Econometrica, Vol. 56, No. 4 (July, 1988), 931–954

ROOT-N-CONSISTENT SEMIPARAMETRIC REGRESSION

BY P. M. ROBINSON[1]

One type of semiparametric regression on an $\mathscr{R}^p \times \mathscr{R}^q$-valued random variable (X, Z) is $\beta' X + \theta(Z)$, where β and $\theta(Z)$ are an unknown slope coefficient vector and function, and X is neither wholly dependent on Z nor necessarily independent of it. Estimators of β based on incorrect parameterization of θ are generally inconsistent, whereas consistent nonparametric estimators deviate from β by a larger probability order than $N^{-1/2}$, where N is sample size. An estimator generalizing the ordinary least squares estimator of β is constructed by inserting nonparametric regression estimators in the nonlinear orthogonal projection on Z. Under regularity conditions $\hat{\beta}$ is shown to be $N^{1/2}$-consistent for β and asymptotically normal, and a consistent estimator of its limiting covariance matrix is given, affording statistical inference that is not only asymptotically valid but has nonzero asymptotic first-order efficiency relative to estimators based on a correctly parameterized θ. We discuss the identification problem and $\hat{\beta}$'s efficiency, and report results of a Monte Carlo study of finite-sample performance. While the paper focuses on the simplest interesting setting of multiple regression with independent observations, extensions to other econometric models are described, in particular seemingly unrelated and nonlinear regressions, simultaneous equations, distributed lags, and sample selectivity models.

KEYWORDS: Regression, semiparametric model, kernel nonparametric estimators, root N-consistent estimation, central limit theorem, SUR model, linear simultaneous equations, distributed lags, heteroskedasticity, sample selectivity.

1. INTRODUCTION

STATISTICAL INFERENCE on a multidimensional random variable commonly focuses on functionals of its distribution that are either purely parametric or purely nonparametric. A reasonable parametric model affords precise inferences, a badly misspecified one, possibly seriously misleading ones, while nonparametric modeling is associated both with greater robustness and lesser precision. An intermediate strategy employs a semiparametric form, such as the regression function

$$E(Y|X, Z) = \beta' X + \theta(Z) \quad \text{almost surely (a.s.)}, \tag{1.1}$$

where (X, Y, Z) is an $\mathscr{R}^p \times \mathscr{R} \times \mathscr{R}^q$-valued observable random variable, β is a $\mathscr{R}^p$-valued unknown parameter, and θ is an unknown real function. In (1.1), X, Z, and β are column vectors and the prime indicates transposition. As usual, (1.1) might be the outcome of logging a multiplicative model.

Versions of (1.1) have been studied by Cosslett (1984), Shiller (1984), Wahba (1984, 1985), Stock (1985), Engle et al. (1986), N. Heckman (1986), Rice (1986), Schick (1986). The statistical objectives in these papers vary, as do the motivating applications. In most, though not all, of them Z is a scalar nonstochastic design variable, typically a time index. Our own aim is precise estimation of β when Z

[1] This article is based on research funded by the Economic and Social Research Council (ESRC) reference number: B00232156. I thank Miguel Delgado for carrying out the simulations reported in Section 6, and two referees for many incisive and constructive comments which have stimulated substantial improvements. A previous version was circulated under the title "Adaptive Semiparametric Regression."

is stochastic and of arbitrary dimension, indeed the value of q nontrivially influences our methodology and theory. The components of β may have interesting economic significance, and some hypotheses of interest may be expressible purely in terms of β, in which event the building of a full parametric model may be of secondary importance. Good estimates of β can assist also in parameterizing θ. We picture a practitioner faced with a large cross-sectional data set including many candidate explanatory variables, who on the basis of economic theory or past experience with similar data feels able to parameterize only some of them. Very crudely, (1.1) describes a qualitative unevenness in prior information. It is possible also to rationalize (1.1) as emerging from some econometric models involving latent variables: extending models developed by J. Heckman (1976) and others, a dependent variable is censored or truncated when a latent variable of possibly unknown distributional form exceeds a (possibly unknown) function of Z; extending a model of Zellner (1970), a linear regression includes both observed and latent variables, where the latter are an unknown function of Z. It is also possible to interpret β as the coefficients of the "surprise" component of X, that is the part that cannot be predicted using Z. Both (1.1) and the conditions we impose on it are restrictive in terms of direct applications, but we also describe how some of these conditions might be relaxed and how more general semiparametric models than (1.1) might be estimated.

Under regularity conditions, ordinary least squares (OLS) regression of Y on X alone consistently and efficiently estimates β when $E(X\theta(Z)) = 0$, as when $E(X) = 0$ and X and Z are statistically independent. Such orthogonality is present in certain experimental designs and models containing dummy variables, as well as in some modeling strategies in which Z is not fully or parsimoniously specified, for example orthogonal polynomial and trigonometric regression. Orthogonality can be checked, but it is exceptional, particularly when the explanatory variables include stochastic ones or are large in number. The bias of OLS in the presence of a nonorthogonal omitted variable is explained in elementary econometric textbooks. In much applied work there is an understandable tendency to include candidate explanatory variables in an *ad hoc*, typically linear, fashion, resulting again in biased estimation. Rigorous statistical analysis of parametric estimators in the presence of model misspecification is possible; under typical regularity conditions OLS estimators of β based on incorrect parameterization of θ are asymptotically normal about $\beta + B$ after $N^{1/2}$ norming, where N is the number of observations and the "asymptotic bias" B reflects the unknown θ (see, e.g., White (1982)). Some analysis of B may be possible, allowing speculation about the direction of bias and the signs of β's elements relative to $\beta + B$'s. The omission of many variables, or a "large" discrepancy between the true θ and the misspecified one, does not necessarily result in incorrect conclusions. On the other hand some applied studies indicate high sensitivity of parameter estimators to misspecification of the rest of the model. Automatic or semi-automatic algorithms help bridge the gap between theory and model specification (see, e.g., Amemiya (1980), Stone (1981), and references therein). For example, stepwise regression selects a parsimonious model with good explanatory power while keeping some variables (i.e., X) in the regression irrespective of their

t ratios, though it searches only over linear models. Specification tests are available, but failure to reject correct specification does not necessarily inspire confidence in the null hypothesis, and rejection necessitates continuing the model search.

Consistency for β in the presence of unknown θ is possible, however. Perhaps the most obvious source is nonparametric estimation of $e(x, z) = E(Y|X = x, Z = z)$ at a point (x, z). Let $\hat{e}(x, z)$ be (say) a Nadaraya–Watson kernel estimator of $e(x, z)$ with differentiable kernel (see, e.g., Prakasa Rao (1983, pp. 33–37, 180–200, 239–247, and Section 2 below)); when X and Z do not overlap, $\hat{e}_x = (\partial/\partial x)\hat{e}(x, z)$ estimates β consistently under quite general conditions; see, e.g., Schuster and Yakowitz (1979). Unfortunately $\hat{e}$ and $\hat{e}_x$ are not $N^{1/2}$-consistent, because the asymptotically correct centering at β is due to a "bandwidth" parameter approaching 0, with the effect that, asymptotically, only a vanishingly small proportion of the data, "near" (x, z), is used. Indeed, the greater $p + q$, the further we fall short of $N^{1/2}$-consistency, and $\hat{e}_x$ converges even slower than $\hat{e}$; Stone (1982) discusses optimal rates of convergence in nonparametric regression and its derivatives. Estimators that are consistent but not $N^{1/2}$-consistent generate inferences which, though asymptotically valid, have zero efficiency relative to ones based on $N^{1/2}$-consistent estimators, and while the latter comparison presents an exaggeratedly pessimistic impression of the finite-sample reality, it is debatable whether nonparametric estimators should necessarily be preferred to the "$N^{1/2}$-inconsistent" ones based on incorrectly parameterizing θ. Averaging $\hat{e}_x$ over n (x, z)-values only improves rates of convergence if n increases with N, for example

$$\beta^* = \sum_{i=1}^{N} \hat{e}_x(x_i, z_i)\omega_i, \tag{1.2}$$

where x_i, z_i might be either the observed X's and Z's or a sequence of representative design points, and the ω_i are probability weights, e.g., $\omega_i \equiv N^{-1}$. (It seems β^* is $N^{1/2}$-consistent for β under suitable conditions, and thus competitive with the estimator $\hat{\beta}$ developed below. One might establish β^*'s limiting distribution and compare its efficiency with $\hat{\beta}$'s.)

Other modifications of nonparametric regression should be mentioned. Elbadawi et al. (1983) and Gallant (1985) approximate their models by infinite series, the early terms representing the parametric part (our $\beta' X$), the remaining ones (a trigonometric expansion) representing the nonparametric part (our θ). The hope is that few of the latter terms will be required, and that β will be estimated with good precision. However, β is not really on a different footing from the coefficients of the trigonometric expansion, and consistency relies on the number of terms in the series, hence the number of parameters, going slowly to infinity with N. While the estimators of Elbadawi et al. (1983) and Gallant (1985) might well be better in finite samples than pure nonparametric ones, they converge slower than $N^{1/2}$ unless the true regression is approximated at a fast enough rate as $N \to \infty$. (Actually, identification of β requires strong restrictions on θ; see Section 4 below.) Stone's (1982, 1985) results imply that nonparametric

estimators exploiting the additive structure of (1.1) can achieve faster rates of convergence than pure nonparametric regression on X and Z, but his estimators do not exploit the partial parameterization of (1.1), and fall short of $N^{1/2}$-consistency. Projection pursuit regression (Friedman and Stuetzle (1981)) entails some structural restriction of θ, and it is not clear whether it can produce $N^{1/2}$-consistency.

