi_{19} n\sqrt{n}}{n^{2\mu} \log n} \right\}.$$

Since $\mathcal{C}_j$ is a finite set, we can apply the Bonferroni inequality to obtain that for some χ_{20},

$$\Pr[A \cap B] \leq \chi_{20} n^{(\varphi-\phi)} [\log n]^{\rho} \frac{E[|X_N|^{\rho} \left\| X_D \right\|^{3\rho}]}{n^{(\frac{3}{2}-2\mu)\rho}} = o(n^{-2\mu}),$$

provided $(3-4\mu)\rho > 4\mu + 2(\varphi - \phi)$. Applying (A.**??**), **4.1** follows.

We now establish **4.2**. This probability is bounded by

$$\chi_{21} n^{(\varphi-\phi)} \Pr[|n^{(\mu-\pi)} \frac{\partial T^*}{\partial h}(h_k)| > \frac{\chi_{22} n^{(\varphi-\mu)}}{\log n}] + o(n^{-2\mu}),$$

for some h_k, such that $h_k n^{\pi} \equiv \gamma_k \in \mathcal{C}_k$. Employing

$$|\frac{\partial T^*}{\partial h}| \leq \chi_{24} \left\{ |\frac{\partial X_N}{\partial h}| + \frac{\left\| \frac{\partial X_D}{\partial h} \right\| |X_N|}{\sqrt{n}} + .. + \frac{\left\| \frac{\partial X_D}{\partial h} \right\| \left\| X_D \right\| |X_N|}{n} \right\},$$

and Markov's inequality, we have to establish

4.2.1 $Sup_n E[|n^{(\mu-\pi)} \frac{\partial X_N}{\partial h}|^{\rho_1^*}] < \infty$, $(\varphi-\mu)\rho_1^* > 2\mu + \varphi - \phi$

4.2.2 $Sup_n E[\left\| n^{(\mu^*-\pi)} \frac{\partial X_D}{\partial h} \right\|^{\rho_2^*} |X_N|^{\rho_2^*}] < \infty$, $(\varphi - \mu + \frac{1}{2} - [\mu^* - \mu])\rho_2^* > 2\mu + \varphi - \phi$

4.2.3 $Sup_n E[\left\| n^{(\mu^*-\pi)} \frac{\partial X_D}{\partial h} \right\|^{\rho_3^*} \left\| X_D \right\|^{\rho_3^*} |X_N|^{\rho_3^*}] < \infty$, $(\varphi-\mu+1-[\mu^*-\mu])\rho_3^* > 2\mu+\varphi-\phi$,

all quantities evaluated at a bandwidth h_k. By Cauchy-Schwarz,

$$E[\left\| n^{(\mu^*-\pi)} \frac{\partial X_D}{\partial h} \right\|^{\rho_2^*} |X_N|^{\rho_2^*}] \leq E^{1/2}[\left\| n^{(\mu^*-\pi)} \frac{\partial X_D}{\partial h} \right\|^{2\rho_2^*}] E^{1/2}[|X_N|^{2\rho_2^*}]$$
$$E[\left\| n^{(\mu^*-\pi)} \frac{\partial X_D}{\partial h} \right\|^{\rho_3^*} \left\| X_D \right\|^{\rho_3^*} |X_N|^{\rho_3^*}] \leq E^{1/2}[\left\| n^{(\mu^*-\pi)} \frac{\partial X_D}{\partial h} \right\|^{2\rho_3^*}] E^{1/4}[\left\| X_D \right\|^{4\rho_3^*}] E^{1/4}[|X_N|^{4\rho_3^*}]$$

Therefore, provided M_2 moments exist, the result follows by application of Lemma 10.

