Econometrica, Vol. 57, No. 6 (November, 1989), 1403–1430

SEMIPARAMETRIC ESTIMATION OF INDEX COEFFICIENTS

BY JAMES L. POWELL, JAMES H. STOCK, AND THOMAS M. STOKER[1]

This paper gives a solution to the problem of estimating coefficients of index models, through the estimation of the density-weighted average derivative of a general regression function. We show how a normalized version of the density-weighted average derivatives can be estimated by certain linear instrumental variables coefficients. Both of the estimators are computationally simple, root-N-consistent and asymptotically normal; their statistical properties do not rely on functional form assumptions on the regression function or the distribution of the data. The estimators, based on sample analogues of the product moment representation of the average derivative, are constructed using nonparametric kernel estimators of the density of the regressors. Asymptotic normality is established using extensions of classical U-statistic theorems, and asymptotic bias is reduced through use of a higher-order kernel. Consistent estimators of the asymptotic variance-covariance matrices of the estimators are given, and a limited Monte Carlo simulation is used to study the practical performance of the procedures.

KEYWORDS: Index restrictions, semiparametric estimation, density-weighted average derivatives, kernel estimation, root-N-consistency, limited dependent variables.

1. INTRODUCTION

A PROBLEM OF SUBSTANTIAL PRACTICAL INTEREST concerns the estimation of coefficients in index models. In this paper we give a solution to this problem through the estimation of the density-weighted average derivative of a general regression function. To fix ideas, let y denote a dependent variable and x a vector of independent variables, where the true regression function is $E(y|x) = g(x)$, and x is distributed with density $f(x)$. The density-weighted average derivative vector is defined as[2]

$$(1.1) \qquad \delta \equiv E\left(f(x) \frac{\partial g}{\partial x} \right).$$

Our approach is nonparametric: we propose an estimator of δ whose properties can be derived under weak restrictions on the joint distribution of (y, x). In particular, no functional form assumptions are applied to $g(x)$ or $f(x)$.

Weighted average derivatives are of practical interest because they are proportional to coefficients in index models. Suppose that the model explaining y implies that $g(x)$ can be written in the "single index" form

$$(1.2) \qquad g(x) = G(x'\beta)$$

for some univariate function $G(.)$. For instance, this form arises in various

[1] The authors gratefully acknowledge funding for this research from National Science Foundation grants. The authors wish to thank L. Hansen, J. Hausman, J. Heckman, C. Manski, S. Marron, D. McFadden, P. Robinson, C. Sims, and the two anonymous referees for helpful comments, and M. Kennet for computational assistance. An earlier version of this paper, entitled "Semiparametric Estimation of Weighted Average Derivatives," was presented at the 1986 North American Summer Meetings of the Econometric Society.

[2] Note that we have not required the weighting function to have expectation unity—later we introduce this normalization, so that the estimators are properly averaged derivatives.

models of limited dependent variables (see Ruud (1986) and Stoker (1986) for examples), where the structure of $G(.)$ is determined by an underlying distribution of unobserved stochastic terms. For our purposes, two features of such index models are relevant. First, since any nonzero rescaling of the coefficients β can be absorbed into the definition of $G(.)$, of interest are estimators of β up to scale. Second, the single index form is manifested as a restriction on the derivatives of the regression function $g(x)$, as

$$(1.3)\qquad \frac{\partial g}{\partial x}=\frac{\partial G(x'\beta)}{\partial x}=\left(\frac{dG}{d(x'\beta)}\right)\beta$$

so that $\partial g/\partial x$ is proportional to β for each value of x. Thus any weighted average of the derivatives $\partial g/\partial x$ will also be proportional to β. Let $\omega(x)$ be a weighting function; then

$$\delta_\omega\equiv E[\omega(x)\,\partial g/\partial x]=E[\omega(x)\,dG/d(x'\beta)]\beta\equiv\gamma_\omega\beta$$

is proportional to β, provided $\gamma_\omega\neq 0$. Given this flexibility, we are free to set $\omega(x)=f(x)$ (the density of x) and focus on estimating the density-weighted average derivative δ of (1.1), where for the index model (1.2) we have

$$(1.4)\qquad \delta\equiv E\left(f(x)\frac{\partial g}{\partial x}\right)=E\left(f(x)\frac{dG}{d(x'\beta)}\right)\beta\equiv\gamma\beta.$$

As seen later, the choice of density-weighting is made because it permits an estimator to be proposed whose properties can be analyzed and understood in a straightforward fashion.

While any scaling normalization permits identification of the index coefficients, a natural choice would be to impose the condition $E[w(x)]=1$ on the weighting function $w(x)$. For the density weighted coefficients δ, this yields the rescaled coefficients $\delta^*=\delta/E[f(x)]$. For example, if $g(x)=\alpha+x'\beta$, then $\delta=E[f(x)]\beta$, but $\delta^*=\beta$. Since the components of δ^* are comparable to linear model coefficients, their values may be more easily interpreted than those of the components of δ.

In this paper we propose an estimator $\hat{\delta}_N$ of the density-weighted average derivative δ, where N is the sample size. We show that $\hat{\delta}_N$ is a $\sqrt{N}$-consistent, asymptotically normal estimator of δ, and give a consistent estimator of its asymptotic variance-covariance matrix. We also propose an estimator $\hat{d}_N$ of δ^*, by a straightforward modification of $\hat{\delta}_N$, and show that $\hat{d}_N$ has analogous distributional properties.

The estimators are based on sample analogues of a product-moment representation of density-weighted average derivatives. The representation involves derivatives of the density of x, which are estimated nonparametrically using the kernel density estimation technique of Parzen (1962) and others (see Prakasa-Rao (1983) for a survey). The estimator $\hat{\delta}_N$ is based on an appropriate average of the estimated density derivatives. The estimator $\hat{d}_N$ is the slope coefficient vector of y regressed on x, where the estimated density derivatives are used as instrumental variables. Each of these estimators is computed directly from the observed data, requiring no computational techniques for maximization or other types of equation solving.

The verification of the statistical properties of the estimators is of theoretical interest, because it involves reconciling the relatively slow convergence properties of nonparametric density estimators with the classical properties of sample averages. The key to establishing $\sqrt{N}$-consistency and asymptotic normality of $\hat{\delta}_N$ is noting that $\hat{\delta}_N$ can be written as a U-statistic: this structure permits proper accounting of the "overlaps" in the density derivative estimators that comprise $\hat{\delta}_N$. The U-statistic structure also motivates a natural estimator of the asymptotic variance of $\hat{\delta}_N$. The statistical properties of $\hat{d}_N$ follow in a straightforward fashion from those of $\hat{\delta}_N$.

We study the practical performance of the estimators via a limited Monte Carlo analysis. The instrumental variables estimator $\hat{d}_N$ performs well in small samples, and displays better operating performance than $\hat{\delta}_N$ for the modelling situations studied.

Sections 2, 3, and 4 give the formal analysis of the estimators and related results, with all proofs not developed in the text removed to Appendix 1. Specifically, Section 2 presents our assumptions and briefly reviews some properties of kernel estimators. Section 3 proposes the estimator $\hat{\delta}_N$ of density-weighted average derivatives, establishes $\sqrt{N}$-consistency and asymptotic normality, and gives the consistent estimator of its asymptotic variance-covariance matrix. Section 4 introduces the instrumental variables estimator $\hat{d}_N$ and discusses its properties. Following the theoretical discussion, Section 5 presents some Monte Carlo evidence on the performance of the estimators, and Section 6 gives some concluding remarks.

2. NOTATION, ASSUMPTIONS, AND TECHNICAL BACKGROUND

2.1. *The Basic Framework and Approach to Estimation*

We consider an empirical problem where y denotes a dependent variable and x a k-vector of independent variables. The data consists of N observations (y_i, x_i'), $i = 1, \ldots, N$, which is assumed to be an i.i.d. random sample from a distribution that is absolutely continuous with respect to a σ-finite measure ν, with (Radon-Nikodym) density $F(y, x)$. The marginal density of x is denoted as $f(x)$, and the regression function of y given x is denoted as $g(x) \equiv E(y|x)$.

The main structural assumptions can now be stated as follows:

ASSUMPTION 1: *The support Ω of f is convex (possibly unbounded) subset of R^k with nonempty interior Ω_0. The underlying measure ν can be written in product form as $\nu = \nu_y \times \nu_x$, where ν_x is Lebesgue measure on R^k.*

ASSUMPTION 2: *The density function f is continuous in the components of x for all $x \in R^k$, so that $f(x) = 0$ for all $x \in \partial\Omega$, where $\partial\Omega$ denotes the boundary of Ω.*

Furthermore, f is continuously differentiable in the components of x for all $x \in \Omega_0$ and g is continuously differentiable in the components of x for all $x \in \bar{\Omega}$, where $\bar{\Omega}$ differs from Ω_0 by a set of measure zero.

ASSUMPTION 3: *The components of the random vector $\partial g/\partial x$ and random matrix $[\partial f/\partial x][y, x']$ have finite second moments. Also, $\partial f/\partial x$ and $\partial(gf)/\partial x$ satisfy the following Lipschitz conditions: For some $m(x)$,*

$$\left\| \frac{\partial f(x+\nu)}{\partial x} - \frac{\partial f(x)}{\partial x} \right\| < m(x)\|\nu\|,$$

$$\left\| \frac{\partial [f(x+\nu)\cdot g(x+\nu)]}{\partial x} - \frac{\partial [f(x)\cdot g(x)]}{\partial x} \right\| < m(x)\|\nu\|,$$

with $E[(1+|y|+\|x\|)m(x)]^2 < \infty$. Finally, $v(x) = E(y^2|x)$ is continuous in x.

Assumption 1 restricts x to be a continuously distributed random variable, where no component of x is functionally determined by other components of x. Continuity of x is useful in this context because of the generality of the dependent variables considered (for example, Manski (1988) points out how continuity of the regressors is useful for identification of binary response models with index restrictions). Assumption 2 is a boundary condition, that allows for unbounded x's (where $\Omega = R^k$ and $\partial\Omega = \varnothing$) and gives the smoothness conditions on f and g. Assumption 3 imposes standard bounded moment and dominance conditions.