In most of the earlier work relating to (1.1) that was referenced above, $N^{1/2}$-consistency of estimation of β is not established, indeed the emphasis is sometimes as much if not more on estimating θ. The exceptions are N. Heckman (1986) and Rice (1986), who assume Z is a scalar nonstochastic design variable on the unit interval, the "observations" on which get dense as $N \to \infty$, and Schick (1986), who assumes Z is a scalar uniform random variable. Our setting of stochastic multi-dimensional Z, of quite general distributional form, is more suited to econometric applications. Like N. Heckman and Schick we establish not only $N^{1/2}$-consistency but asymptotic normality of our estimator (which differs from theirs and Rice's), and also we give a consistent estimator of the covariance matrix in the limiting distribution, providing the usual basis for large-sample interval estimation and hypothesis testing. The only information on finite-sample properties we present is the outcome of some Monte Carlo simulations.

We compare and contrast our problem and results with ones in the "adaptive estimation" literature. Authors such as Bickel (1982) and Manski (1984) presented asymptotically efficient estimators of linear and nonlinear regression estimators in the presence of residuals of unknown distributional form, while Carroll (1982), Robinson (1985) presented regression estimators that achieve the asymptotic Gauss–Markov bound in the presence of residuals suffering from heteroskedasticity of unknown form. Like these authors, we insert nonparametric shape estimators of the nonparametric component in a standard "parametric" estimator. Unlike them, we are unable to claim efficiency of our semiparametric estimator, since the "orthogonality" between the parametric and nonparametric components of their models (see Begun et al. (1983)) is in general lacking in ours, and we merely isolate some parametric θ for which our approach happens to be as efficient as one which uses information on θ's form.

2. ESTIMATOR OF β

The model (1.1) implies that $Y - E(Y|Z) = \beta'(X - E(X|Z)) + U$, where $E(U|X, Z) = 0$ a.s., suggesting that estimators of the regression functions $E(X|Z)$, $E(Y|Z)$ be inserted prior to application of a standard rule, such as no-intercept OLS. While a variety of nonparametric regression estimators is available (two reviews are Prakasa Rao (1983, pp. 239–256), Collomb, (1985)), the technical difficulties described in Section 3 below are conveniently overcome by a subset of the Nadaraya–Watson kernel estimators. Introduce even functions k: $\mathscr{R} \to \mathscr{R}$ and K: $\mathscr{R}^q \to \mathscr{R}$ related by

$$K(z) = \prod_{i=1}^{q} k(z_i), \tag{2.1}$$

where z_i is z's ith element. Let a be a positive constant. For a vector-valued sequence $A_1, \ldots, A_N$, introduce the notation

$$(2.2) \qquad \bar{A}_i = (Na^q)^{-1} \sum_{j=1}^{N} A_j K_{ij}, \quad K_{ij} = K\left(\frac{Z_i - Z_j}{a}\right),$$

and define, with $1_j \equiv 1$, $\hat{f}_i = \bar{1}_i$, $\hat{X}_i = \bar{X}_i/\hat{f}_i$, $\hat{Y}_i = \bar{Y}_i/\hat{f}_i$. Under conditions set out in Section 3, $\hat{f}_i$ "estimates" $f(Z_i)$, the probability density function (pdf) of Z with random argument Z_i, while $\hat{X}_i$ and $\hat{Y}_i$ "estimate" $E(X_i|Z_i)$ and $E(Y_i|Z_i)$. As in some other applications of kernel regression estimators, $\hat{X}_i$ and $\hat{Y}_i$ cause technical difficulty owing to the random denominator $\hat{f}_i$, which can be small; we "trim" out small $\hat{f}_i$ as do, e.g., Bickel (1982), Manski (1984). For constant $b > 0$ define $I_i = I(|\hat{f}_i| > b)$, where I is the usual indicator function; then estimate β by

$$(2.3) \qquad \hat{\beta} = S_{X-\hat{X}}^{-1} S_{X-\hat{X}, Y-\hat{Y}},$$

where for scalar or column-vector sequences A_i and B_i, we define $S_{AB} = N^{-1}\sum_{i=1}^{N} A_i B_i' I_i$ and $S_A = S_{AA}$. Notice that

$$(2.4) \qquad S_{A-\hat{A}, B-\hat{B}} = (A_1, \ldots, A_N) \times \{\operatorname{diag}(I_1, \ldots, I_N) DD' \operatorname{diag}(I_1, \ldots, I_N)\}(B_1, \ldots, B_N)',$$

where D is the N-rowed identity matrix minus the matrix with (i, j)th element $K_{ij}/\hat{f}_j$, so $\hat{\beta}$ has a generalized least squares (GLS) interpretation, as well as a no-intercept OLS one. Because $K_{ij} = K_{ji}$, $K_{ii} \equiv K(0)$, only $\frac{1}{2}N(N-1)$ distinct K_{ij} need be computed; nevertheless (2.3) entails $O(p^2qN^2)$ operations.

If the $\hat{X}_i$, $\hat{Y}_i$ are replaced in $\hat{\beta}$ by the linear OLS predictors of the X_i, Y_i, we have the OLS estimator $\tilde{\beta}$, say, that corresponds to taking $\theta(Z)$ linear in Z; indeed if we take $k(u) \propto I(|u| \leqslant 1)$ and a large enough, $\hat{\beta}$ reduces to OLS that assumes $\theta(Z)$ constant. This similarity of $\hat{\beta}$ to a standard parametric estimator (not shared by β^* in (1.2), for example) seems attractive in view of $\tilde{\beta}$'s well known optimality properties, and it extends to the structure of formulae for standard errors (see the theorem in Section 3), the only additional statistic needed to calculate $N^{1/2}(\hat{\beta} - \beta)$'s estimated covariance matrix $\hat{\sigma}^2 S_{X-\hat{X}}^{-1}$ being

$$\hat{\sigma}^2 = S_{Y-\hat{Y}-\hat{\beta}'(X-\hat{X})} = S_{Y-\hat{Y}} + 2S_{Y-\hat{Y}, X-\hat{X}}\hat{\beta} + \hat{\beta}' S_{X-\hat{X}}\hat{\beta},$$

which estimates $\sigma^2 = V(Y|X, Z)$, assuming the residuals from (1.1) are conditionally homoskedastic. The extension of $\hat{\beta}$ to more general semiparametric models is analogous to $\tilde{\beta}$'s in parametric models, as will be indicated in Section 7. $\hat{\beta}$ and $\tilde{\beta}$ differ in $\hat{\beta}$'s use of residuals from the best (in least squares sense) predictors of Y and X given Z, rather than the best linear predictors, and in computational terms the difference is immense, increasing rapidly with N and q. $\hat{\beta}$ is likely to be more expensive of computer time than nonlinear least squares (NLLS) when θ is nonlinear in parameters, though its closed form structure is an advantage, it is straightforward to program, and it avoids the need to choose a vector of starting values for the iterations and the possibilities of slow or nonexistent convergence. To compare with other semiparametric treatments of (1.1), Wahba (1984, 1985),

Shiller (1984), Engle et al. (1986), N. Heckman (1986), and Rice (1986) use spline estimation; Stock (1985) uses (untrimmed) kernel estimation, but his focus is not β; Schick (1986) uses the kernel idea, but his estimator for his version of (1.1) is quite different in form. Comparing $\hat{\beta}$ with $N^{1/2}$-consistent estimators proposed for other problems, Bickel (1982), Manski (1984), Robinson (1987), Powell et al. (1986), Schick (1986), and others, all employ, for technical reasons, an element of "sample-splitting," which in our case might entail replacing N in (2.2) by $M < N$, then constructing $S_{X-\hat{X}}$, $S_{X-\hat{X},Y-\hat{X}}$ by summing only over the remaining $N-M$ observations. By avoiding this device, $\hat{\beta}$ makes fuller use of the data.

The dependence of $\hat{\beta}$ on the user-supplied numbers a and b is an undesirable feature shared with other semiparametric estimators that employ nonparametric "shape" estimation. The Theorem sets conditions on a and b's rate of decay as $N \to \infty$ that are virtually useless to the practitioner. It is not obvious how sensitive $\hat{\beta}$ is to a and b, but the effects of extreme choices, while possibly not as catastrophic as in the case of pure nonparametric estimation, are liable to be serious: "large" a can induce bias, "small" a, imprecision, because $1/a$ can be thought of like the dimensionality of a parameterization of θ; a "large" b loses efficiency, a "small" b allows $\hat{X}_i$ and $\hat{Y}_i$ with small denominators $\hat{f}_i$ to exert undue influence. Automatic methods such as cross-validation offer an alternative to trial-and-error choice of a, and it is easy to suggest suitable cross-validating objective functions, but we will not discuss the details because our theorem unfortunately does not cater to data-driven a to b. In connection with a, when $q > 1$ some refinement in $\hat{\beta}$ is desirable because of likely scale differences in Z's elements, indicating that K's argument in (2.2) should be replaced by $a^{-1}(Z_i - Z_j)$ where a is here either a diagonal or a positive definite matrix (in the latter case K is a more general multivariate function than (2.1)). The conditions on a in our Theorem are straightforwardly generalized in the manner of conditions of Cacoullos (1966) for diagonal a, and Robinson (1983) for matrix a. We have not bothered to treat this extension explicitly because our conditions and proofs are already somewhat complicated, and merely note that it suffices, in the diagonal-a case, for each diagonal element to decay as $N \to \infty$ at the same rate. One alternative to multidimensional a is scaling the Z_i, via the estimated standard deviations or covariance matrix, though our conditions do not automatically require that Z have infinite variance.