PROOF OF COROLLARY

The local polynomial estimators $\widehat{\mathcal{D}_k^q\theta}(Z_i)$ and $\widehat{\mathcal{D}_k^q G}_2(Z_i)$ are both linear in y with weights $\left\{w_{ij}^q\right\}$, and can be decomposed as

$$\widehat{\mathcal{D}_k^q\theta}(Z_i) = \mathcal{D}_k^q\theta(Z_i) + B_{\theta i}^q + V_{\theta i}^q \ ; \ \widehat{\mathcal{D}_k^q G}_2(Z_i) = \mathcal{D}_k^q G_2(Z_i) + B_{\eta i}^q + V_{\eta i}^q,$$

where $V_{di}^q = \sum_j w_{ij}^q \eta_{di} = O_P(n^{-1/2}h^{*-(P+q)/2})$ and $B_{di}^q = O(h^{*(q^*-q)})$ from Lemma 1. Furthermore,

$$\begin{aligned}
\widehat{\zeta_q^{\eta\theta}} - \zeta_q^{\eta\theta} &= n^{-1}\textstyle\sum_{i=1}^n \left\{B_{\theta i}^q B_{\eta i}^q + \mathcal{D}_k^q\theta(Z_i)B_{\eta i}^q + \mathcal{D}_k^q G_2(Z_i)B_{\theta i}^q\right\} \\
&+ n^{-1}\textstyle\sum_{i=1}^n \left\{\mathcal{D}_k^q\theta(Z_i)V_{\eta i}^q + \mathcal{D}_k^q G_2(Z_i)V_{\theta i}^q\right\} \\
&+ n^{-1}\textstyle\sum_{i=1}^n \left\{B_{\theta i}^q V_{\eta i}^q + B_{\eta i}^q V_{\theta i}^q\right\} \\
&+ n^{-1}\textstyle\sum_{i=1}^n V_{\theta i}^q V_{\eta i}^q.
\end{aligned}$$

The first row is deterministic and $O(h^{*(q^*-q)})$. The second is $O_P(n^{-1/2})$ by interchanging summations. The third is $O_P(n^{-1/2}h^{*(q^*-q)})$, while the last is $O_P(n^{-1}h^{*-(P+q)/2})$. For $\sqrt{n}$ consistency of $\widehat{\zeta_q^{\eta\theta}}$ it is necessary that $nh^{*2(q^*-q)} \to 0$ and $nh^{*(P+q)} \to \infty$.

We apply the proof technique of Theorem 1. Let $B = \{|\gamma'(\psi_0^*)| < \chi_{25}\}$, then

$$\{|\widehat{\zeta_q^{\eta\theta}} - \zeta_q^{\eta\theta}| < \chi_{26}\} \cup \{\left\|\widehat{\Sigma} - \Sigma\right\| < \chi_{26}\} \subset B$$

for some χ_{26}, by the continuity of γ_0'. Using a version of Lemma 9 for $\widehat{\zeta_q^{\eta\theta}} - \zeta_q^{\eta\theta}$, we obtain $\Pr[B^c] = o(n^{-2\mu})$. We apply (A.**??**), and the result follows.

REFERENCES

[1] AHN, H., AND J.L. POWELL (1993): "Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics*, 58, 3-30.

[2] ANDREWS, D.W.K., (1989a): "Semiparametric Econometric Models: I Estimation." Cowles Foundation Discussion paper no. 908.

[3] ANDREWS, D.W.K., (1989b): "Semiparametric Econometric Models: II Testing," Cowles Foundation Discussion paper no. 910.

[4] ANDREWS, D.W.K., (1991a): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models." *Econometrica*, 59, 307-346.

[5] ANDREWS, D.W.K., (1991b): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica*, 59, 817-858.

[6] BANKS, J., R. BLUNDELL, AND A. LEWBEL, (1992): "Quadratic Engel Curves, Welfare Measurement and Consumer Demand." IFS Working Paper 92-14.

[7] BARTLETT, M.S., (1963): "Statistical Estimation of Density Functions," *Sankhya Ser A*, *25, 145-154.*

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

[9] BICKEL, P. J., C. A. J. KLAASSEN, Y. RITOV, J. A. WELLNER (1993): Efficient and Adaptive Inference in Semiparametric Models, Forthcoming Monograph. Baltimore: Johns Hopkins University Press.

[10] BIERENS, H.J., (1987): "Kernel Estimators of Regression Functions." in *Advances in Econometrics: Fifth World Congress,* Vol 1, ed. by T.F. Bewley. Cambridge University Press.