Our approach to the estimation of δ of (1.1) is based on a product-moment representation of the density-weighted average derivative (alternative approaches are discussed in Section 6). This representation is based on the following multivariate application of integration by parts:

$$E\left(f(x)\frac{\partial g}{\partial x}\right) = \int \frac{\partial g}{\partial x} f(x)^2\,dx = -2\int g(x)\frac{\partial f}{\partial x} f(x)\,dx = -2E\left(y\frac{\partial f}{\partial x}\right) \tag{2.1}$$

where the boundary terms in the integration by parts formula vanish by Assumption 2. We formalize the result as Lemma 2.1:

LEMMA 2.1: *Given Assumptions* 1–3,

$$\delta = E\left(f(x)\frac{\partial g}{\partial x}\right) = -2E\left(y\frac{\partial f}{\partial x}\right). \tag{2.2}$$

We propose to estimate δ by the sample analogue of (2.2), where $\partial f/\partial x$ is replaced by a consistent nonparametric estimate. Specifically, let $\hat{f}(x)$ be an estimator of $f(x)$, and let $\partial\hat{f}(x)/\partial x$ denote the associated estimator of its derivative. Then an estimator of δ can be formed as the sample product-moment of (2.2), namely $(-2/N)\Sigma[y_i\,\partial\hat{f}(x_i)/\partial x]$. Our specific estimator $\hat{\delta}_N$ of δ uses a kernel estimator of the marginal density $f(x)$. We now review kernel density estimators and some of their properties.

2.2. *Kernel Estimators*: *Notation and Pointwise Convergence Properties*

There are a number of methods for estimating an unknown function nonparametrically; in this paper, we use kernel estimators, which arise from a particular method of local averaging.[3] A kernel estimator of the density $f(x)$ can be written in the form

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{h}\right)^{k} K\left(\frac{x - x_i}{h}\right), \tag{2.3}$$

where the "kernel" $K(.)$ is a weighting function and the "band (or window) width" $h = h_N$ is a smoothing parameter that depends on the sample size N. The contribution to $\hat{f}(x)$ of data points that are close to x is determined by $K(.)$, where "closeness" is determined by the bandwidth h. The asymptotic properties of $\hat{f}(x)$ refer to the limiting properties obtained as the sample size N increases and the bandwidth h declines. We assumed that a kernel $K(.)$ is chosen that obeys Assumption 4:

Assumption 4: *The support Ω_K of $K(u)$ is a convex (possibly unbounded) subset of R^k with nonempty interior, with the origin as an interior point. $K(u)$ is a bounded differentiable function such that $\int K(u)\,du = 1$ and $\int uK(u)\,du = 0$. $K(u) = 0$ for all $u \in \partial\Omega_K$, where $\partial\Omega_K$ denotes the boundary of Ω_K. $K(u)$ is a symmetric function*; *$K(u) = K(-u)$ for all $u \in \Omega_K$.*

We denote the derivative of K as $K'(u) \equiv \partial K/\partial u$. The symmetry of K implies that K' is anti-symmetric: $K'(-u) = -K'(u)$ for all $u \in \Omega_K$.[4]

Detailed studies of the pointwise properties of kernel density estimators can be found in the statistical literature. Derivations of the properties cited below can be found in Silverman (1986); see also Fryer (1977), Tapia and Thompson (1978), Spiegelman and Sacks (1980), Stone (1984), and Bierens (1983) among many others. For our purposes, some of the known pointwise properties on the convergence of $\hat{f}(x)$ to $f(x)$ are of interest for interpreting our results. The bias of $\hat{f}(x)$ is at most $O(h)$, and thus converges to 0 as $N \to \infty$ and $h \to 0$. The variance of $\hat{f}(x)$ is $O(1/Nh^k)$, and therefore converges to 0 if $Nh^k \to \infty$. Consequently, if $Nh^k \to \infty$ and $Nh^{k+2} \to 0$, the mean square error of $\hat{f}(x)$ is $O(1/Nh^k)$. These properties imply that the maximal rate of convergence of $\hat{f}(x)$ to $f(x)$ is $\sqrt{N}h^k$, which is strictly slower than $\sqrt{N}$ since $h \to 0$.

[3] Other methods include nearest neighbor estimation, as studied by Stone (1977) and others. A survey of several methods can be found in Prakasa-Rao (1983). For a review of nonparametric estimators in the context of econometric problems, see McFadden (1985).

[4] While we do not require it, if $K(.)$ is assumed to have bounded support, the Lipschitz condition of Assumption 3 can be weakened to hold only for ν in a neighborhood of 0.

Some intuition for these properties is available from examining (2.3). If h were fixed as $N \to \infty$, $\hat{f}(x)$ would be an ordinary sample average, whose variance was $O(1/N)$ for standard reasons. But unless $f(x)$ were linear nearby x, the bias $E[\hat{f}(x)] - f(x)$ would not vanish. To eliminate the bias, averaging is done over effectively smaller areas via $h \to 0$. This implies that the variance collapses more slowly as $O(1/Nh^k)$, where h^k reflects the effective area over which averaging is performed. To have a more accurate pointwise approximation as N increases, $\hat{f}(x)$ must converge to $f(x)$ at a rate less than $\sqrt{N}$.[5]

The same slow pointwise convergence is displayed by kernel density derivative and kernel regression function estimators. The density derivative estimator associated with $\hat{f}(x)$ is obtained by differentiating (2.3) as

$$(2.4)\qquad \frac{\partial \hat{f}(x)}{\partial x} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{h}\right)^{k+1} K'\left(\frac{x-x_i}{h}\right).$$

By analogous reasoning for the density estimator, the density derivative estimator $\partial\hat{f}/\partial x$ obeys $MSE(\partial\hat{f}/\partial x) = O(1/Nh^{k+1})$ as $h \to 0$, $Nh^{k+1} \to \infty$, and $Nh^{k+3} \to 0$. Moreover, kernel estimators of the regression function $g(x)$ can be defined, which exhibit the same convergence properties as kernel density estimators.

Such slow rates of convergence imply that precise pointwise nonparametric characterizations of density functions, regression functions, and derivatives of such functions will be feasible only for extremely large data sets. These problems are particularly severe for higher dimensional applications (larger k), reflecting a particular embodiment of the "curse of dimensionality" cited by Huber (1985), McFadden (1985) and others.

We have raised these issues to place our results in a particular context. In the next section, we produce a $\sqrt{N}$-consistent and asymptotically normal estimator of the weighted average derivative δ, that is based on averages of kernel density derivative estimators. Consequently, our results give an example of how the slow convergence rates of pointwise estimators can be speeded up when they are averaged to estimate a finite parameter vector, thereby avoiding the curse of dimensionality.

3. THE WEIGHTED AVERAGE DERIVATIVE ESTIMATOR

3.1. *The Estimator and Its Interpretation*

We define $\hat{\delta}_N$, the estimator of the weighted average derivative $\delta = E[f(x)\,\partial g/\partial x]$, as the sample analogue of the product-moment representation (2.2):

$$(3.1)\qquad \hat{\delta}_N = \frac{-2}{N}\sum_{i=1}^{N}\left(\frac{\partial \hat{f}_i(x_i)}{\partial x}\right) y_i,$$

[5] Stone (1980) makes this intuition concrete in his display of optimal pointwise rates of convergence for nonparametric estimators.

where $\hat{f}_i(x)$ is the kernel density estimator

$$\text{(3.2)} \qquad \hat{f}_i(x) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{1}{h}\right)^k K\left(\frac{x - x_j}{h}\right).$$

(Note that $\hat{f}_i(x)$ differs from $\hat{f}(x)$ of (2.4) by omitting x_i in the estimation of the density $f(x)$, which is the most convenient formulation for our technical development.[6]) Thus to compute the ith summand of $\hat{\delta}_N$, the density is estimated as $\hat{f}_i(x)$, the derivative is estimated as $\partial \hat{f}_i(x)/\partial x$, and the summand formed as $(\partial \hat{f}_i(x_i)/\partial x) y_i$.

Provided that $\partial \hat{f}_i(x)/\partial x$ is consistent for $\partial f(x)/\partial x$ (in some sense), $\hat{\delta}_N$ will be a consistent estimator of δ by the law of large numbers. In Section 3.2 we show that $\sqrt{N}[\hat{\delta}_N - E(\hat{\delta}_N)]$ has a limiting normal distribution. In Section 3.3 we indicate how the asymptotic bias is controlled by structuring the kernel $K(.)$, and show that $\sqrt{N}(\hat{\delta}_N - \delta)$ has a limiting normal distribution.

A useful representation of $\hat{\delta}_N$ is obtained by first inserting (3.2) into (3.1):

$$\text{(3.3)} \qquad \hat{\delta}_N = \frac{-2}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{1}{h}\right)^{k+1} K'\left(\frac{x_i - x_j}{h}\right) y_i$$

and then writing $\hat{\delta}_N$ in the standard "U-statistic" form as

$$\text{(3.4)} \qquad \hat{\delta}_N = -\binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{1}{h}\right)^{k+1} K'\left(\frac{x_i - x_j}{h}\right)(y_i - y_j),$$

using the fact that $K'(u) = -K'(-u)$.[7] This information permits a direct analysis of the asymptotic properties of $\hat{\delta}_N$, as will be seen. Note that for any value of the bandwidth h, calculating $\hat{\delta}_N$ only involves a computation of order N^2.

Examination of the components of $\hat{\delta}_N$ yields a natural "slope" interpretation. Let the subscript l denote a particular component of a k-vector, as in $x_i = (x_{1i}, \ldots, x_{li}, \ldots, x_{ki})'$. The lth component of $\hat{\delta}_N$ can be written via (3.4) as

$$\text{(3.5)} \qquad \hat{\delta}_{lN} = -\binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{1}{h}\right)^{k} w_l\left(\frac{x_i - x_j}{h}\right)\left(\frac{y_i - y_j}{x_{il} - x_{jl}}\right)$$

[6] By leaving x_i out of the averaging in $\hat{f}_i$, the variation of $\hat{f}_i(x_i)$ around $f(x_i)$ becomes independent of x_i. This conveys the technical advantages of "sample splitting" (c.f. Bickel (1982)) to our analysis, without a significant cost in data usage.

Note in addition that omitting x_i has only a minor impact on the value of the estimator. In particular, utilizing (2.4) would only alter (3.3) to add zero terms (as Assumption 4 implies $K'(0) = 0$) and change the leading factor from $2/[N(N-1)]$ to $2/N^2$.

[7] This represents the main use of the symmetry of $K(.)$. One could dispense with the symmetry of K by using a "symmetrized" representation of the kernel, as described in Serfling (1980, p. 172).

where $w_l(u) = -u_l \partial K/\partial u_l$ is a weighting function. An application of integration by parts gives $\int w_l(u)\,du = \int K(u)\,du = 1$. Equation (3.5) shows that $\hat{\delta}_{lN}$ is a weighted combination of the slopes $(y_i - y_j)/(x_{il} - x_{jl})$, $i, j = 1, \ldots, N$, with low weight given to observations with $\|x_i - x_j\|$ large.[8] Consequently, the estimator $\hat{\delta}_N$ embodies the intuitive feature of combining the slope (derivative) estimates $(y_i - y_j)/(x_{il} - x_{jl})$ for all i, j, l.