Finally, we can use $\hat{\beta}$ to form "estimators" of $\theta(Z_i)$, $\tilde{\theta}(Z_i) = \hat{Y}_i - \hat{\beta}'\hat{X}_i$; predictors of Y_i (conditional on X_i, Z_i), $\tilde{Y}_i = \hat{\beta}'X_i + \tilde{\theta}(Z_i)$; and estimated residuals, $\tilde{U}_i = Y_i - \tilde{Y}_i$, $1 \leqslant i \leqslant N$. (In fact, $\hat{\sigma}^2 = S_{\tilde{U}\cdot}$.) Given (1.1), $\tilde{Y}_i$ and $\tilde{U}_i$ should improve on predictors and residuals based on pure nonparametric regression, though we make no study of their properties.

3. CONDITIONS AND THEOREM

With the definitions $U = Y - \beta'X - \theta(Z)$, $\hat{\theta}_i = \bar{\theta}_i/\hat{f}_i$, $\hat{U}_i = \bar{U}_i/\hat{f}_i$, write

$$Y_i - \hat{Y}_i = \beta'\left(X_i - \hat{X}_i\right) + \left(\theta_i - \hat{\theta}_i\right) + \left(U_i - \hat{U}_i\right). \tag{3.1}$$

The component $\theta_i - \hat{\theta}_i$ of the "residual" in (3.1) presents a bias problem, because it is hard to see how $N^{1/2}$-consistency of $\hat{\beta}$ can be established in the absence of the property $S_{X-\hat{X},\theta-\hat{\theta}} = o_p(N^{-1/2})$. Assuming the conditional expectation $\xi(z) = E(X|Z=z)$ exists, and defining $V = X - \xi(Z)$, it is sufficient that $S_{V-\hat{V},\theta-\hat{\theta}} = o_p(N^{-1/2})$ and $S_{\xi-\hat{\xi},\theta-\hat{\theta}} = o_p(N^{-1/2})$. The last relationship is troublesome to establish. After centering the $\hat{\xi}_i - \xi_i$ and $\hat{\theta}_i - \theta_i$ in $S_{\xi-\hat{\xi},\theta-\hat{\theta}}$ at expectations conditional on the Z_i, it is not difficult to show that the resulting expression is indeed $o_p(N^{-1/2})$ so long as a does not approach O too rapidly as $N \to \infty$, and this type of condition on a is required elsewhere in the proof in any case. However, this centering introduces a term reflecting the bias of the kernel "estimators" $\hat{\theta}_i$ and $\hat{\xi}_i$ of θ_i and ξ_i. Such bias can be made arbitrarily small by setting a small enough, to establish $S_{\xi-\hat{\xi},\theta-\hat{\theta}} \xrightarrow{p} 0$ and eventually $\hat{\beta} \xrightarrow{p} \beta$. However, achieving the more ambitious goals of $S_{\xi-\hat{\xi},\theta-\hat{\theta}} = o_p(N^{-1/2})$, and $N^{1/2}$-consistency of $\hat{\beta}$, simply by making a approach 0 suitably fast as $N \to \infty$ may not be possible because of the aforementioned limitations on a's convergence. In fact, as in much statistical theory for kernel estimators (see, e.g., Cacoullos (1966), Stone (1982)) the upper bound on a's rate of decay as $N \to \infty$ strengthens as the dimensionality q of Z increases, so much so that unless q is suitably small, $N^{1/2}$-consistency requires special measures to ensure an a-sequence satisfying the competing restrictions even exists.

We adopt the "higher-order" kernel approach to bias-reduction proposed by Bartlett (1963) for nonparametric probability and spectral density estimators, since developed by many authors and featured prominently in the kernel literature: a sufficiently smooth function behaves locally like a polynomial of sufficiently high order, and if this property is exploited by a kernel with enough zero "moments," the bias decreases sufficiently rapidly with a.

DEFINITION 1: K_l, $l \geq 1$, is the class of even functions k: $\mathscr{R} \to \mathscr{R}$ satisfying

$$\int_{\mathscr{R}} u^i k(u)\, du = \delta_{i0} \qquad (i = 0, \ldots, l-1), \tag{3.2}$$

$$k(u) = O\left(\left(1 + |u|^{l+1+\varepsilon}\right)^{-1}\right), \quad \text{some} \quad \varepsilon > 0, \tag{3.3}$$

where δ_{ij} is Kronecker's delta.

The requirement that k be bounded and integrate to 1 makes $\hat{f}_i$ a sensible estimator of $f(Z_i)$. For given l satisfying (3.2), (3.3) has a slightly stronger tail condition on k than $\int |u^l k(u)|\, du < \infty$, which is usually employed in the higher-order kernel literature (see, e.g., (23) on p. 44 of Prakasa Rao (1983)), but kernels used in practice usually have compact support or decay exponentially. Some of the kernel literature emphasizes weak conditions on k as a priority, but for implementation it suffices that the conditions admit a convenient k, and practical experience suggests less sensitivity to k than to a. If (3.2) holds for some odd l it holds for $l+1$ also under (3.3). $\mathscr{K}_l$ contains no nonnegative functions when $l \geq 3$, indicating the potential for negative estimates of the density of f, although this seems of little concern in our context. As indicated by a number of authors

(e.g., Prakasa Rao (1983, p. 44)) a $k \in \mathscr{K}_l$ is straightforwardly constructed. Consider, for even $l \geqslant 2$,

$$k(u) = \sum_{j=0}^{1/2(l-2)} c_j u^{2j} \psi(u), \tag{3.4}$$

where ψ is even. Given that we can evaluate the moments $m_{2j} = \int u^{2j} \psi(u)\, du$, $0 \leqslant j \leqslant \frac{1}{2}(l-2)$, as readily we may when $\psi(u) = \frac{1}{2} I(|u| \leqslant 1)$ or $\psi(u) = (2\pi)^{-1/2} \exp(-\frac{1}{2}u^2)$, substitution of the c_j satisfying the linear system of $\frac{1}{2}(l-2)$ simultaneous equations $\sum_{j=0}^{1/2(l-2)} c_j m_{2(i+j)} = \delta_{i0}$, $0 \leqslant i \leqslant \frac{1}{2}(l-2)$, into (3.4) produces a $k \in \mathscr{K}_l$, if $\psi(u) = O((1 + |u|^{2l-1+\varepsilon})^{-1})$.

The classes $\mathscr{K}_l$ confer increasingly small bias on nonparametric kernel estimators as l increases, but also increasingly large asymptotic variance, the latter varying directly with $\int k(u)^2\, du$. However, the asymptotic distribution in the Theorem below is independent of k, detecting no advantage or disadvantage in a $\mathscr{K}_l$ when l is chosen arbitrarily greater than required. Nevertheless, in finite samples $\hat{\beta}$ may inherit variance properties of the kernel estimators from which it is formed, as might be revealed by a closer approximation to the distribution of $\hat{\beta}$. Thus, while increasing l cannot shrink, and may well widen, the band of a-sequences satisfying our Theorem, we caution against too generous a choice of l. It is interesting that whereas the classes $\mathscr{K}_l$ play useful roles of bias-reduction and of widening the spectrum of admissible bandwidths in nonparametric estimation, they are decisive in our problem, which requires dealing with a greater ($N^{1/2}$) norming than in the central limit theorem for q-variate nonparametric estimators (($Na^q)^{1/2}$). A related bias-reduction device is the "generalized jacknife" method suggested by Schucany and Sommers (1977) for kernel density estimators, later developed by other authors, which would require $q+1$ bandwidth numbers to be selected, instead of our single a. In fact, Schucany and Sommers' approach is used in a different semiparametric estimation problem from ours by Powell et al. (1986), who extend Stoker's (1986) work on the model $E(Y|X) = F(\beta' X)$ where F is unknown, and there are no functional relationships between components of X. Stoker independently rediscovered a result used previously by Beran (1977) and Cox (1985) in other semiparametric and nonparametric problems, that $h(X)$ and X's score function have covariance $E((\partial/\partial X)h(X))$, to suggest a simple estimator of β up to undetermined scale that depends on finite parameterization of X's score function. In a spirit similar to (1.2), Powell et al. relax the latter requirement by using nonparametric kernel estimation of the derivative of X's density, solving a bias problem analogous to ours via an extension of Schucany and Sommers' approach.

The potential of the $\mathscr{K}_l$ classes to produce $N^{1/2}$-consistency, or to widen the band of admissible a-sequences, will not be realized unless the functions θ, ξ, and f are collectively sufficiently smooth, and all else being equal it seems reasonable to suppose that the smoother they are, the better $\hat{\beta}$ will be. Let $|\cdot|$ denote Euclidean norm.

DEFINITION 2: $\mathcal{G}_\mu^\alpha$, $\alpha > 0$, $\mu > 0$, is the class of functions g: $\mathcal{R}^q \to \mathcal{R}$ satisfying: g is $(m-1)$-times partially differentable, for $m - 1 \leqslant \mu \leqslant m$ and all z; for some $\rho > 0$, $\sup_{y \in \mathcal{S}_{z\rho}} |g(y) - g(z) - Q(y, z)| / |y - z|^\mu \leqslant h(z)$ for all z, where $\mathcal{S}_{z\rho} = \{y\colon\ |y - z| < \rho\}$; $Q = 0$ when $m = 1$; Q is a $(m-1)$th-degree homogeneous polynomial in $y - z$ with coefficients the partial derivatives of g at z of orders 1 through $m - 1$ when $m > 1$; and $g(z)$, its partial derivatives of order $m - 1$ and less, and $h(z)$, have finite αth moments.