[11] BIERENS, H.J., AND H.A. POTT-BUTER, (1990): "Specification of Household Engel Curves by Nonparametric Regression." *Econometric Reviews*, 9, 123-184.

[12] CARROLL, R.J., AND W.HÄRDLE, (1989): "Second Order Effects in Semiparametric Weighted Least Squares Regression." *Statistics,* 2, 179-186.

[13] CAVANAGH, C.J. (1989): "The Cost of Adapting for Heteroscedasticity in Linear Regression Models." Unpublished Manuscript, Harvard University.

[14] CHEN, H. (1988): "Convergence Rates for Parametric Components in a Partly Linear Model," *Annals of Statistics*, 16, 136-146.

[15] CHESHER, A., AND I. JEWITT, (1987): "The Bias of a Heteroskedasticity Consistent Covariance Matrix." *Econometrica,* 55, 1217-1222.

[16] CHESHER, A., AND R. SPADY (1991): "Asymptotic Expansions of the Information Matrix Test Statistic." *Econometrica,* 59, 787-816.

[17] DELGADO, M. AND T. STENGOS (1994): "Semiparametric Specification Testing of Non-nested Econometric models." *Review of Economic Studies,* 61, 291-303.

[18] DELGADO, M. (1992): "Semiparametric Generalised Least Squares in the Multivariate Nonlinear Regression Model." *Econometric Theory*, 8, 203-222.

[19] ENGLE, R. F., C. W. J. GRANGER, J. RICE, AND A. WEISS (1986): "Semiparametric Estimates of the Relationship Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310-320.

[20] FAN, J. (1992): "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association*, 87, 998-1004.

[21] FAN, J. and I.GIJBELS (1992): "Variable Bandwidth and Local Linear Regression Smoothers." *Annals of Statistics*, Forthcoming .

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

[23] HALL, P., (1983): "Large Sample Optimality of Least Squares Cross-validation in Density Estimation." *Annals of Statistics*, 11, 1156-1174.

[24] HÄRDLE, W., P.HALL, AND H. ICHIMURA (1992): "Optimal Smoothing in Single Index Models." CORE Discussion Paper no. 9107.

[25] HÄRDLE, W., J. HART, J. S. MARRON, AND A. B. TSYBAKOV (1992): "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association,* 87, 218-226.

[26] HÄRDLE, W., AND O. LINTON (1994): "Applied nonparametric methods," *The Handbook of Econometrics*, Vol. IV*,* eds D.F. McFadden and R.F. Engle III. North Holland.

[27] HÄRDLE, W., AND T. M. STOKER (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986-995.

[28] HÄRDLE, W. (1990): *Applied Nonparametric Regression.* Cambridge University Press

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

[30] HSIEH, D.A., AND C.F. MANSKI (1987): "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression." *Annals of Statistics*, 15, 541-551.

[31] JONES, M.C., J.S. MARRON, AND S.J. SHEATHER (1992): "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," Australian Graduate School of Management, University of NSW, Working Paper 92-014.

[32] LEE, A.J., (1990): *U-Statistics: Theory and Practice.* Marcel Dekker.

[33] LEE, L-F, M.R. ROSENZWEIG, AND M.M. PITT, (1992) "The Effects of Improved Nutrition, Sanitation, and Water Purity on Child Health in High-Mortality Populations," Unpublished manuscript.

[34] LINDSAY, B.G. (1989): "On the Determinants of Moment Matrices." *Annals of Statistics*, 17, 711-721.

[35] LINTON, O.B. (1991): *Edgeworth Approximation in Semiparametric Regression Models* PhD thesis, Department of Economics UC Berkeley.

[36] LINTON, O.B. (1992): "Edgeworth Approximation for Generalised Minimum Contrast Estimators in Semiparametric Regression Models." Nuffield College Discussion paper no 73.

[37] MARRON, J.S. AND M.P.WAND (1992): "Exact Mean Integrated Squared Error." *Annals of Statistics*, 20, 712-736.

[38] MIKOSCH, T. (1991): "Functional Limit Theorems for Random Quadratic Forms." *Stochastic Processes and their Applications,* 37, 81-98.

[39] MÜLLER, H. G. (1988): *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics, Vol. 46. Heidelberg/New York: Springer-Verlag.