3.2. *Asymptotic Normality*

In this section we establish that $\sqrt{N}[\hat{\delta}_N - E(\hat{\delta}_N)]$ has a limiting normal distribution with mean 0 and variance-covariance matrix Σ_δ, and obtain an explicit formula for Σ_δ. As indicated above, the limiting distribution of $\hat{\delta}_N$ involves a faster rate of convergence than the separate density derivative estimators. This follows from the fact that each data point is used in the estimation of several density derivative values. These overlaps in the local averaging of the density derivative estimates are reflected directly in the U-statistic representation (3.4), and correspondingly the results of this section follow from a general result on the asymptotic behavior of U-statistics. We first prove the general result as Lemma 3.1, which extends the classical theorems of Hoeffding (1948) (see Serfling (1980) for a recent reference). We then apply this result to the average derivative estimator $\hat{\delta}_N$.

Begin by considering a general "second-order" U-statistic of the form

$$(3.6)\qquad U_N \equiv \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_N(z_i, z_j)$$

where $\{z_i, i = 1, \ldots, N\}$ is an i.i.d. random sample and p_N is a k-dimensional symmetric kernel; i.e. $p_n(z_i, z_j) = p_N(z_j, z_i)$. Also define

$$(3.7)\qquad r_N(z_i) = E[p_N(z_i, z_j) | z_i],$$

$$(3.8)\qquad \theta_N = E[r_N(z_i)] = E[p_N(z_i, z_j)],$$

$$(3.9)\qquad \hat{U}_N = \theta_N + \frac{2}{N} \sum_{i=1}^{N} [r_N(z_i) - \theta_N],$$

where we assume that θ_N exists. $\hat{U}_N$ is called the "projection" of the statistic U_N (Hoeffding (1948)).

Our first result is Lemma 3.1, which establishes the asymptotic equivalence of U_N and $\hat{U}_N$:[9]

LEMMA 3.1: *If* $E[\|p_N(z_i, z_j)\|^2] = o(N)$, *then* $\sqrt{N}(U_N - \hat{U}_N) = o_p(1)$.

[8] In general, the weight w_l is nonmonotonic in the difference $\|x_i - x_j\|$, where $\|.\|$ denotes the standard Euclidean norm on R^k, namely $\|u\|^2 = \Sigma u_l^2$. For instance, if $k = 1$ and $K(.)$ is a standard normal density function, the weighting function $w(.)$ is bimodal, with modes at $-2h$ and $2h$.

[9] This lemma is an extension of the results in Serfling (1980, p. 186–188), to the case where $p_N(z_i, z_j)$ varies with N.

Because the projection $\hat{U}_N$ is an average of independent random variables, we could establish directly the asymptotic normality of U_N, assuming the regularity conditions of a central limit theorem.

Consequently, Lemma 3.1 provides sufficient technical machinery to establish the asymptotic normality of the average kernel estimator δ_N centered about $E(\hat{\delta}_N)$. Let $z_i \equiv (y_i, x_i')'$, and rewrite (3.4) as

$$(3.10) \qquad \hat{\delta}_N = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_N(z_i, z_j)$$

with

$$(3.11) \qquad p_N(z_i, z_j) \equiv -\left(\frac{1}{h}\right)^{k+1} K'\left(\frac{x_i - x_j}{h}\right)(y_i - y_j),$$

where p_N varies with N through h. Moreover, define

$$(3.12) \qquad v(x) \equiv E(y^2|x).$$

To apply Lemma 3.1, we require conditions on the bandwidth such that $E[\|p_N(z_i, z_j)\|^2] = o(N)$. We have

$$\begin{aligned}(3.13) \qquad & E\left[\|p_N(z_i, z_j)\|^2\right] \\ &= \int \left(\frac{1}{h}\right)^{2k+2} \left\|K'\left(\frac{x_i - x_j}{h}\right)\right\|^2 \left[v(x_i) + v(x_j) - 2g(x_i)g(x_j)\right] \\ &\quad \times f(x_i)f(x_j)\, dx_i\, dx_j \\ &= \left(\frac{1}{h}\right)^{k+2} \int \|K'(u)\|^2 \left[v(x_i) + v(x_i + hu) - 2g(x_i)g(x_i + hu)\right] \\ &\quad \times f(x_i)f(x_i + hu)\, dx_i\, du \\ &= O(h^{-(k+2)}) = O\left[N(Nh^{k+2})^{-1}\right]\end{aligned}$$

where the second equality uses the change-of-variables from (x_i, x_j) to $(x_i, u = (x_j - x_i)/h)$, with Jacobian h^{-k}, and the third equality uses the continuity of v, g, and f. Consequently, we have $E[\|p_N(z_i, z_j)\|^2] = o(N)$ if and only if $Nh^{k+2} \to \infty$ as $h \to 0$.

Thus under this condition, the asymptotic distribution of $[\hat{\delta}_N - E(\hat{\delta}_N)]$ is the same as that of the sample average $(2/N)\Sigma\{r_N(z_i) - E[r_N(z_i)]\}$. To characterize the distribution, define

$$(3.14) \qquad r(z_i) = f(x_i)\frac{\partial g(x_i)}{\partial x} - [y_i - g(x_i)]\frac{\partial f(x_i)}{\partial x}.$$

For $p_N(z_i, z_j)$ of (3.11), we have

$$
\begin{aligned}
(3.15)\qquad r_N(z_i) &= E\left[p_N(z_i, z_j)|z_i\right]\\
&= -\int\left(\frac{1}{h}\right)^{k+1} K'\left(\frac{x_i - x}{h}\right)(y_i - g(x))f(x)\,dx\\
&= \int\left(\frac{1}{h}\right)K'(u)(y_i - g(x_i + hu))f(x_i + hu)\,du\\
&= \int\frac{\partial(gf)(x_i + hu)}{\partial x}K(u)\,du - y_i\int\frac{\partial f(x_i + hu)}{\partial x}K(u)\,du\\
&= r(z_i) + \int\left(\frac{\partial(gf)(x_i + hu)}{\partial x} - \frac{\partial(gf)(x_i)}{\partial x}\right)K(u)\,du\\
&\quad - y_i\int\left(\frac{\partial f(x_i + hu)}{\partial x} - \frac{\partial f(x_i)}{\partial x}\right)K(u)\,du\\
&\equiv r(z_i) + t_N(z_i).
\end{aligned}
$$

Now since

$$
\begin{aligned}
(3.16)\qquad \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\{r_N(z_i) - E[r_N(z_i)]\} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{r(z_i) - E[r(z_i)]\}\\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{t_N(z_i) - E[t_N(z_i)]\},
\end{aligned}
$$

the limiting distribution of $\sqrt{N}[\hat{\delta}_N - E(\hat{\delta}_N)]$ is the same as that of $(2/\sqrt{N})\Sigma_i\{r(z_i) - [r(z_i)]\}$, provided the last term of (3.16) converges in probability to zero. But, by Assumption 3, this last term has second moment that is bounded above by $4h^2\{E[(1 + |y|)m(x)]^2[\int \|u\| \cdot |K(u)|du]^2\} = O(h^2)$, so it does converge to zero in probability. Application of the Lindeberg-Levy Central Limit Theorem to (3.16) then yields the result:

THEOREM 3.1: *Given Assumptions* 1–4, *if* $h \to 0$ *and* $Nh^{k+2} \to \infty$, *then the average derivative estimator* $\hat{\delta}_N$ *of* (3.1) *is such that* $\sqrt{N}[\hat{\delta}_N - E(\hat{\delta}_N)]$ *has a limiting multivariate normal distribution with mean* 0 *and variance-covariance matrix* Σ_δ, *where*

$$(3.17)\qquad \Sigma_\delta \equiv 4E[r(z_i)r(z_i)'] - 4\delta\delta',$$

is the variance-covariance matrix of $2r(z_i)$.

Note that the condition $Nh^{k+2} \to \infty$ places an upper bound on the rate that the bandwidth h converges to 0. Moreover, note that the asymptotic variance-covariance matrix Σ_δ does not depend on the kernel $K(.)$, and thus does not depend on the weighting used in the local averaging. This is in contrast to the (pointwise) asymptotic variances of the kernel density and density derivative estimators, which do depend on the kernel $K(.)$ (see Silverman (1986) and Prakasa-Rao (1983) among others).

3.3. *Asymptotic Bias*

Unlike the asymptotic variance, the asymptotic bias of $\hat{\delta}_N$ does depend on the kernel function $K(.)$. In this section we analyze the asymptotic bias, showing how it will vanish at rate $\sqrt{N}$ when a certain type of kernel function is used.

We begin by introducing conditions under which the bias can be expanded as a Taylor series in the bandwidth h as

$$(3.18)\qquad E(\hat{\delta}_N) - \delta = b_1 h + b_2 h^2 + \dots + b_{P-1} h^{P-1} + O(h^P)$$

where $P = (k+4)/2$ if k is even and $P = (k+3)/2$ if k is odd. By Young's version of Taylor's Theorem (c.f. Serfling (1980), among others), this representation is possible if the first P derivatives of $E(\hat{\delta}_N)$ with respect to h exist at $h = 0$. Write $E(\hat{\delta}_N)$ as

$$(3.19)\qquad E(\hat{\delta}_N) = -2\int \left(\frac{1}{h}\right)^{k+1} K'\left(\frac{x_1 - x_2}{h}\right) g(x_1) f(x_1) f(x_2)\, dx_1\, dx_2$$

$$= 2\left(\frac{1}{h}\right)\int K'(u) g(x) f(x) f(x+hu)\, dx\, du$$

$$= -2\int K(u) g(x) f(x) \frac{\partial f(x+hu)}{\partial x}\, dx\, du.$$

Expansion of the last integral in (3.19) gives the representation (3.18), with the lth coordinate of b_p given as

$$(3.20)\qquad b_{lp} = \frac{-2}{p!} \sum_{l_1,\dots,l_p=1}^{k} \int u_{l_1} \dots u_{l_p} K(u) g(x) \frac{\partial^{p+1} f(x)}{\partial x_{l_1} \dots \partial x_{l_p} \partial x_l} f(x)\, dx\, du.$$

Thus a sufficient condition for the existence of the expansion (3.18) is the following assumption:

ASSUMPTION 5: *Let $P = (k+4)/2$ if k is even and $P = (k+3)/2$ if k is odd. All partial derivatives of $f(x)$ of order $P+1$ exist. The expectation $E[y(\partial^p f(x)/\partial x_{l_1} \dots \partial x_{l_p})]$ exists for all $p \leqslant P+1$. All moments of $K(u)$ of order P exist.*

The source of asymptotic bias of $\hat{\delta}_N$ lies in the leading $P-1$ terms of the expansion (3.18). To see this, multiply (3.18) by $\sqrt{N}$ as

$$(3.21)\qquad \sqrt{N}\left[E(\hat{\delta}_N) - \delta\right] = b_1\sqrt{N}h + b_2\sqrt{N}h^2 + \dots + b_{P-1}\sqrt{N}h^{P-1} + O(\sqrt{N}h^P).$$

For Theorem 3.1, we require $Nh^{k+2} \to \infty$, so that the $\sqrt{N}h$ through $\sqrt{N}h^{P-1}$ terms explode, and we can choose h such that $\sqrt{N}h^P \to 0$. Therefore, the asymptotic bias will vanish if and only if the coefficients b_ρ, $\rho = 1, \ldots, P-1$, vanish.