The functions in $\mathcal{G}_\mu^\alpha$ are thus expanded in a Taylor series with a local Lipschitz condition on the remainder, (α, μ) depending simultaneously on smoothness and moment properties. Bounded functions in $\mathrm{Lip}(\mu)$ (the Lipschitz class of degree μ) for $0 < \mu \leqslant 1$ are in $\mathcal{G}_\mu^\infty$; for $\mu > 1$, $\mathcal{G}_\mu^\infty$ contains the bounded and $(m-1)$-times boundedly differentiable functions whose $(m-1)$th partial derivatives are in $\mathrm{Lip}(\mu - m + 1)$. In applying $\mathcal{G}_\mu^\alpha$ to f, we take $\alpha = \infty$, but we allow for $\alpha < \infty$ in Definition 2 because we have no wish to require that Z, ξ or θ are a.s. bounded. For example, a degree-m polynomial in Z is in $\mathcal{G}_\infty^\alpha$ when $E|Z|^{m\alpha} < \infty$.

THEOREM: *Let the following conditions hold*: (i) (X_i, Y_i, Z_i), $i = 1, 2, \ldots$, *are independent and distributed as* (X, Y, Z); (ii) (1.1) *is true*; (iii) U *is independent of* X, Z; (iv) $E(U^2) = \sigma^2 < \infty$; (v) $E|X|^4 < \infty$; (vi) Z *admits a pdf* $f \in \mathcal{G}_\lambda^\infty$, *for some* $\lambda > 0$; (vii) $\xi \in \mathcal{G}_\mu^2$, *for some* $\mu > 0$; (viii) $\theta \in \mathcal{G}_\nu^4$, *for some* $\nu > 0$; (ix) *as* $N \to \infty$, $Na^{2q}b^4 \to \infty$, $Na^{2\min(\lambda+1,\mu)+2\min(\lambda+1,\nu)}b^{-4} \to 0$, $a^{\min(\lambda+1,2\lambda,\mu,\nu)}b^{-2} \to 0$, $b \to 0$; (x) $k \in \mathcal{K}_{\max(l+m-1,\, l+n-1)}$, *for the integers* l, m, n *such that* $l - 1 < \lambda \leqslant l$, $m - 1 < \mu \leqslant m$, $n - 1 < \nu \leqslant n$. *Then the condition*

$$(3.5)\qquad \Phi \equiv E\left[\{X - E(X|Z)\}\{X - E(X|Z)\}'\right] \text{ is positive definite}$$

is necessary and sufficient for $N^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2\Phi^{-1})$ *and* $\hat{\sigma}^2 S_{X-\hat{X}}^{-1} \xrightarrow{p} \sigma^2\Phi^{-1}$.

The proof of this theorem is presented in the form of Appendices. Notice that (ix) and (x) are to be satisfied simultaneously, for $\lambda, \mu, \nu, l, m, n$ satisfying the stated inequalities, so that, for example, when $k \in \mathcal{K}_2$ only, the lower bounds on a's rate of decay are no better than $Na^8b^{-4} \to 0$, $a^2b^{-2} \to 0$, no matter the degree of smoothness prevailing. While (ix) prevents b from converging to O too fast, there is nothing to stop it converging arbitrarily slowly. A necessary condition for reconciling the components of (ix) is

$$(3.6)\qquad \lambda > \tfrac{1}{2}q - 1, \quad \lambda + \mu > q - 1, \quad \lambda + \nu > q - 1, \quad \mu + \nu > q.$$

Conditions (vi)–(viii) are complicated but it is not hard to find examples satisfying them, as the discussion of Definition 2 indicated, and some simple ones are used in the simulations of Section 6. Although some smoothness in f, ξ, θ is needed even when $q = 1$, this need not amount to differentiability, and for other smallish values of q (vi)–(viii) may not be excessive. Very smooth f, $\xi(f, \theta)$ can compensate for a not-very-smooth $\theta(\xi)$. In view of (3.6), a necessary condition for (x) is that $k \in \mathcal{K}_{q-1}$. Given sufficient smoothness in f, ξ and θ, when $q \leqslant 3$, $\mathcal{K}_2$ (which includes all even, bounded pdfs with finite fifth moments) admits

suitable a and b sequences, although the greater the order of $\mathscr{K}$ the greater the range of a, b sequences satisfying (x). The main restriction on the explanatory variables is that discrete components of Z (but not X) are ruled out. In fact it is not difficult to allow Z to have components that are discrete with finite support, and we can see how to achieve some further relaxation when $q \leqslant 3$, as well as a variety of trade-offs between conditions, but still the difference between our conditions on explanatory variables and unknown functional forms and the weaker ones of Robinson (1987) for a different semiparametric regression problem is considerable, and warrants further investigation.

4. IDENTIFICATION

The necessary and sufficient condition (3.5) is an identification condition, unfortunately a very restrictive one. It prohibits β from including an "intercept" coefficient; only "slope" coefficients can be estimated. This is less a drawback of $\hat{\beta}$ than a consequence of the generality of the semiparametric model (1.1): $\beta'X + \theta(Z) = (\alpha + \beta'X) + \{\theta(Z) - \alpha\}$, for all α, and $\theta(Z)$ may be redefined as $\theta(Z) - \alpha$. It is possible to identify α if the model is restricted further; for example Schick (1986) assumes θ integrates to zero and Z is uniformly distributed, and in fact considers the efficient estimation of α under further conditions.

More generally, (3.5) prevents any element of X from being a.s. perfectly predictable by Z in the least squares sense. This rules out such important cases as an unknown regression function of a single variable Z, with $\beta'X$ representing a truncated Taylor expansion and θ taking care of the remainder (c.f. White, 1980). Such models could be said to be more nonparametric than semiparametric (they are "seminonparametric" in Gallant's (1985) terminology), and again it is the unrestricted nature of θ which excludes them, not our method of estimation, because $\beta'X + \theta(Z) = \{\beta'X + \eta(Z)\} + \{\theta(Z) - \eta(Z)\}$, for all $\eta(Z)$. While β is not identified in the linear model

$$Y = \alpha + \beta'X + \gamma'Z + U \tag{4.1}$$

if any X element is linear in Z, (1.1) forbids more general forms of dependence, and it is only to be expected that this more loosely specified model would entail stronger identification conditions. Notice that (nonlinear) functional relationships among X elements are not ruled out. Notice also that identification may be possible even if X uniquely defines Z, when the converse is not true: for example, if $p = q = 1$ and $Z = X^2$, then $\xi(z) = \sqrt{z}(1 - 2P)$ and $\Phi = 4P(1 - P)E(X^2)$, where $P = P(X \leqslant 0)$, so it is necessary and sufficient that X be neither nonnegative nor nonpositive. Given that no elements of the prediction error $X - \xi(Z)$ are a.s. zero, the additional condition implied by (3.5) is their lack of multicollinearity, which fails if X itself is collinear.

5. EFFICIENCY OF $\hat{\beta}$

Suppose θ is a known, partially differentiable function of Z and of a r-dimensional unknown parameter vector δ, $\theta(Z; \delta)$. If (β^+, δ^+) is a NLLS estimator of

(β, δ), then it is well known that under regularity conditions the covariance matrix in the limiting normal distribution of $N^{1/2}(\beta^+ - \beta)$ is

$$(5.1) \qquad \sigma^2\left(C_X - C_{X\partial}C_\partial^{-1}C_{\partial X}\right)^{-1},$$

where $C_X = E(XX')$, $C_{X\partial} = E\{X(\partial/\partial\delta)'\theta(Z;\delta)\}$, $C_\partial = E\{(\partial/\partial\delta) \times \theta(Z;\delta)(\partial/\partial\delta)'\theta(Z;\delta)\}$. Note that (5.1) is the asymptotic Gauss–Markov bound in case (4.1), and in the nonlinear case is minimal with respect to the class of weighted NLLS estimator, when U is conditionally homoskedastic, as we have assumed.

By the Schwarz inequality, $(5.1) \leqslant \sigma^2\Phi^{-1}$, so β^+ is at least as efficient as $\hat{\beta}$. There is equality between $\sigma^2\Phi^{-1}$ and (5.1) if and only if $E\{E(X|Z)E(X|Z)'\} = C_{X\partial}C_\partial^{-1}C_{\partial X}$, that is if

$$(5.2) \qquad E(X|Z) = \Gamma(\partial/\partial\delta)\theta(Z;\delta), \quad \text{a.s.},$$

for some $p \times r$ matrix Γ. Of course (5.2) includes the case of $\theta(Z)$ actually constant, so that at least no efficiency has been lost by our elaborate estimator $\hat{\beta}$ relative to OLS estimation of slope coefficients, which is all that is required then. If, more generally, $\theta(Z;\delta) = \alpha + \gamma'Z$, (5.2) can be written

$$(5.3) \qquad E(X|Z) = \Gamma_1 + \Gamma_2 Z, \quad \text{a.s.},$$

the necessary and sufficient condition for $\hat{\beta}$ to attain the Gauss–Markov bound with respect to (4.1). It immediately follows that $\hat{\beta}$ is then also asymptotically as efficient as the maximum likelihood estimator based on (4.1) when the distribution of Y given X, Z is normal. Often (5.3) is assumed in parametric estimation of "surprise" models.