[40] NAGAR, A.L., (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica,* 27, 575-595.

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

[42] NEWEY, W.K. (1990b): "Efficient Instrumental Variables Estimation of Nonlinear Models." *Econometrica,* 58, 809-837.

[43] NEWEY, W.K., J.L. POWELL, AND J.R. WALKER, (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results." *American Economic Review,* 80, 324-328.

[44] PARK, B.U., AND J.S. MARRON, (1990): "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85, 66-72.

[45] PFANZAGL, J., (1980): "Asymptotic Expansions in Parametric Statistical Theory," in: *Developments in Statistics*, v3, (P.R.Krishnaiah ed) Academic Press.

[46] PHILLIPS, P.C.B AND J.Y. PARK (1988): "On the Formulation of Wald Tests for Nonlinear Restrictions." *Econometrica,* 56, 1065-1083.

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

[48] RICE, J. (1986): "Convergence Rates for Partially Splined Models." *Statistics and Probability Letters*, 4, 203-208.

[49] ROBINSON, P. M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown form." *Econometrica*, 56, 875-891.

[50] ROBINSON, P. M. (1988a): "The Stochastic Difference between Econometric Statistics," *Econometrica*, 56, 531-548.

[51] ROBINSON, P. M. (1988b): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931-954.

[52] ROBINSON, P. M. (1988c): "Semiparametric Econometrics: A Survey," *Journal of Applied Econometrics*, 3, 35-51.

[53] ROBINSON, P. M. (1989): "Hypothesis Testing in Semiparametric and Non-parametric Models for Economic Time Series." *Review of Economic Studies*, 56, 511-534.

[54] ROBINSON, P. M. (1991a): "Automatic Frequency Domain Inference on Semi-parametric and Nonparametric Models." *Econometrica*, 59, 1329-1364.

[55] ROBINSON, P. M. (1991b): "Best Nonlinear Three-stage Least Squares Estimation of certain Econometric Models." *Econometrica*, 59, 755-786.

[56] ROTHENBERG, T., (1984a): "Approximating the Distributions of Econometric Estimators and Test Statistics." Ch.14 in: *Handbook of Econometrics*, Vol 2, ed. by Z. Griliches and M.Intriligator. North Holland.

[57] ROTHENBERG, T., (1984b): "Approximate Normality of Generalized Least Squares Estimates." *Econometrica*, 52, 811-825.

[58] ROTHENBERG, T., (1984c): "Hypothesis Testing in Linear Models when the Error Covariance Matrix is Nonscalar." *Econometrica*, 52, 827-842.

[59] ROTHENBERG, T., (1988): "Approximate Power Functions for some Robust Tests of Regression Coefficients." *Econometrica*, 56, 997-1019.

[60] RUPPERT, D. AND M.P. WAND (1992): "Multivariate Weighted Least Squares Regression." Rice University, Department of Statistics Technical Report no 92-4.

[61] SARGAN, J.D., AND W.M.MIKHAIL, (1971): "A General Approximation to the Distribution of Instrumental Variables Estimates." *Econometrica*, 39, 131-169.

[62] SHEATHER, S.J., AND M.C. JONES, (1991): "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society Series B*, 53, 683-690.

[63] SILVERMAN, B. (1986): *Density estimation for statistics and data analysis.* London, Chapman and Hall.

[64] SPECKMAN, P. (1988): "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society Series B*, 50, 413-436.

[65] SRINAVASAN, T.N. (1970): "Approximations to Finite Sample Moments of Estimators whose Exact Sampling Distributions are unknown," *Econometrica,* 38, 533-541.

[66] STOCK, J. H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84, 567-576.

[67] STOCK, J. H. (1991): "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

[68] STOKER, T.M, (1989): "Tests of Average Derivative Constraints," *Review of Economic Studies,* 56, 535-552.

[69] STOKER, T.M, (1993): "Smoothing Bias in Derivative Estimation." *Journal of the American Statistical Association,* Forthcoming.

[70] STONE, C.J. (1977): "Consistent Nonparametric Regression," *Annals of Statistics,* 5, 595-620.