Moreover, equation (3.20) indicates how the kernel function can be chosen so that the asymptotic bias vanishes; namely choose $K(.)$ whose "moments" of order $P-1$ or less are zero. In particular, we assume $K(.)$ is of order P, by assuming the following:

ASSUMPTION 6: *The kernel function* $K(.)$ *obeys*

$$\int K(u)\,du = 1, \qquad \int u_1^{l_1} \ldots u_k^{l_{\rho'}} K(u)\,du = 0 \quad \text{for} \quad l_1 + \ldots + l_{\rho'} < P, \quad \text{and}$$

$$\int u_1^{l_1} \ldots u_k^{l_{\rho'}} K(u)\,du \neq 0 \quad \text{for} \quad l_1 + \ldots + l_{\rho'} = P.$$

When $P > 2$ (or $k > 1$), the kernel $K(.)$ must take on positive and negative values, because its second moments must be zero. Thus bias is controlled by using positive and negative weights in the local averaging.

We summarize the above discussion on asymptotic bias as follows:

THEOREM 3.2: *Given Assumptions* 1–6, *if* h *obeys* $Nh^{2P} \to 0$ *as* $N \to \infty$, *then* $\sqrt{N}[E(\hat{\delta}_N) - \delta] = o(1)$.

Thus we have also shown the following theorem:

THEOREM 3.3: *Given Assumptions* 1–6, *if* h *obeys* $Nh^{k+2} \to \infty$ *and* $Nh^{2P} \to 0$ *as* $N \to \infty$, *then the average derivative estimator* $\hat{\delta}_N$ *of* (3.1) *is such that* $\sqrt{N}(\hat{\delta}_N - \delta)$ *has a limiting multivariate normal distribution with mean* 0 *and variance covariance matrix* Σ_δ *of* (3.17).

Unlike in the demonstration of asymptotic normality, the lack of asymptotic bias of $\hat{\delta}_N$ is directly associated with the pointwise bias properties of its nonparametric components. In particular, under our assumptions, the pointwise bias of the density derivative estimates, $E[\partial \hat{f}_i(x)/\partial x] - \partial f(x)/\partial x$, can be shown to be $O(h^P)$ (c.f. Silverman (1986), among others). Correspondingly, the average bias $E(\hat{\delta}_N) - \delta$ is $O(h^P)$, so that the asymptotic bias $\sqrt{N}[E(\hat{\delta}_N) - \delta]$ is $O(\sqrt{N}h^P)$, vanishing as $Nh^{2P} \to 0$.

Higher order kernels can be constructed in a number of ways; for instance, see Gasser, Mueller, and Mammitzsch (1985) or Robinson (1986). For our Monte Carlo analysis, we construct such kernels by taking weighted differences of density functions with varying spreads. This choice has an alternative interpretation in terms of a "generalized jackknife" method of bias control, as outlined in Appendix 2.

A practical concern with the use of higher order kernels may exist when the sample size N is small (relative to the dimension k). In particular, the estimates of the density function from (3.2) can be quite variable, when they are based on averaging a very small number of observations with positive and negative weights. In such cases, it may be practically advantageous to stabilize the density estimates somewhat by using a positive kernel: $K(.)$ such that $K(u) \geqslant 0$ for $u \in \Omega_K$, $\int K(u)\,du = 1$, and $\int uK(u)\,du = 0$. We examine this issue as part of the Monte Carlo analysis in Section 5.

3.4. *Measurement of Precision*

In addition to giving the basis for asymptotic normality, the U-statistic structure of $\hat{\delta}_N$ also suggests a natural estimator of the asymptotic variance-covariance matrix Σ_δ.[10] Recall that Σ_δ is the covariance matrix of $2r(z_i)$, where $r(z_i)$ of (3.18) is the limit of $r_N(z_i) = E[p_N(z_i, z_j)|z_i]$, with $p_N(z_i, z_j)$ the U-statistic component defined in (3.15). A kernel estimator of $r(z_i)$ is obtained directly as the sample analogue of $r_N(z_i)$, namely

$$(3.22)\qquad \hat{r}_N(z_i) \equiv \frac{-1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} p_N(z_i, z_j)$$

$$= \frac{-1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\left(\frac{1}{h}\right)^{k+1} K'\left(\frac{x_i - x_j}{h}\right)(y_i - y_j).$$

The asymptotic variance covariance matrix Σ_δ is estimated using $\hat{r}_N(z_i)$, $i = 1,\ldots,N$, via

$$(3.23)\qquad \hat{\Sigma}_\delta = 4\frac{\sum_i \hat{r}_N(z_i)\hat{r}_N(z_i)'}{N} - 4\hat{\delta}_N\hat{\delta}_N'.$$

The consistency of $\hat{\Sigma}_\delta$ for Σ_δ is established as

THEOREM 3.4: *Under the conditions of Theorem* 3.3, $\hat{\Sigma}_\delta$ *is a consistent estimator of* Σ_δ.

Hypothesis tests on the values of some or all of the components of δ can be performed with standard Wald statistics using $\hat{\delta}_N$ and $\hat{\Sigma}_\delta$. In particular, if $R\delta = \delta_0$ is a coefficient restriction of interest, where R is a $k_1 \times k$ matrix of full rank $k_1 \leqslant k$, then the limiting distribution of $N(R\hat{\delta}_N - \delta_0)'(R\hat{\Sigma}_\delta R')^{-1}(R\hat{\delta}_N - \delta_0)$ is χ^2 with k_1 degrees of freedom. For example, for single index models with $g(x) = G(x'\beta)$ and $\delta = \gamma\beta$, the (scale-free) hypothesis that $R\beta = 0$ is equivalent to $R\delta = \delta_0$ with $\delta_0 = 0$, which can be tested in this manner. This accommodates zero restrictions on some or all of the components of β, as well as equality restrictions.

[10] This procedure was suggested by an anonymous referee, whom the authors gratefully acknowledge.

4. SCALING BY INSTRUMENTAL VARIABLES REGRESSION

As discussed in the introduction, the scaling normalization implicit in the definition of $\delta = E[f(x)\,\partial g/\partial x]$ may not always yield the most easily interpreted estimates, so we now consider the rescaled parameter vector $\delta^* = \delta/E[f(x)]$. It is not difficult to propose estimators of δ^*: if $\bar{f}_N$ is any consistent estimator of $E[f(x)]$, then it is clear that δ^* is consistently estimated by $\hat{\delta}_N/\bar{f}_N$. In this section we discuss a particular modification of $\hat{\delta}_N$ which estimates δ^* by the slope coefficients of a linear instrumental variables regression of y_i on x_i.[11]

The motivation for the estimator begins by noting that $\partial x'/\partial x = I_k$, where I_k is the $k \times k$ identity matrix, and applying Lemma 2.1 to $I_k E[f(x)]$ as

$$(4.1) \qquad I_k E[f(x)] = -2E\left(\frac{\partial f}{\partial x} x'\right).$$

Therefore, $I_k E[f(x)]$ is consistently estimated by the density-weighted average derivative estimator replacing y_i by x_i' as

$$(4.2) \qquad \hat{\delta}_{xN} = \frac{-2}{N} \sum_{i=1}^{N} \left(\frac{\partial \hat{f}_i(x_i)}{\partial x}\right) x_i'$$

where $\hat{f}_i(x)$ is defined in (3.2). Further note that by combining (2.2) and (4.1), δ^* can be written as

$$(4.3) \qquad \delta^* = \left(E\left(\frac{\partial f}{\partial x} x'\right)\right)^{-1} E\left(\frac{\partial f}{\partial x} y\right).$$

Equation (4.3) motivates the following estimator of δ^*. Consider the slope coefficients of the linear equation:

$$(4.4) \qquad y_i = x_i' \hat{d}_N + \hat{u}_i \qquad (i = 1, \ldots, N)$$

estimated using the estimated density derivatives $\partial \hat{f}_i(x_i)/\partial x$ as instrumental variables, or

$$(4.5) \qquad \hat{d}_N = \hat{\delta}_{xN}^{-1} \hat{\delta}_N = \left(\sum_{i=1}^{N} \left(\frac{\partial \hat{f}_i(x_i)}{\partial x}\right) x_i'\right)^{-1} \left(\sum_{i=1}^{N} \left(\frac{\partial \hat{f}_i(x_i)}{\partial x}\right) y_i\right).$$

The above remarks motivate how $\hat{d}_N$ is a consistent estimator of δ^*. We have omitted a constant term from the linear equation (4.4) to most easily motivate the instrumental variables formula (4.5). However, it should be noted that the value of the slope coefficient vector $\hat{d}_N$ is unaffected by the inclusion of a constant, since $\Sigma_i[\partial \hat{f}_i(x_i)/dx] = 0$ by construction.