The intuition behind efficiency condition (5.3) is seen by rewriting (4.1) as $Y = (\alpha + \beta'\Gamma_1) + \beta'V + (\beta'\Gamma_2 + \gamma')Z + U$, under (5.3). By construction, Z and V are orthogonal and $E(V) = 0$, so were V observable, regressing Y on V would asymptotically efficiently estimate β; the Theorem demonstrates that $\hat{\beta}$ is asymptotically as efficient as this regression. When $\hat{\beta}$ is not efficient in this sense, and indeed no element of the vector equality (5.3) is true, an approximate level-α Hausman (1978)-type specification test consists of rejecting (4.1) if (with $\overline{Z}_i' = (1, Z_i')$)

$$(5.4) \qquad N\hat{\sigma}^{-2}\left(\hat{\beta} - \tilde{\beta}\right)'\Big[S_{X-\hat{X}}^{-1} - N\Big\{\sum X_i\tilde{Z}_i'\left(\sum \tilde{Z}_i\tilde{Z}_i'\right)^{-1}\sum \tilde{Z}_i X_i'\Big\}^{-1}\Big]^{-1}\left(\hat{\beta} - \tilde{\beta}\right)$$

exceeds the $100(1-\alpha)$th percentile of the χ_p^2 distribution. If desired, $\hat{\sigma}^2$ could be replaced in (5.4) by the residual mean square in the OLS regression fit of (4.1). Computationally, (5.4) is far more expensive than statistics based on parametric omitted variables, and it should be less powerful in the direction of such alternatives, but if $\hat{\beta}$ has already been computed (5.4) entails little extra work and might be expected to enjoy reasonable power against a range of alternatives.

Necessary and sufficient conditions on X and Z for (5.3) are given by Kagan et al. (1973, pp. 11, 12). One interesting case of (5.3) is (X, Z) multivariate

normal, but normality is not necessary, except for special structures (Kagan et al. (1973, Sec. 10.5)). An estimation strategy is suggested in relation to a tentatively specified linear regression model

$$(5.5) \qquad E(Y|W) = \alpha + \gamma' W$$

where γ and W are $r \times 1$. Denoting jth element by subscript j, form $\hat{\gamma}$ such that $\hat{\gamma}_j$ is $\hat{\beta}$ with $p = 1$, $q = r - 1$; let $X = W_j$ and Z be W with W_j deleted. Then $\hat{\gamma}_j$ estimates γ_j robustly in the sense of being $N^{1/2}$-consistent even if the functional dependence on the W_k, $k \neq j$, has been misrepresented by (5.5). Moreover, if (5.5) is correct, $\hat{\gamma}$ is as efficient asymptotically as the OLS estimator of γ if the regression of W_j on all W_k, $k \neq j$, is linear, for each j, for example if W is normal.

6. SIMULATIONS

Finite-sample theory for semiparametric estimators such as $\hat{\beta}$ is not on the horizon, even under much more precise distributional assumptions than ours; indeed little is known about the finite-sample distribution of the nonparametric regression estimators of which $\hat{\beta}$ is composed. To gain some idea of finite-sample performance and the influence of such factors as dimensionality of Z and order of kernel, a small simulation study was conducted, in double precision FORTRAN on the University of London's Amdahl computer. Such vast variation of design is possible that the results are in no sense representative, and we would only wish to add that $\hat{\beta}$ is invariant to location shifts in X, Y and Z, while $\hat{\beta} - \beta$ (on which all the summary statistics we report depend) is invariant to β. Four different models with varying q $(= 1, 5, 10)$ and θ (and satisfying the regularity conditions of the Theorem) were selected, and three sample sizes, $N = 25$, 50, and 200. Because computing time varies greatly with N and q, as indicated above, the numbers of replications were on a sliding scale, from 100,000 when $q = 1$ or 5 and $N = 25$, to a mere 1000 when $q = 5$ or 10 and $N = 200$. We obtained a and b by inspecting the results for various values used on training samples, the only constraint that was initially imposed being that a and b be monotonic over N and q in a fashion that roughly reflects condition (ix) of the Theorem. There was no serious attempt at optimal choice but we avoided values which entailed extreme bias or variability, and used the same values for model (4.1) and model (6.1) below. We report results only for three different kernels, selected in order to gauge the implications of kernel order. Kernels 1–3 are in $\mathscr{K}_2$, $\mathscr{K}_4$, and $\mathscr{K}_6$ respectively, and given by (3.4) with $l = 2, 4, 6$, respectively, and $\psi(u) = (2\pi)^{-1/2} \exp(-\frac{1}{2}u^2)$. Most of the calculations were also repeated for the three corresponding kernels formed from $\psi(u) = \frac{1}{2} I(|u| \leqslant 1)$; these are quicker to compute, but having compact support, unless N and/or a are large enough relative to q it does happen on occasion that $\hat{X}_i \equiv X_i$, when the estimator breaks down.

In (4.1) we took X and Z to be scalar random variables from a bivariate normal population with zero means, variances 4 and 3, and covariance 2; U to be

TABLE I

MODEL (4.1)

N	a	b	r	$\hat{\beta}(1)$	$\hat{\beta}(2)$	$\hat{\beta}(3)$	
25	1.65	.01	10^5	−.1213	−.0254	−.0095	BIAS
				.9141	.9750	.9128	$\sqrt{}$EFFICIENCY
50	1.25	.005	5×10^4	−.0697	−.0094	−.0040	BIAS
				.9020	.9626	.9376	$\sqrt{}$EFFICIENCY
200	0.75	.001	10^4	−.0161	−.0007	.0003	BIAS
				.9607	.9722	.9696	$\sqrt{}$EFFICIENCY

standard normal; and $\alpha = \beta = \gamma = 1$. Subroutine G05DDF from the NAG library generated the observations. Let $\hat{\beta}$ be the OLS (i.e., maximum likelihood) estimator of β based on the true model: for (4.1) it is unbiased when $N \geqslant 3$. (Intercept OLS of Y and X alone, denoted $\bar{\beta}$, is inconsistent.) While $\hat{\beta}$ is not unbiased for finite N, it is as efficient as $\tilde{\beta}$ in (4.1) (see Section 5), so these are relatively favorable circumstances for $\hat{\beta}$, especially as $q = 1$ only. The results are presented in Table I, where r is the number of replications. In each table we report the simulation biases of the $\hat{\beta}$ estimates, formed from kernels 1–3, and headed $\hat{\beta}(i)$, $i = 1, 2, 3$, and the *ratio* of $\tilde{\beta}$'s simulation standard deviation to $\hat{\beta}(i)$'s, called $\sqrt{}$efficiency. The biases in Table I are mostly negative, and decrease a bit as kernel order increases. The $\sqrt{}$efficiencies are not as good as the asymptotic ones.

Table II contains corresponding results for the model

$$(6.1) \qquad Y = \alpha + \beta X + \gamma Z^2 + U,$$

under the same specification as before. Now $\hat{\beta}$ is no longer as efficient asymptotically as OLS $\tilde{\beta}$ based on (6.1) (its asymptotic relative efficiency is 2/3). ($\bar{\beta}$ happens to be consistent, unbiased when $N \geqslant 2$, and asymptotically efficient for this model.) The biases are all positive and increase a bit as kernel order increases. The $\sqrt{}$efficiencies are sometimes better, sometimes worse, than the asymptotic ones, though not surprisingly uniformly worse than Table I's.

Finally we tested the method against Z's of much higher dimension, extending (6.1) to

$$(6.2) \qquad Y = \alpha + \beta X + \sum_{j=1}^{q} \gamma_j Z_{(j)}^2 + U,$$

TABLE II

MODEL (6.1)

N	a	b	r	$\hat{\beta}(1)$	$\hat{\beta}(2)$	$\hat{\beta}(3)$	
25	1.65	.01	10^5	.0188	.0191	.0204	BIAS
				.7862	.7761	.7271	$\sqrt{}$EFFICIENCY
50	1.25	.005	5×10^4	.0075	.0079	.0083	BIAS
				.8754	.8606	.8375	$\sqrt{}$EFFICIENCY
200	0.75	.001	10^4	.0018	−.0019	.0019	BIAS
				.9356	.9299	.9201	$\sqrt{}$EFFICIENCY

TABLE III

MODEL (6.2), $q = 5$

N	a	b	r	$\hat{\beta}(1)$	$\hat{\beta}(2)$	$\hat{\beta}(3)$	
25	3	.0001	10^5	.2774	.1600	.1276	BIAS
				.3638	.3527	.3246	$\sqrt{}$EFFICIENCY
50	2.4	.00005	25×10^3	.1743	.0988	.0813	BIAS
				.3743	.3716	.3231	$\sqrt{}$EFFICIENCY
200	1.5	.00001	10^3	.0693	.0399	.0285	BIAS
				.4349	.4245	.3653	$\sqrt{}$EFFICIENCY

TABLE IV

MODEL (6.2), $q = 10$

N	a	b	r	$\hat{\beta}(1)$	$\hat{\beta}(2)$	$\hat{\beta}(3)$	
25	4.5	10^{-8}	25×10^3	.6688	.3559	.2523	BIAS
				.2788	.2231	.1941	$\sqrt{}$EFFICIENCY
50	3.25	5×10^{-9}	10^4	.3357	.1621	.1070	BIAS
				.2181	.1972	.1726	$\sqrt{}$EFFICIENCY
200	2.25	10^{-9}	10^3	.1663	.0728	.0485	BIAS
				.2081	.2039	.1785	$\sqrt{}$EFFICIENCY

where α, β and the γ_j are all 1; U is as before; and X and the $Z_{(j)}$ are equicorrelated identically distributed $N(1,3)$ variables, with correlation $2/3$. The asymptotic relative efficiency of $\hat{\beta}$ to $\tilde{\beta}$ increases from $8/9$ when $q = 1$, to 1 as $q \to \infty$. Because $E(Z_{(j)}) \neq 0$, $\bar{\beta}$ is inconsistent. Results for cases $q = 5$ and 10 are presented in Tables III and IV. The biases are uniformly positive and mostly very bad, especially in Table IV, though bias does improve materially with increase in N and, more interestingly, with kernel order. The role played by the higher-order kernels in the asymptotic theory does therefore seem to have implications for finite-sample practice. However, they do produce larger variances, as surmised in Section 3, though even for kernel 1 the $\sqrt{}$efficiencies are anything from half (when $q = 5$) to less than a quarter (when $q = 10$) of that predicted by asymptotic theory. These figures are only slightly influenced by $\tilde{\beta}$'s variances being mostly a bit lower than the asymptotic ones. Evidently the nonparametric kernel estimates are so bad for these sample sizes and high-dimensional Z's as to seriously inflate $\hat{\beta}$'s variability.