The limiting distribution of $\hat{d}_N$ can be derived by writing its departure from δ^* in terms of the large sample residuals from the equation (4.4). In particular,

[11] Alternatively, we could rescale by dividing by the sample average of the estimated density function $\hat{f}(x_i)$, which would yield an estimator with a different large-sample distribution. We consider only the instrumental variables rescaling here, since this approach yields an estimator which is unbiased (conditioned on the regressors) when the true model is linear in the index $x'\beta$.

define $u = y - x'\delta^*$ and $u_i = y_i - x_i'\delta^*$, $i = 1, \dots, N$, so that by construction we have

$$(4.6) \qquad \hat{d}_N - \delta^* = \hat{\delta}_{xN}^{-1}\left(-2N^{-1}\sum_{i=1}^{N}\left(\frac{\partial \hat{f}_i(x_i)}{\partial x}\right)u_i\right).$$

By Theorem 3.3, the term in brackets consistently estimates the density-weighted average derivative of $D(x) \equiv E(u|x) = g(x) - x'\delta^*$. But this average derivative is 0; namely with $\partial D/\partial x = \partial g/\partial x - \delta^*$, we have $E[f(x)\,\partial D/\partial x] = 0$. Thus, $\hat{d}_N$ is consistent, as $\operatorname{plim} \hat{d}_N - \delta^* = E[f(x)]^{-1}E[f(x)\,\partial D/\partial x] = 0$. Moreover, for the same reason, $\hat{\delta}_{xN}$ can be replaced by its probability limit $I_k E[f(x)]$, giving

$$(4.7) \qquad \sqrt{N}\,[\hat{d}_N - \delta^*] = E[f(x)]^{-1}\sqrt{N}\left(-2N^{-1}\sum_{i=1}^{N}\left(\frac{\partial \hat{f}_i(x_i)}{\partial x}\right)u_i\right) + o_p(1).$$

In this form, it is clear that the limiting distribution of $\sqrt{N}(\hat{d}_N - \delta^*)$ follows from an immediate application of Theorem 3.3. The asymptotic covariance matrix is

$$(4.8) \qquad \Sigma_d \equiv 4E[r_d(z_i)r_d(z_i)']$$

where the component $r_d(z_i)$ is derived by repeating the derivation (3.14–15) with y replaced by $u = y - x'\delta^*$ and then multiplying by $E[f(x)]^{-1}$ as

$$(4.9) \qquad r_d(z_i) = E[f(x)]^{-1}\left(f(x_i)\left(\frac{\partial g(x_i)}{\partial x} - \delta^*\right) - [y_i - g(x_i)]\frac{\partial f(x_i)}{\partial x}\right).$$

We summarize this discussion of the asymptotic properties of $\hat{d}_N$ as follows:

COROLLARY 4.1: *Given Assumptions* 1–6, *if* h *obeys* $Nh^{k+2} \to \infty$ *and* $Nh^{2P} \to 0$ *as* $N \to \infty$, *then the estimator* $\hat{d}_N$ *of* (4.5) *is such that* $\sqrt{N}(\hat{d}_N - \delta^*)$ *has a limiting multivariate normal distribution with mean* 0 *and variance covariance matrix* Σ_d *of* (4.8).

Moreover, the matrix Σ_d is consistently estimated by using the estimated residuals from equation (4.4): let $\hat{u}_i = y_i - x_i'\hat{d}_N$, $i = 1, \dots, N$, denote these residuals, and in accordance with (3.22) define

$$(4.10) \qquad \hat{r}_{dN}(z_i) \equiv \hat{\delta}_{xN}^{-1}\left(\frac{-1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N}\left(\frac{1}{h}\right)^{k+1}K'\left(\frac{x_i - x_j}{h}\right)(\hat{u}_i - \hat{u}_j)\right)$$

and $\hat{\Sigma}_d$ by

$$(4.11) \qquad \hat{\Sigma}_d = 4\frac{\sum_i \hat{r}_{dN}(z_i)\hat{r}_{dN}(z_i)'}{N}.$$

We then have the following corollary:

COROLLARY 4.2: *Under the conditions of Corollary* 4.1, $\hat{\Sigma}_d$ *is a consistent estimator of* Σ_d.

Inferences on some or all of the values of δ^* can be carried out using Wald statistics constructed from $\hat{d}_N$ and $\hat{\Sigma}_d$, as outlined in Section 3.4.

5. FINITE SAMPLE BEHAVIOR

In order to evaluate the practical performance of the approach discussed in the previous sections, in this section we present the results of a small-scale simulation study of the proposed estimators. While it is not possible to completely characterize the sampling behavior of the weighted average derivative or instrumental variables estimator under the general conditions imposed in Section 2, the results presented below are quite suggestive of the applicability of the large-sample theory to finite samples.

Four models consistent with the "single index" specification (1.2) were used in the study; the dependent variable y_i in each case was generated from an underlying linear model with two covariates,

$$y_i^* = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad (i = 1, \dots, N), \tag{5.1}$$

with true values $\alpha = 0$ and $\beta_1 = \beta_2 = 1$ held constant across designs. In the "linear" specifications, it is assumed that $y_i = y_i^*$, i.e., the true latent variable is observed; in the "binary response" models, only an indicator variable denoting positivity of y_i^* is observed, so that $y_i = 1(y_i^* > 0)$, where "$1(A)$" denotes the indicator function of the event "A". For each of these two model specifications, two conditional distributions of the error terms ε_i were used. In the "homoskedastic" designs, ε_i was assumed to be independent of $\{x_i\}$ and i.i.d., with a standard Gaussian distribution. The "heteroskedastic" designs assumed the error distribution was multiplicatively heteroskedastic, that is, $\varepsilon_i = \sigma_i \cdot \nu_i$, where ν_i is an i.i.d. standard Gaussian sequence and $\sigma_i^2 = \exp\{x_i'\beta + k\}$, where k is a constant chosen so that, given the distribution of the regressors, $E[\sigma_i^2]$ is equal to one. For each of these model/distribution pairs, it is easy to verify the relationship (1.2); in all but the "heteroskedastic binary" case, the corresponding function $G(\cdot)$ is nondecreasing in the argument $x'\beta$.

Given the imposed symmetry in the way the covariates enter the model (with equal coefficients), it is important that the covariates *not* be identically distributed. Otherwise, when a scale normalization is imposed, any estimation method which is symmetric in the two covariates will tend to be median unbiased. For the results reported below, the first covariate x_{i1} was assumed to have a χ_3^2 distribution, standardized to have zero mean and unit variance; the second covariate was independently distributed and standard normal.

The weighted average derivative estimators and corresponding instrumental variables estimators for all designs are calculated using two kernel functions. The "not bias-corrected" kernel is a standard multivariate normal density function, with zero mean and identity covariance matrix. This kernel does not satisfy Assumption 6 for these models, so the asymptotic bias will not be $o(1/\sqrt{N})$; still, the magnitude of this bias for finite samples is an open question. The "bias-corrected" kernel is constructed using the "generalized jackknife" approach described in Appendix 2: that is, it is a linear combination of $P = 4$ multivariate normal density functions, with weights and bandwidths chosen to ensure the conditions of Assumption 6. In the notation of Appendix 2, $(\psi_1, \psi_2, \psi_3) = (2, 3, 4)$ and $\eta = 0$, so that $c_N \equiv c = (1.5, -1.0, 0.25)'$. Fixing these kernel functions, the only remaining free parameter in the estimators $\hat{\delta}_N$ of (3.1) and $\hat{d}_N$ of (4.7) is the bandwidth, h.

Tables I through IV report summary statistics for various estimators under the four model/error distribution combinations—homoskedastic linear, homoskedastic binary response, heteroskedastic linear, and heteroskedastic binary response, respectively. The simulations reported here took the sample size $N = 50$, the number of covariates $k = 2$, and the bandwidth parameter $h = 1.0$. All designs were replicated 400 times; summary statistics reported for these replications include the sample mean (MEAN), standard deviation (SD), and root-mean-squared-error (RMSE), as well as the lower quartile (LQ), median (MEDIAN), upper quartile (UQ), and median absolute error (MAE). All simulations and calculations were performed using the GAUSS programming language on microcomputers.

In order to ensure comparability of the estimated coefficients across estimators and designs, all estimated slope coefficients were rescaled to have the sum of their absolute values equal to 2, which is the sum of magnitudes of the true coefficients β_0. For this normalization, if both estimated coefficients are positive (as is typically the case), the deviations of the two coefficient estimates from their true values will be of equal magnitude and opposite sign; that is, $\hat{\beta}_1 - \beta_1 = \beta_2 - \hat{\beta}_2$ when $\hat{\beta}_1$, $\hat{\beta}_2 > 0$. This normalization was preferred to examination of ratios of estimated slope coefficients, which have ill-behaved sample moments; it was also preferred to normalization of the Euclidian length of the coefficients, which induces more asymmetry in the sampling distribution of the coefficient estimators about the true values. Except for comparison of magnitudes of sample moments, though, the qualitative conclusions below do not depend on the particular normalization chosen.

For all designs, the (rescaled) classical least squares estimator was calculated and summarized, to provide a standard for comparison; in Tables II and IV, the behavior of the probit maximum likelihood estimator (under the assumption of homoskedastic Gaussian errors) is also summarized. While the least squares estimator is not consistent for the "binary response" models, it is often justified as a computationally-convenient estimator for the homoskedastic binary response model when the expectation function $G(x'\beta)$ is not frequently close to zero or one (see, e.g., Amemiya (1981)). In Tables I and II, then, the semiparametric estimators can be compared to properly-specified maximum likelihood estimators, while in Tables III and IV misspecified maximum likelihood estimators are the standard for comparison.

TABLE I

FINITE-SAMPLE BEHAVIOR OF ESTIMATORS FOR HOMOSKEDASTIC LINEAR MODEL

Weighted Average Derivative, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.11	0.35	0.37	0.90	1.12	1.34	0.24
1.00	0.86	0.41	0.43	0.65	0.88	1.10	0.24

Weighted Average Derivative, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.12	0.38	0.40	0.89	1.11	1.34	0.25
1.00	0.84	0.45	0.48	0.65	0.87	1.11	0.26

Instrumental Variables, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.01	0.36	0.36	0.78	0.99	1.21	0.21
1.00	0.96	0.42	0.43	0.78	1.00	1.21	0.21

Instrumental Variables, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.01	0.38	0.38	0.77	0.99	1.24	0.23
1.00	0.94	0.48	0.48	0.76	1.00	1.22	0.24

Least Squares:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.01	0.29	0.29	0.83	1.02	1.18	0.17
1.00	0.98	0.32	0.32	0.81	0.98	1.17	0.18

TABLE II

FINITE-SAMPLE BEHAVIOR OF ESTIMATORS FOR HOMOSKEDASTIC BINARY RESPONSE MODEL

Weighted Average Derivative, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.07	0.44	0.45	0.81	1.08	1.35	0.27
1.00	0.88	0.50	0.51	0.64	0.92	1.18	0.27

Weighted Average Derivative, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.06	0.46	0.47	0.79	1.08	1.36	0.29
1.00	0.87	0.53	0.55	0.62	0.92	1.20	0.29

Instrumental Variables, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	0.97	0.44	0.44	0.71	0.97	1.25	0.27
1.00	0.97	0.52	0.52	0.74	1.02	1.29	0.27

Instrumental Variables, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	0.96	0.46	0.46	0.70	0.95	1.26	0.28
1.00	0.96	0.56	0.56	0.72	1.04	1.29	0.28

Least Squares:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.00	0.37	0.37	0.77	1.01	1.23	0.23
1.00	0.99	0.37	0.37	0.76	0.99	1.23	0.23

Probit Maximum Likelihood:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	0.97	0.37	0.38	0.73	0.96	1.22	0.24
1.00	1.01	0.38	0.38	0.78	1.04	1.27	0.24

TABLE III

FINITE-SAMPLE BEHAVIOR OF ESTIMATORS FOR HETEROSKEDASTIC LINEAR MODEL

Weighted Average Derivative, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.08	0.42	0.43	0.88	1.10	1.32	0.23
1.00	0.83	0.50	0.53	0.66	0.88	1.11	0.23

Weighted Average Derivative, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.09	0.42	0.43	0.89	1.09	1.33	0.23
1.00	0.83	0.48	0.51	0.64	0.89	1.10	0.24

Instrumental Variables, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	0.98	0.41	0.41	0.78	0.99	1.21	0.22
1.00	0.92	0.53	0.54	0.78	1.00	1.22	0.22

Instrumental Variables, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	0.99	0.41	0.41	0.78	0.98	1.20	0.22
1.00	0.93	0.51	0.51	0.74	1.01	1.21	0.22

Least Squares:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	0.85	0.63	0.65	0.60	0.90	1.23	0.31
1.00	0.67	0.93	0.99	0.35	0.99	1.25	0.33

TABLE IV

FINITE-SAMPLE BEHAVIOR OF ESTIMATORS FOR HETEROSKEDASTIC BINARY RESPONSE MODEL

Weighted Average Derivative, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.16	0.37	0.40	0.93	1.17	1.42	0.28
1.00	0.81	0.42	0.46	0.57	0.83	1.07	0.28

Weighted Average Derivative, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.14	0.40	0.42	0.91	1.14	1.42	0.29
1.00	0.81	0.45	0.49	0.56	0.84	1.08	0.29

Instrumental Variables, Not Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.06	0.37	0.38	0.80	1.05	1.31	0.26
1.00	0.90	0.44	0.45	0.68	0.95	1.19	0.26

Instrumental Variables, Bias-Corrected:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.04	0.40	0.40	0.78	1.03	1.32	0.26
1.00	0.90	0.47	0.48	0.67	0.96	1.22	0.26

Least Squares:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.15	0.39	0.42	0.90	1.11	1.42	0.25
1.00	0.82	0.43	0.47	0.58	0.87	1.09	0.25

Probit Maximum Likelihood:

TRUE	MEAN	SD	RMSE	LQ	MEDIAN	UQ	MAE
1.00	1.17	0.39	0.42	0.92	1.12	1.44	0.24
1.00	0.81	0.43	0.47	0.55	0.87	1.08	0.24

The first two entries of Table I summarize the behavior of the weighted average derivative estimators for the base design. As a glance down the "MEAN" column indicates, these estimators are not very well-behaved for any of the models in the "base design." The estimates are significantly biased away from their true values, and this bias causes a noticeable increase in the RMSE over the standard deviation of the estimator. Moreover, the bias is not due to asymmetry of the sampling distribution of the estimator; the "MEDIAN" column follows the same pattern as the sample means. Finally, looking across the tables, there seems to be no systematic improvement in either bias or mean-squared error between the "bias-corrected" and "not bias-corrected" versions of the estimator.

In contrast, the instrumental variables estimators $\hat{d}_N$ calculated for the same data are quite well behaved in terms of bias, as the next two entries of Table I illustrate. The (mean or median) bias of the finite-sample distribution of the instrumental variables estimators is not significantly different from zero (at a 5 percent level), even for the "not bias-corrected" kernel. This suggests that the instrumental variables "correction" to the weighted average derivative estimator $\hat{\delta}_N$ is a much more important means of bias-correction than the generalized jackknife in practice. Heuristically, the behavior of the instrumental variables estimator $\hat{d}_N$ relative to $\hat{\delta}_N$ is analogous to the behavior of the least squares estimator relative to the "product moment" estimator $N^{-1}\Sigma x_i y_i$: for the linear models studied here, both are consistent up to scale (since $E[x_i x_i'] = I$), but least squares is also conditionally (on $\{x_i\}$) unbiased for the unscaled regression coefficients. As for the "generalized jackknife" correction for bias, it yields no systematic reduction in bias in the results of Table I through IV, but is systematically more variable.

Comparing instrumental variables and least squares directly, the results of Tables I and II indicate a higher precision of the least squares estimators for both homoskedastic models, with the behavior of probit maximum likelihood being quite similar to least squares (in Table II). For the heteroskedastic linear model of Table III, though, the least squares coefficients are very poorly behaved relative to the instrumental variables estimators. While the unnormalized least squares and instrumental variables coefficients would be unbiased for β_0 (conditionally on the regressors) in this model, the former would also be very dispersed, with a substantial proportion of negative values; imposition of the normalization in this case yields a sampling distribution of the coefficients with large negative biases in both coefficients. The instrumental variables estimators in the design of Table IV have a comparable precision (in terms of RMSE or MAE) to either least squares or the probit maximum likelihood estimator. However, for the heteroskedastic binary response model, the least squares and probit estimators are significantly biased, while the instrumental variables estimator is substantially less biased (albeit with a larger dispersion). Since the bias of least squares need not decline as the sample size increases, this suggests that, in larger samples, the instrumental variables estimator will dominate least squares or probit, provided the bandwidth is chosen to shrink with the sample size at the appropriate rate.

Several variations on the design of Table I were used to investigate the effects of the sample size, bandwidth parameter, and number of covariates on the sampling behavior of the estimators; while the results of these simulations are not summarized here, they are fairly unsurprising in view of the foregoing theoretical and simulation results. Doubling the sample size (with a commensurate reduction in the bandwidth parameter) does not lead to a substantial improvement of the precision of the semiparametric relative to the parametric estimators. This indicates that the variance reduction due to a larger sample size is partially offset by the reduction in the bandwidth parameter of the instrumental variables estimator, at least for these designs and sample sizes. As the bandwidth parameter is doubled, the standard deviations of its sampling distributions decline, though for the binary response model this is accompanied by an increase in the magnitude of the bias. Still, the RMSE and the MAE of the least squares and instrumental variables estimators are quite close, and the change in precision of the latter is not dramatic. Finally, inclusion of an additional covariate (with true coefficient equal to zero) causes a general increase in the sampling variability of all the estimators; this effect is more pronounced for the binary response models. Also, the magnitude of the bias of the instrumental variables estimators increases, since the bandwidth is held fixed as the dimensionality of the estimated joint density function of the regressors increases. Again, the changes in the summary measures of dispersion are not dramatic.

While it is difficult to arrive at general conclusions about finite-sample performance on the basis of the small number of models investigated here, the results do suggest a number of working hypotheses which may be useful as a guide to practical application of the proposed procedures:

(i) the instrumental variables "rescaling" of the weighted average derivative estimator is an important bias-reduction adjustment to the approach, even if a further scaling restriction is imposed;

(ii) the estimators using the "bias-corrected" kernels are not systematically less biased than the estimators based upon standard positive kernels, and have higher dispersion; and

(iii) the measures of dispersion of the (not bias-corrected) instrumental variables estimator are of comparable magnitude to the least squares estimator, but in some cases the latter estimator is substantially biased, while the magnitude of the bias of the instrumental variables estimator is small across all designs considered.

6. CONCLUDING REMARKS

In this paper we have proposed an estimator $\hat{\delta}_N$ of the density-weighted average derivative $\delta = E[f(x)\,\partial g/\partial x]$, as a solution to the problem of estimating the coefficients β up to scale in single index models with $g(x) = G(x'\beta)$. This estimator is based on averaging of nonparametric kernel estimates of the derivatives of the density $f(x)$, and can be computed directly from the data, requiring no computational techniques for maximization or other types of equation solving. We have shown that $\hat{\delta}_N$ is $\sqrt{N}$-consistent and asymptotically normal, and give a consistent estimator of its asymptotic variance-covariance matrix. We have also proposed a general estimator $\hat{d}_N$ of the correctly-scaled weighted average $\delta^* = E[f(x)\,\partial g/\partial x]/E[f(x)]$, as the estimated slope coefficients of the linear regression of y regressed on x, using estimated density derivatives as instrumental variables. The practical performance of these estimators was studied via Monte Carlo simulation, where $\hat{d}_N$ was seen to display better performance in small samples.

In broader statistical terms, our results have an interesting role in the general theory of estimation, as a bridge between known distributional properties of nonparametric estimators. On the one hand, nonparametric pointwise estimates of density functions or regression functions are consistent for the true values at rates that are necessarily slower than $\sqrt{N}$ (c.f. Stone (1980) and McFadden (1985) among others). On the other hand, central limit theory states that sample average statistics (as well as related estimators of finite parameter vectors, such as maximum likelihood) are $\sqrt{N}$-consistent for their limits. Our results give a nontrivial situation where averaging of nonparametric pointwise estimates permits $\sqrt{N}$-consistency (and asymptotic normality) to be attained for finite vectors (δ and δ^*), while maintaining no model specific restrictions. The issues of large data requirements surrounding nonparametric characterizations of density functions and regression functions do not apply to the estimation of these weighted average derivatives.

Related to our work are several recent applications of kernel techniques to econometric estimation. Robinson (1988) shows how $\sqrt{N}$-consistent, asymptotically normal estimators can be obtained for coefficients in semilinear models, using a higher order kernel for bias control but an alternate proof of normality. In similar vein, Powell (1987) derives a $\sqrt{N}$-consistent estimator for coefficients in models of selected samples, presuming that the coefficients of the selection equation have been estimated (for instance, by the methods proposed in this paper). Stock (1989) proves asymptotic normality for a specific average kernel estimator (centered around its mean), and analyzes the asymptotic bias via simulation. For single index models, Ichimura (1987) proposes a least-squares approach to estimating β up to scale that uses kernel estimation in the optimizing conditions, and Han (1987) has proposed estimation of index coefficients on the basis of maximizing rank correlation.

Our characterization of the asymptotic distributions of $\hat{\delta}_N$ and $\hat{d}_N$ uses an extension of the classical U-statistic theory of Hoeffding (1948). The simple U-statistic structure of $\hat{\delta}_N$ arises because of the choice of the density $f(x)$ as a weighting function. Kernel estimation of average derivatives with other weighting functions does not give rise to a simple U-statistic—rather complex technical and practical problems arise from averaging nonlinear combinations of kernel estimators. In this vein, Härdle and Stoker (1988) give results on a trimmed kernel estimator of the unweighted average derivative $E[\partial g/\partial x]$. The same problems arise with statistics using kernel regression function estimators—for instance, δ

could be estimated using $(1/N)\Sigma_i \hat{f}(x_i)\,\partial\hat{g}(x_i)/\partial x$, where $\hat{g}(x)$ is a kernel estimator of the regression function $g(x)$. The relationship of these approaches to our results raises interesting questions for future research.

Our results pose a large number of practical future research questions as to the best way to implement the average derivative estimator. While we have established the proper asymptotic behavior of the kernel bandwidth to establish attractive statistical properties for $\hat{\delta}_N$ and $\hat{d}_N$, future research is necessary to indicate the best way to set bandwidth size in applications, such as whether for average derivative estimators there exist desirable "cross-validation" techniques (c.f. Silverman (1986) for a survey of these methods). Similarly, further study is required on the impact of the choice of kernel function in small data samples.