7. EXTENSIONS

We indicate some extensions of our semiparametric model and estimator that are of possible econometric interest, without giving full details or regularity conditions (which have not been worked out), but noting limitations as well as positive features.

1. *Seemingly unrelated regression.* A system of J partly linear semiparametric "seemingly unrelated" regressions is $Y_{(j)} = \beta_{(j)}' X_{(j)} + \theta_j(Z_{(j)}) + U_{(j)}$, $1 \leqslant j \leqslant J$, where the θ_j are unknown functions and $X_{(j)}$, $Z_{(j)}$ all comprise elements of a

vector W, independent of $U_* = (U_{(1)}, \ldots, U_{(J)})$, such that a W-element might appear in X in one subset of the equations and in Z in another, disjoint, subset. Given N observations distributed as $(W, Y_{(1)}, \ldots, Y_{(J)})$, the efficiency of J separate estimators of the form (2.3) can be improved upon when $\Sigma = E(U_* U_*')$ is not diagonal, by analogy with Zellner (1963).

2. *Simultaneous equations.* Consider the structural equation

$$(7.1) \qquad Y = \alpha' Y_* + \gamma' X_* + \theta(Z) + U,$$

where Y_* is not uncorrelated with U but X_* and Z are independent of U (so nonlinearities of unknown form in endogenous variables are not allowed, though (7.1) could be completely nonparametric in exogenous variables). Replacing the conditional expectations in the projection form of (7.1) by nonparametric "estimators" gives $Y - \hat{Y} = \alpha'(Y_* - \hat{Y}_*) + \gamma'(X_* - \hat{X}_*) + U$. A valid instrument for $Y_* - \hat{Y}_*$ is a vector function of an observable vector W that includes Z and is independent of U, such that the covariance matrix in the limiting distribution of our resulting $N^{1/2}$-consistent estimator of α and γ exists. The most efficient instrument is $\tilde{Y}_* - \hat{Y}_*$, where $\tilde{Y}_*$ is a nonparametric "estimator" of $E(Y_* | W)$, which is of unknown form if the structural equations for Y, Y_* and any other endogenous variables contain nonlinearities in the endogenous and/or exogenous variables of unknown form, or even if the form of nonlinearity is known but information on Y_*'s conditional distribution given W is insufficient to parameterize $E(Y_* | W)$. (When $\theta(Z)$ is absent but Y_* still has nonparametric reduced form our estimator is similar to Newey's (1986) for nonlinear equations with known structural form but unknown reduced form.) For a full system or a subsystem of equations like (7.1), whose residuals are not all uncorrelated, a further improvement in efficiency is possible via an analogue of three stage least squares.

3. *Nonlinear regression.* Generalize (1.1) to $E(Y|X, Z) = g(X; \gamma_0) + \theta(Z)$, where g is a known function of X and the unknown s-dimensional parameter γ_0. We might estimate γ_0 by $\hat{\gamma}$ minimizing $\Sigma_i [\Sigma_j \{Y_i - Y_j - g(X_i; \gamma) + g(X_j; \gamma)\} K_{ij}]^2 I_i / \hat{f}_i^2$ over admissible γ's. The prospect of a grid search over s dimensions to obtain a starting value for iterations is daunting, and it seems desirable that representation (2.4) be used in both the search and iterations after storing DD'. In the class $g(X; \gamma_0) = \alpha h(\beta' X)$ for α an unknown scalar, we may estimate β up to an unknown scale δ, say, using derivatives of nonparametric regression as described in Section 1 or by Powell et al. (1986); then after concentrating out α we need only search over δ.

4. *Time Series.* It remains to be seen to what extent $N^{1/2}$-consistency holds when the data are serially dependent but stationary, not only for $\hat{\beta}$ but for analogues of parametric methods for improving efficiency in the presence of serially dependent residuals. One time series model of interest is the partly rational distributed lag

$$(7.2) \qquad Y_i - \sum_{j=1}^{p} \beta_j Y_{i-j} = \theta(Z_i) + U_i, \quad \left| 1 - \sum_{j=1}^{p} \beta_j s^j \right| \neq 0, \quad |s| \geqslant 1,$$

where Z_i is independent of U_j for all i and j. When Z_i consists of lagged values

of a single variable Z_{1i}, and θ is linear, (7.2) approximates a quite general linear distributed lag in Z_{1i} in a uniform frequency-domain sense, but no such strong result justifies approximating (7.2) by a linear form. When U_i is serially independent the asymptotic covariance matrix of $\hat{\beta}$ can be derived from (3.5), where β is automatically identified. A sufficient condition for $\hat{\beta}$ to be as efficient as OLS when θ is actually linear is that Z_i is stationary Gaussian (see Section 5). When U_i is serially dependent, $\hat{\beta}$ is inconsistent, but a natural extension of Liviatan's (1963) instrumental variables estimator is possible. Other time series models that might be treated are partly linear stationary autoregressions, such as $Y_i = \beta Y_{i-1} + \theta(Y_{i-2}) + U_i$.

5. *Heteroskedasticity.* Assumption (iii) of the Theorem, that U is independent of the explanatory variables, is familiar, but too strong for many econometric applications, and in fact it can be relaxed to a milder assumption on conditional moments, at the cost of some strengthening of other conditions. Under conditional heteroskedasticity ($V(U|X, Z) = \sigma^2(X, Z)$, say) $\hat{\beta}$ will still be $N^{1/2}$-consistent under appropriate conditions. A parametric form for $\sigma^2(X, Z)$ seems implausible since the conditional mean is semiparametric, but following Eicker (1963), a consistent estimator of $\Xi = E[\{X - \xi(Z)\}\{X - \xi(Z)\}'\sigma^2(X, Z)]$ in $\hat{\beta}$'s limiting covariance matrix $\Phi^{-1}\Xi\Phi^{-1}$ should be $\hat{\Xi} = N^{-1}\Sigma_i(X_i - \hat{X}_i)(X_i - \hat{X}_i)'\tilde{U}_i^2 I_i$, in the presence of heteroskedasticity of unknown form. A heteroskedastic (1.1) arises naturally from the semiparametric sample selectivity model

$$(7.3) \qquad Y_{(1)} = \beta' X + \beta_{(1)}' Z_{(1)} + U_{(1)}, \quad Y_{(2)} = \theta_2\big(Z_{(1)}, Z_{(2)}\big) + U_{(2)},$$

where we observe $Y_{(1)}$ when and only when $Y_{(2)} \geqslant 0$, so the second (decision) equation in (7.3) imparts sample selectivity when $U_{(1)}$ and $U_{(2)}$ are not independent, and where $U_{(1)}$ and $U_{(2)}$ are in any case independent of the disjoint vectors of explanatory variables X, $Z_{(1)}$ and $Z_{(2)}$. In the Tobit and some other models, all explanatory variables in the first (outcome) equation, are present also in the decision equation, in which case $\beta' X$ is absent and our approach is inapplicable. On the other hand, we do not assume a parametric conditional distribution of $U_{(1)}$ given $U_{(2)}$, and allow the decision equation to be nonparametric, in which sense (7.3) is more general than J. Heckman's (1976) model. (Some further generalization of (7.3) is possible.) With $Y = Y_{(1)} | Y_{(2)} \geqslant 0$, $Z = (Z_{(1)}, Z_{(2)})$,

$$\theta(Z) = \beta_{(1)}' Z_{(1)} + E\big(U_{(1)} | U_{(2)} \geqslant -\theta_2(Z)\big),$$

we obtain (1.1), and also

$$V(Y|X, Z) = V\big(U_{(1)} | U_{(2)} \geqslant -\theta_2(Z)\big) = \sigma^2(Z),$$

so we can use $\hat{\beta}$ as before, and (5.4) as a test for absence of sample selectivity, but we must allow for heteroskedasticity of unknown form in estimating $\hat{\beta}$'s covariance matrix if the test rejects. J. Heckman's (1976) estimator is also based on Y's regression function, but a parametric version. For other work on semiparametric inference in limited dependent variable models, see e.g. Manski (1975), Cosslett (1983, 1984), Powell (1984), Chamberlain (1986). Irrespective of (1.1)'s origin, we may improve upon $\hat{\beta}$'s efficiency in the presence of residual heteroskedasticity of

unknown form, by GLS-type estimators employing nonparametric estimators of $\sigma^2(X, Y)$, c.f. Carroll (1982), Robinson (1985).