Department of Economics, University of Wisconsin, Madison, WI 53706, U.S.A.;
Kennedy School of Government, Harvard University, Cambridge, MA 02138, U.S.A.;
and
Sloan School of Management, M.I.T., Cambridge, MA 02139, U.S.A.

Manuscript received September, 1986; final revision received March, 1989.

APPENDIX 1

OMITTED PROOFS

PROOF OF LEMMA 2.1: Let x_1 denote the first component of x, and x_0 the other components, so that $x=(x_1,x_0')'$. For a given value of x_0, denote the range of x_1 as $\omega(x_0)=\{x_1\,|\,(x_1,x_0')'\in\Omega\}$. Now apply Fubini's Theorem (c.f. Billingsley (1979), among others) to write $E(f(x)\,\partial g/\partial x_1)$ as

$$\text{(A1.1)}\qquad \int_\Omega \frac{\partial g(x)}{\partial x_1}(f(x))^2\,dx=\int\left(\int_{\omega(x_0)}\frac{\partial g(x)}{\partial x_1}(f(x))^2\,dx_1\right)dx_0 .$$

The result follows from the validity of the following equation:

$$\text{(A1.2)}\qquad \int_{\omega(x_0)}\frac{\partial g(x)}{\partial x_1}(f(x))^2\,dx_1=-2\int_{\omega(x_0)}g(x)\frac{\partial f(x)}{\partial x_1}f(x)\,dx_1 .$$

By inserting (A1.2) into (A1.1), $E(f(x)\,\partial g/\partial x_1)=-2E(g(x)\,\partial f/\partial x_1)$ is established, and by iterated expectation, $E(g(x)\,\partial f/\partial x_1)=E(y(\partial f/\partial x_1))$.

To establish (A1.2), note first that the convexity of Ω implies that $\omega(x_0)$ is either a finite interval $[a,b]$ (where a, b depend on x_0), or an infinite interval of the form $[a,\infty),(-\infty,b]$ or $(-\infty,\infty)$. Supposing first that $\omega(x_0)=[a,b]$, integrate the left-hand side of (A1.2) by parts (c.f. Billingsley (1979)) as

$$\text{(A1.3)}\qquad \begin{aligned}\int_a^b\frac{\partial g(x)}{\partial x_1}(f(x))^2\,dx_1&=-2\int_a^b g(x)\frac{\partial f(x)}{\partial x_1}f(x)\,dx_1\\ &\quad+g(b,x_0)(f(b,x_0))^2-g(a,x_0)(f(a,x_0))^2.\end{aligned}$$

The latter two terms represent gf^2 evaluated at boundary points, which vanish by Assumption 2, so that (A1.2) is established for $\omega(x_0)=[a,b]$.

For the unbounded case $\omega(x_0)=[a,\infty)$, note first that the existence of $E(f(x)y)$, $E(f(x)\,\partial g/\partial x_1)$, and $E(g(x)\,\partial f/\partial x_1)$ respectively imply the existence of $E(f(x)g(x)\,|\,x_0)$,

$E(f(x)\,\partial g/\partial x_1 | x_0)$, and $E(g(x)\,\partial f/\partial x_1 | x_0)$ (c.f. Kolmogorov (1950)). Now consider the limit of (A1.3) over intervals $[a, b]$, where $b \to \infty$, rewritten as

$$\begin{aligned}\lim_{b\to\infty} g(b, x_0)(f(b, x_0))^2 &= g(a, x_0)(f(a, x_0))^2 + \lim_{b\to\infty}\int_a^b \frac{\partial g(x)}{\partial x_1} f(x)^2\,dx_1 \\ &\quad + 2\lim_{b\to\infty}\int_a^b g(x)\frac{\partial f(x)}{\partial x_1} f(x)\,dx_1 \\ &= g(a, x_0)(f(a, x_0))^2 + f_0(x_0)E\left(\frac{\partial g}{\partial x_1} f(x) | x_0\right) \\ &\quad + 2f_0(x_0)E\left(g(x)\frac{\partial f}{\partial x_1} | x_0\right)\end{aligned}$$

so that $C \equiv \lim g(b, x_0)f(b, x_0)^2$ exists, where $f_0(x_0)$ is the marginal density of x_0. Now suppose that $C > 0$. Then there exists scalars ε and B such that $0 < \varepsilon < C$ and for all $b \leqslant B$, $|g(b, x_0)f(b, x_0)^2 - C| < \varepsilon$. Therefore $g(x_1, x_0)f(x_1, x_0)^2 > (C - \varepsilon)I_{[B,\infty)}$, where $I_{[B,\infty)}$ is the indicator function of $[B, \infty)$. But this implies that

$$f_0(x_0)E(g(x)f(x)|x_0) = \int g(x_1, x_0)f(x_1, x_0)^2\,dx_1 > (C - \varepsilon)\int I_{[B,\infty)}\,dx_1 = \infty,$$

which contradicts the existence of $E(g(X)f(x)|x_0)$. Consequently, $C > 0$ is ruled out. $C < 0$ similarly contradicts the existence of $E(g(x)f(x)|x_0)$.

Since $C \equiv \lim g(b, x_0)f(b, x_0)^2 = 0$, and $g(a, x_0)f(a, x_0)^2 = 0$ by Assumption 2, equation (A1.2) is valid for $\omega(x_0) = [a, \infty)$. Analogous arguments establish the validity of (A1.2) for $\omega(x_0) = (-\infty, a]$ and $\omega(x_0) = (-\infty, \infty)$. *Q.E.D.*

Proof of Lemma 3.1: We prove the equivalence by showing that $NE[\|U_N - \hat{U}_N\|^2] = o(1)$, where $\|U_N - \hat{U}_n\|^2 = (U_n - \hat{U}_n)'(U_N - \hat{U}_n)$. Define

$$q_N(z_i, z_j) = [p_N(z_i, z_j) - r_N(z_i) - r_N(z_j) + \theta_n],$$

so that

$$U_N - \hat{U}_N = \binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} q_N(z_i, z_j).$$

The expectation of the squared length of the vector $U_N - \hat{U}_N$ is

$$E[\|U_N - \hat{U}_N\|^2] = \binom{N}{2}^{-2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{l=1}^{N-1}\sum_{m=l+1}^{N} E[q_N(z_i, z_j)'q_N(z_l, z_m)].$$

Because z_i, $i = 1, \dots, N$, are independent vectors, all terms with $(i, j) \neq (l, m)$ have zero expectations (if $i \neq l$ and $j \neq m$ this is obvious, and writing out the expectation for $i = l$, $j \neq m$ gives a quick verification). Therefore,

$$E[\|U_N - \hat{U}_N\|^2] = \binom{N}{2}^{-2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E[\|q_N(z_i, z_j)\|^2].$$

The number of terms in this double summand is $O(N^2)$. The nonzero expectations are each $O(E\|q_N(z_i, z_j)\|^2) = O(E\|p_N(z_i, z_j)\|^2) = o(N)$, the latter equality by assumption. Consequently,

$$\begin{aligned}NE[\|U_N - \hat{U}_N\|^2] &= N\binom{N}{2}^{-2} O(N^2)o(N) \\ &= o(1),\end{aligned}$$

as required. *Q.E.D.*

Proof of Theorem 3.4: Because $\hat{\delta}_N$ is consistent for δ, we focus on the leading term $N^{-1}\Sigma \hat{r}_N(z_i)\hat{r}_N(z_i)'$, and establish that it is consistent for $E[r(z)r(z)']$. Letting $p_{Nl}(z_i, z_j)$ and $r_{Nl}(z_i)$ denote the lth components of $p_N(z_i, z_j)$ and $r_N(z_i)$ defined in (3.11) and (3.15), we first note that

$$\begin{aligned} E\|\hat{r}_N(z_i) - r_N(z_i)\|^2 &= \sum_{l=1}^{k} E[\operatorname{Var}(\hat{r}_{Nl}(z_i)|z_i)] \\ &= \frac{1}{N-1}\sum_{l=1}^{k} E[\operatorname{Var}(p_{Nl}(z_i, z_j)|z_i)] \\ &\leqslant \frac{1}{N-1} E[\|p_N(z_i, z_j)\|^2] = O(1/Nh^{k+2}), \end{aligned}$$

so that $E[\|\hat{r}_N(z_i) - r_N(z_i)\|^2] = o(1)$ under the condition imposed on the bandwidth h. Furthermore, the argument following (3.16) above implies that $t_N(z_i)$, defined in equation (3.15), has

$$E[\|t_N(z_i)\|^2] \equiv E[\|r_N(z_i) - r(z_i)\|^2] = o(h^2) = o(1),$$

since $h \to 0$. So

$$E[\|\hat{r}_N(z_i) - r(z_i)\|^2] = o(1),$$

which implies

$$E[\|\hat{r}_N(z_i)\hat{r}_N(z_i)' - r(z_i)r(z_i)'\|] = o(1),$$

where, for a matrix A, $\|A\| \equiv [\operatorname{trace}(A'A)]^{1/2}$. Markov's inequality and the SLLN thus yield

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{n} \hat{r}_N(z_i)\hat{r}_N(z_i)' &= \frac{1}{N}\sum_{i=1}^{n} r(z_i)r(z_i)' + o_p(1) \\ &= E[r(z_i)r(z_i)'] + o_p(1), \end{aligned}$$

as desired.

Proof of Corollary 4.2: *Since $\hat{d}_N$ and $\hat{\delta}_{xN}$ are consistent for δ^* and $I_k E[f(x)]$ and do not vary with i, it is straightforward to modify the proof of Theorem* 3.4 *above to apply here, replacing "$y_i - y_j$" with "$(x_i - x_j)'$" where necessary.*

APPENDIX 2

BIAS CONTROL VIA JACKKNIFING

In this appendix, we outline a different method of bias control than that used in the exposition, namely a generalized jackknife. Following the description of this approach, we point out how certain higher dimensional kernels can be constructed using the jackknife formulae.

Standard jackknifing procedures, as introduced by Quenouille (1949), are based on the fact that the bias in many estimators depends on sample size, and that the bias can be estimated by taking differences in estimators computed from samples of varying sizes (see Efron (1982) for a recent exposition). As in Schucany and Sommers (1977) and Bierens (1987), here we utilize the bandwidth h in the role of sample size. In particular, one can remove asymptotic bias of $\hat{\delta}_N$ by subtracting from it a weighted sum of $P - 1$ kernel estimators with differing bandwidths. Suppose that $K(.)$ is a standard positive kernel, and for $\rho = 1, \ldots, P-1$, let $\hat{\delta}_{\rho N}$ denote the estimator

$$\text{(A2.1)} \qquad \hat{\delta}_{\rho N} = \frac{-2}{N}\sum_{i=1}^{N}\left(\frac{\partial \hat{f}_{\rho i}(x_i)}{\partial x}\right) y_i$$

where $\hat{f}_{\rho i}(x)$ is the kernel density estimator with bandwidth h_ρ:

$$\hat{f}_{\rho i}(x) = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{1}{h_\rho}\right)^k K\left(\frac{x - x_j}{h_\rho}\right). \tag{A2.2}$$

A "jackknifed" estimator $\bar{\delta}_N$ is defined as

$$\bar{\delta}_N = \frac{\hat{\delta}_N - \sum_\rho c_{\rho N} \hat{\delta}_{\rho N}}{1 - \sum_\rho c_{\rho N}}, \tag{A2.3}$$

where $c_{\rho N}$, $\rho = 1, \dots, P-1$, are a set of weights that vary with N.

The estimator $\bar{\delta}_N$ will display no asymptotic bias if the bandwidths h_ρ and the weights $c_{\rho N}$, $\rho = 1, \dots, P-1$ are chosen so that the leading terms of the bias expansion of $\Sigma_\rho c_{\rho N} \hat{\delta}_{\rho N}$ match those of $\hat{\delta}_N$. To achieve this, set $h_\rho = \psi_\rho h^{1-\eta}$, where $0 \leqslant \eta < 1/(k+1)$ and where ψ_ρ, $\rho = 1, \dots, P-1$ are distinct positive constants. Let $c_N = (c_{1N}, \dots, c_{P-1,N})'$ be defined as

$$c_N = \Psi^{-1} \begin{pmatrix} h^\eta \\ \vdots \\ h^{(P-1)\eta} \end{pmatrix}; \qquad \Psi = \begin{pmatrix} \psi_1 & \cdots & \psi_{P-1} \\ \vdots & & \vdots \\ \psi_1^{P-1} & \cdots & \psi_{P-1}^{P-1} \end{pmatrix}. \tag{A2.4}$$

The properties of the "jackknifed" estimator $\bar{\delta}_N$ are summarized as follows:

THEOREM A1: *Given Assumptions* 1–5, *if h obeys $Nh^{k+2} \to \infty$ and $Nh^{2[P-\eta(P-1)]} \to 0$ as $N \to \infty$, and if $\eta > 0$, then the "jackknifed" estimator $\bar{\delta}_N$ of* (*A*2.3) *is such that $\sqrt{N}(\bar{\delta}_N - \delta)$ has a limiting multivariate normal distribution with mean* 0 *and variance covariance matrix Σ_δ of* (3.17).

PROOF: $Nh^{k+2} \to \infty$ implies that $Nh_\rho^{k+2} = N\psi_\rho^{k+2} h^{(1-\eta)(k+2)} \to \infty$, so Theorem 3.1 implies that $\sqrt{N}[\hat{\delta}_{\rho N} - E(\hat{\delta}_{\rho N})]$ has a limiting normal distribution for each $\rho = 1, \dots, P-1$. If $\tilde{\delta}_N = (\delta_N', \hat{\delta}_{1N}', \dots, \hat{\delta}_{P-1,N}')$, the Cramer-Wold device implies that $\sqrt{N}[\tilde{\delta}_N - E(\tilde{\delta}_N)]$ has a limiting normal distribution with mean 0 and variance-covariance matrix $\Sigma_{\tilde{\delta}}$. Consequently, since $0 < \eta < 1/(k+1) < 1/(P-1)$, we have $c_N \to 0$ as $h \to 0$, and $\sqrt{N}[\bar{\delta}_N - E(\bar{\delta}_N)]$ has a limiting normal distribution with mean 0 and variance covariance matrix $\Sigma_\delta = \lim_{N \to \infty} [(1, c_N')' \otimes I_k]' \Sigma_{\tilde{\delta}} [(1, c_N')' \otimes I_k]$, where I_k is the $k \times k$ identity matrix.

The result follows from $\lim \sqrt{N}[E(\bar{\delta}_N) - \delta] = 0$. This is verified directly: First evaluate (3.18) for each $\hat{\delta}_{\rho N}$ as

$$\begin{pmatrix} E(\hat{\delta}_{1N}) - \delta \\ \vdots \\ E(\hat{\delta}_{P-1,N}) - \delta \end{pmatrix} = \Psi' \begin{pmatrix} b_1 h^{1-\eta} \\ \vdots \\ b_{P-1} h^{(P-1)(1-\eta)} \end{pmatrix} + O(h^{P(1-\eta)})$$

so that

$$c_N' \begin{pmatrix} E(\hat{\delta}_{1N}) - \delta \\ \vdots \\ E(\hat{\delta}_{P-1,N}) - \delta \end{pmatrix} = (h^\eta, \dots, h^{(P-1)\eta})(\Psi^{-1})'\Psi' \begin{pmatrix} b_1 h^{1-\eta} \\ \vdots \\ b_{P-1} h^{(P-1)(1-\eta)} \end{pmatrix} + O(h^{P(1-\eta)})$$

$$= b_1 h + b_2 h^2 + \dots + b_{P-1} h^{P-1} + O(h^{P(1-\eta)})$$

where Ψ^{-1} exists because the constants ψ_ρ, $\rho = 1, \dots, P$, are distinct. Consequently, we have that $\sqrt{N}[E(\bar{\delta}_N) - \delta] = \sqrt{N}([E(\hat{\delta}_N - \Sigma\, c_{\rho N} E(\hat{\delta}_{\rho N}))/(1 - \Sigma\, c_{\rho N})] - \delta) = O(\sqrt{N}\, h^{P(1-\eta)})$. Since $\eta < 1/(k+1)$, $2P(1-\eta) > k+2$, as required. Thus since h is such that $Nh^{2P(1-\eta)} \to 0$, $\lim \sqrt{N}[E(\hat{\delta}_N) - \delta] = 0$. *Q.E.D.*

The generalized jackknife technique differs theoretically from the use of a higher order kernel in that the configuration of local weights can be varied as the sample size increases. When $\eta = 0$ is set in the above formulae, the local weights do not vary and the "jackknifed" estimator just utilizes a higher order kernel as in the text. In particular, in this case the weights $c_{\rho N}$ no longer vary with N, so $c_{\rho N} = c_\rho$, $\rho = 1, \ldots, P-1$. The estimator $\bar{\delta}_N$ can then be written as

$$\text{(A2.5)} \qquad \bar{\delta}_N = \frac{-2}{N} \sum_{i=1}^{N} \left(\frac{\partial \bar{f}_i(x_i)}{\partial x} \right) y_i$$

with the density estimator $\bar{f}_i(x)$ defined as

$$\text{(A2.6)} \qquad \bar{f}_i(x) = \frac{\hat{f}_i(x) - \sum_\rho c_\rho \hat{f}_{\rho i}(x)}{1 - \sum_\rho c_\rho} = \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{1}{h} \right)^k \bar{K}\left(\frac{x - x_j}{h} \right)$$

using the kernel $\bar{K}(u) = [K(u) - \Sigma_\rho c_\rho \psi_\rho^{-k} K(u/\psi_\rho)]/[1 - \Sigma_\rho c_\rho]$. It is easy to verify that $\bar{K}(.)$ is a higher order kernel (obeying Assumption 6). Consequently, beginning from a standard positive kernel (K here), equation (A2.6) can be used to construct a higher order kernel, as we have done in Section 5.

REFERENCES

AMEMIYA, T. (1981): "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 19, 1483–1536.

BICKEL, P. (1982): "On Adaptive Estimation," *Annals of Statistics*, 10, 647–671.

BIERENS, H. J. (1983): "Uniform Consistency of Kernel Estimators of a Regression Function Under Generalized Conditions," *Journal of the American Statistical Association*, 78, 699–707.

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

BILLINGSLEY, P. (1979): *Probability and Measure*. New York: John Wiley and Sons.

EFRON, B. (1982): *The Jackknife, the Bootstrap and Other Resampling Plans*, CBMS Regional Conference Series in Applied Mathematics, 38, Society for Industrial and Applied Mathematics, Philadelphia.

FRYER, M. J. (1977): "A Review of Some Nonparametric Methods of Density Estimation," *Journal of the Institute of Mathematical Applications*, 20, 335–354.

GASSER, T., H. G. MUELLER, AND V. MAMMITZSCH (1985): "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society*, *Ser. B*, 47, 238–252.

HAN, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303–316.

HÄRDLE, W., AND T. M. STOKER (1988): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," MIT, Sloan School of Management, Working Paper No. 2004-88.

HOEFFDING, W. (1948): "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293–325.

HUBER, P. J. (1985): "Projection Pursuit," *Annals of Statistics*, 13, 435–475.

ICHIMURA, H. (1987): "Estimation of Single Index Models," doctoral dissertation, Massachusetts Institute of Technology.

KOLGOMOROV, A. N. (1950): *Foundations of the Theory of Probability* (German edition, 1933). New York: Chelsea.

MANSKI, C. F. (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729–738.

MCFADDEN, D. (1985): "Specification of Econometric Models," Presidential Address to the Fifth World Congress of the Econometric Society, Cambridge, Massachusetts.

PARZEN, E. (1962): "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, 33, 1065–1076.

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

POWELL, J. L. (1987): "Semiparametric Estimation of Bivariate Latent Variable Models," Department of Economics, University of Wisconsin, Social Systems Research Institute Working Paper No. 8704.

QUENOUILLE, M. (1949): "Approximate Tests of Correlation in Time Series," *Journal of the Royal Statistical Society*, Ser. B, 11, 18–84.

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

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

SCHUCANY, W. R., AND J. P. SOMMERS (1977): "Improvement of Kernel Type Density Estimators," *Journal of the American Statistical Associations*, 72, 420–423.

SILVERMAN, B. W. (1978): "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density Function and Its Derivatives," *Annals of Statistics*, 6, 177–184 (Addendum, 1980, *Annals of Statistics*, 8, 1175–1176).

——— (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

SERFLING, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

SPIEGELMAN, C., AND J. SACKS (1980): "Consistent Window Estimation in Nonparametric Regression," *Annals of Statistics*, 8, 240–246.

STOCK, J. H. (1989): "Nonparametric Policy Analysis," forthcoming, *Journal of the American Statistical Association*.

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

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

——— (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *Annals of Statistics*, 8, 1348–1360.

——— (1984): "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *Annals of Statistics*, 12, 1285–1298.

TAPIA, R. A., AND J. R. THOMPSON (1978): *Nonparametric Probability Density Estimation*. Baltimore: John Hopkins University Press.