6. *Multiplicative and other models.* An alternative, multiplicative rather than additive, semiparametric regression function appears in the model $Y = g(X; \gamma_0)\theta(Z) + U$, say a semiparametric Cobb-Douglas model with additive residuals. Then

$$Y/E(Y|Z) = g(X; \gamma_0)/E(g(X; \gamma_0)|Z) + U.$$

Nonparametric "estimates" of the two denominators can be inserted, then γ_0 estimated by NLLS. One can conceive of more general structures which permit an unknown function of Z to be identified in terms of conditional expectations of various functionals of Y and X.

Department of Economics, London School of Economics, Houghton St., London WC2A 2AE, England

Manuscript received May, 1986; final revision received October, 1987.

APPENDIX A: PROOF OF THEOREM

Necessity of (3.5) is obvious. Rewrite $\hat{\beta}$ and $\hat{\sigma}^2$ using (3.1),

$$\hat{\beta} - \beta = S_{X-\hat{X}}^{-1}\left(S_{X-\hat{X},\theta-\hat{\theta}} + S_{X-\hat{X},U-\hat{U}}\right),$$

$$\hat{\sigma}^2 - \sigma^2 = \left(S_{U-\hat{U}} - \sigma^2\right) + S_{\theta-\hat{\theta}} + \left(\hat{\beta}-\beta\right)' S_{X-\hat{X}}\left(\hat{\beta}-\beta\right) + 2S_{\theta-\hat{\theta},U-\hat{U}}$$
$$-2\left(\hat{\beta}-\beta\right)' S_{X-\hat{X},U-\hat{U}} - 2\left(\hat{\beta}-\beta\right)' S_{X-\hat{X},\theta-\hat{\theta}},$$

where

$$S_{X-\hat{X}} = S_V - S_{V\hat{V}} - S_{\hat{V}V} + S_{\hat{V}\hat{V}} + S_{V,\xi-\hat{\xi}} - S_{\hat{V},\xi-\hat{\xi}} + S_{\xi-\hat{\xi},V} - S_{\xi-\hat{\xi},\hat{V}} + S_{\xi-\hat{\xi}},$$

$$S_{X-\hat{X},\theta-\hat{\theta}} = S_{V,\theta-\hat{\theta}} - S_{\hat{V},\theta-\hat{\theta}} + S_{\xi-\hat{\xi},\theta-\hat{\theta}}, \quad S_{\theta-\hat{\theta},U-\hat{U}} = S_{\theta-\hat{\theta},U} - S_{\theta-\hat{\theta},\hat{U}},$$

$$S_{X-\hat{X},U-\hat{U}} = S_{VU} - S_{\hat{V}U} - S_{V\hat{U}} + S_{\hat{V}\hat{U}} + S_{\xi-\hat{\xi},U} - S_{\xi-\hat{\xi},\hat{U}},$$

$$S_{U-\hat{U}} = S_U - S_{U\hat{U}} - S_{\hat{U}U} + S_{\hat{U}}.$$

The proof is completed by applying Propositions 1–15 established below, which imply via the Cauchy inequality that $S_{V\hat{V}}$, $S_{V,\xi-\hat{\xi}}$, $S_{\hat{V},\xi-\hat{\xi}}$, $S_{U\hat{U}}$, $S_{\theta-\hat{\theta},U}$, and $S_{\theta-\hat{\theta},\hat{U}}$ all $\xrightarrow{p} 0$. The propositions apply the lemmas of Appendix B. We use the abbreviations $E_i(\cdot) = E(\cdot\,|Z_i)$, $\eta = \min(\lambda+1, \mu)$, $\zeta = \min(\lambda+1, \nu)$; C denotes a generic constant.

PROPOSITION 1: $E(S_{\theta-\hat{\theta}}) = O(N^{-1}a^{-q}b^{-2} + a^{2\zeta}b^{-2})$.

PROOF: By identity of distribution, $E(S_{\theta-\hat{\theta}}) = E\{(\theta_1 - \hat{\theta}_1)^2 I_1\} \leqslant N^{-2}a^{-2q}b^{-2}E(T^2)$, where $T = \Sigma t_i$, $t_i = (\theta_1 - \theta_i)K_{1i}$, where $E(T^2) \leqslant 2E\{\Sigma(t_i - t)\}^2 + 2N^2E(t^2)$, where $t = E_1(t_i)$. Conditional on Z_1, the $t_i - t$ are independent with mean 0, so $E\{\Sigma(t_i - t)\}^2 = \Sigma E(t_i - t)^2 < NE(t_2^2) = O(Na^q)$ by k's boundedness and Lemma 3. By Lemma 5, $E(t^2) = O(a^{2(q+\zeta)})$.

PROPOSITION 2: $E|S_{\xi-\hat{\xi}}| = O(N^{-1}a^{-q}b^{-2} + a^{2\eta}b^{-2})$.

PROOF: Use Proposition 1's proof and Cauchy inequality.

PROPOSITION 3: $N^{1/2}S_{\xi-\hat{\xi},\theta-\hat{\theta}} = O_p(N^{-1/2}a^{-q}b^{-2} + N^{1/2}a^{\eta+\zeta}b^{-2})$.

PROOF: By Cauchy inequality and Propositions 1 and 2.

PROPOSITION 4: $S_V = \Phi + O_p(N^{-1/2}a^{-q/2}b^{-1} + a^{\lambda}b^{-1}) + o_p(1)$.

PROOF: Because the V_i are independent and $E|X|^4 < \infty$ implies $E|V|^4 < \infty$, $N^{-1}\Sigma V_i V_i' = \Phi + O_p(N^{-1/2})$ by Chebyshev inequality. By Schwarz inequality

$$E|N^{-1}\Sigma V_i V_i'(1 - I_i)| \leqslant \left\{ E|X|^4 P\left(\hat{f}_1 < b\right)\right\}^{1/2}.$$

With $f_1 = f(Z_1)$,

$$P\left(\hat{f}_1 < b\right) \leqslant P\left(|\hat{f}_1 - f_1| > b\right) + P(f_1 < 2b).$$

By Chebyshev inequality

$$P\left(|\hat{f}_1 - f_1| > b\right) \leqslant 2\left\{ E\left(\hat{f}_1 - \tilde{f}_1\right)^2 + E\left(\tilde{f}_1 - f_1\right)^2\right\}/b^2$$

$$\text{where } \tilde{f}_1 = E_1\left(\hat{f}_1\right) = (Na^q)^{-1}\{K(0) + (N-1)E_1(K_{12})\}.$$

Thus

$$E\left(\tilde{f}_1 - f_1\right)^2 \leqslant 2E\{a^{-q}E_1(K_{12}) - f_1\}^2 + 2(Na^q)^{-2}E\{f_1 + K(0)\}^2$$

$$= O\left(a^{2\lambda} + (Na^q)^{-2}\right),$$

by Lemma 4. Because $\hat{f}_1 - \tilde{f}_1 = (Na^q)^{-1}\Sigma\{K_{1i} - E_1(K_{1i})\}$, whose summands are, conditional on Z_1, independent with zero mean,

$$E\left(\hat{f}_1 - \tilde{f}_1\right)^2 \leqslant (Na^q)^{-2}\Sigma E\left(K_{1i}^2\right) = O(N^{-1}a^{-q}),$$

then Lemma 6 implies $P(\hat{f}_1 < b) \to 0$.

PROPOSITION 5: $S_{\hat{V}} = O_p(N^{-1}a^{-q}b^{-2})$.

PROOF: Because $E(V_1|\mathscr{Z}_N) = 0$, a.s, where $\mathscr{Z}_N = (Z_1, \ldots, Z_N)$, $E|S_{\hat{V}}| \leqslant E(|\hat{V}_1|^2 I_1) \leqslant (Na^q b)^{-2}\Sigma E(|V_i|^2 K_{1i}^2)$, where the sum is

$$K(0)^2 E|V|^2 + (N-1)E\left(|V_2|^2 K_{12}^2\right) \leqslant CE|X|^2 + NE\left\{|V_2|^2 E_2\left(K_{12}^2\right)\right\}$$

$$\leqslant C(1 + Na^q)E|X|^2,$$

by Lemma 2.

PROPOSITION 6: $N^{1/2}S_{V,\theta-\hat{\theta}} = O_p(N^{-1/2}a^{-q/2}b^{-1} + a^{\zeta}b^{-1})$.

PROOF:

$$E|N^{1/2}S_{V,\theta-\hat{\theta}}|^2 = N^{-1}\sum E\left\{|V_i|^2\left(\theta_i - \hat{\theta}_i\right)^2 I_i\right\} \leqslant \left[E|V|^4 E\left\{\left(\theta_1 - \hat{\theta}_1\right)^4 I_1\right\}\right]^{1/2}$$

$$\leqslant (Na^q b)^{-2}\left\{E|X|^4 E(T^4)\right\}^{1/2}.$$

Now $E(T^4) \leqslant C[E\{\Sigma(t_i - t)\}^4 + N^4 E(t^4)]$ by Minkowski inequality, and

$$E\left\{\sum(t_i - t)\right\}^4 \leqslant \sum E\left(t_i^4\right) + \sum_{i \neq j}\sum E\left\{(t_i - t)^2(t_j - t)^2\right\}$$

$$\leqslant NE\left(t_2^4\right) + 8N^2\left[E\left(t_2^2 t_3^2\right) + \left\{E\left(t_2^4\right)E(t^4)\right\}^{1/2} + E(t^4)\right].$$

By Schwarz inequality $E(t_2^2 t_3^2) \leqslant E\{(\theta_1 - \theta_2)^4 K_{12}^2 K_{13}^2\} = E\{(\theta_1 - \theta_2)^4 K_{12}^2 E_1(K_{13}^2)\} = O(a^{2q})$, using Lemmas 2 and 3, and since $E(t^4) = O(a^{4(q+\zeta)})$ by Lemma 5,

$$\text{(A.1)} \qquad E(T^4) = O\left(N^2 a^{2q} + N^4 a^{4(q+\zeta)}\right).$$

PROPOSITION 7: $N^{1/2}S_{\hat{V},\theta-\hat{\theta}} = O_p(N^{-1/2}a^{-q/2}b^{-2} + a^{\zeta}b^{-2})$.

PROOF:

$$\text{(A.2)} \qquad E|N^{1/2}S_{\hat{V},\theta-\hat{\theta}}|^2 \leqslant E\left\{N^{-1}\sum|\hat{V}_i|^2\left(\theta_i - \hat{\theta}_i\right)^2 I_i\right\}$$

$$\text{(A.3)} \qquad + \left|E\left\{N^{-1}\sum\sum_{i \neq j}\hat{V}_i'\hat{V}_j\left(\theta_i - \hat{\theta}_i\right)\left(\theta_j - \hat{\theta}_j\right)I_i I_j\right\}\right|.$$

Because $E(|\hat{V}_1|^2 I_1|\mathscr{Z}_N) \leqslant (Na^q b)^{-2} E(\Sigma |V_i|^2 K_{1i}^2|\mathscr{Z}_N)$, a.s., (A.2)'s right-hand side is bounded by $(Na^q b)^{-4}$ times

$$E\Big(\sum |V_i|^2 K_{1i}^2 T^2\Big) \leqslant CE\big(|V_1|^2 T^2 + N|V_2|^2 t_{12}^2 K_{12}^2 + N|V_2|^2 K_{12}^2 T_1^2\big), \quad \text{where } T_1 = T - t_2.$$

By (A.1), $E(|V_1|^2 T^2) = O(Na^q + N^2 a^{2(q+\zeta)})$. Applying Lemmas 2 and 3 and (A.1),

$$E\big(|V_2|^2 t_2^2 K_{12}^2\big) \leqslant \Big[E\big\{|V_2|^4 E_2\big(K_{12}^4\big)\big\} E\big(t_2^4\big)\Big]^{1/2} = O(a^q),$$

$$\text{(A.4)} \qquad E\big(|V_2|^2 K_{12}^2 T_1^2\big) \leqslant \Big[E\big\{|V_2|^4 E_2\big(K_{12}^2\big)\big\} E\big\{T_1^4 E_1\big(K_{12}^2\big)\big\}\Big]^{1/2} = O(Na^{2q} + N^2 a^{3q+2\zeta}).$$

Thus (A.2)'s right hand side equals $O(N^{-2}a^{-2q}b^{-4} + N^{-1}a^{2\zeta - q})$. Next

$$\text{(A.5)} \qquad E\big(\hat{V}_1' \hat{V}_2 I_1 I_2|\mathscr{Z}_N\big) = (Na^q)^{-2} \hat{f}_1^{-1} \hat{f}_2^{-1} E\Big(\sum |V_i|^2 K_{1i} K_{2i}|\mathscr{Z}_N\Big)$$

so (A.3) is bounded by $N^{-3}a^{-4q}b^{-4}E(\Sigma |V_i|^2 |K_{1i}K_{2i}|T^2)$, in turn by

$$CN^{-3}a^{-4q}b^{-4}E\big\{\big(|V_1|^2 + |V_2|^2\big)\big(t_2^2 + |K_{12}|T_1^2\big) + N|V_3|^2 |K_{13}K_{23}|\big(t_2^2 + t_3^2 + T_2^2\big)\big\},$$

where $T_2 = T_1 - t_3$. As in (A.4) and (A.5), $E(|V_i|^2 t_2^2) = O(a^q)$, $E(|V_i|^2 |K_{12}|T_1^2) = O(Na^{2q} + N^2 a^{3q+2\zeta})$ for $i = 1, 2$. Applying Lemmas 2 and 3

$$E\big(|V_3|^2 |K_{13}K_{23}|t_3^2\big) \leqslant \Big[E\big\{|V_3|^4 E_3\big(K_{13}^2\big) E_3|K_{23}|\big\} E\big\{t_3^4 E_3|K_{23}|\big\}\Big]^{1/2} = O(a^q),$$

and *a fortiori*, $E(|V_3|^2 |K_{13}K_{23}|t_2^2) = O(a^q)$. Applying Lemma 2 and (A.1),

$$E\big(|V_3|^2 |K_{13}K_{23}|T_2^2\big) \leqslant \Big[E\big\{|V_3|^4 E_1|K_{13}|E_3|K_{23}|\big\} E\big\{T_2^4 E_1\big(|K_{13}|E_3|K_{23}|\big)\big\}\Big]^{1/2},$$

which is $O(Na^{3q} + N^2 a^{4q+2\zeta})$. Thus (A.3) $= O(N^{-1}a^{-q}b^{-4} + a^{2\zeta}b^{-4})$.

PROPOSITION 8: $N^{1/2} S_{U, \xi - \hat{\xi}} = O_p(N^{-1/2}a^{-q/2}b^{-1} + a^\eta b^{-1})$.

PROOF: By independence of $\{U_i\}$, $\{Z_i\}$, $E|N^{1/2} S_{U, \xi - \hat{\xi}}|^2 = \sigma^2 E\{\operatorname{tr}(S_{\xi - \hat{\xi}})\}$. Apply Proposition 2.

PROPOSITION 9: $N^{1/2} S_{\hat{U}, \xi - \hat{\xi}} = O_p(N^{-1/2}a^{-q/2}b^{-2} + a^\eta b^{-2})$.

PROOF:

$$\text{(A.6)} \qquad E|N^{1/2} S_{\hat{U}, \xi - \hat{\xi}}|^2 \leqslant E\big(\hat{U}_1^2 |\xi_1 - \hat{\xi}_1|^2 I_1\big) + 2\Big|NE\Big\{\hat{U}_1 \hat{U}_2 \big(\xi_1 - \hat{\xi}_1\big)' \big(\xi_2 - \hat{\xi}_2\big) I_1 I_2\Big\}\Big|.$$

Put $w_i = (\xi_1 - \xi_i) K_{1i}$, $W = \Sigma w_i$, $W_1 = W - w_2$, $W_2 = W_1 - w_3$. The first term on (A.6)'s right-hand side has bound $C(Na^q b)^{-4}$ times

$$\begin{aligned} E\Big(\sum K_{1i}^2 |W|^2\Big) &\leqslant C\big[E|W|^2 + NE|w_2|^2 + NE\big\{|W_1|^2 E_1\big(K_{12}^2\big)\big\}\big] \\ &= O(N^2 a^{2q} + N^3 a^{3q+2\eta}), \end{aligned}$$

using Lemmas 2 and 3 and Proposition 1's proof. The second term of (A.6)'s right-hand side is likewise bounded by $C(Na^q b)^{-4}$ times

$$\begin{aligned} &NE\Big(\sum |K_{1i}K_{2i}|\,|W|^2\Big) \\ &\quad \leqslant CN\big[E\big(|w_2|^2 + |W_1|^2 E_1|K_{12}|\big) \\ &\qquad + NE\big\{|w_2|^2 E_1|K_{13}| + |w_3|^2 E_3|K_{23}| + |W_2|^2 E_1\big(|K_{13}|E_3|K_{23}|\big)\big\}\big] \\ &\quad = O(N^3 a^{3q} + N^4 a^{4q+2\eta}). \end{aligned}$$

PROPOSITION 10: $N^{1/2}S_{U\hat{V}} = O_p(N^{-1/2}a^{-q/2}b^{-1})$.

PROOF: By independence of $\{U_i\}$ and $\{X_i, Z_i\}$, $E|N^{1/2}S_{U\hat{V}}|^2 = \sigma^2 E(|\hat{V}_1|^2 I_1) = O(N^{-1}a^{-q}b^{-2})$ as in Proposition 5's proof.

PROPOSITION 11: $N^{1/2}S_{\hat{U}V} = O_p(N^{-1/2}a^{-q/2}b^{-1})$.

PROOF: Conditioning first on $\{U_i, V_i\}$, then only on $\{V_i\}$,

$$E|N^{1/2}S_{\hat{U}V}|^2 \leqslant E\left(|V_1|^2\hat{U}_1^2 I_1\right) \leqslant C(Na^q b)^{-2}E\left(|V_1|^2\sum K_{1i}^2\right) = O(N^{-1}a^{-q}b^{-2}).$$

PROPOSITION 12: $N^{1/2}S_{\hat{U}\hat{V}} = O_p(N^{-1/2}a^{-q/2}b^{-2})$.

PROOF: $E|N^{1/2}S_{\hat{U}\hat{V}}|^2 \leqslant E(\hat{U}_1^2|\hat{V}_1|^2 I_1) + 2N|E(\hat{U}_1\hat{U}_2\hat{V}_1'\hat{V}_2 I_1 I_2)|$. The first term on the right hand side has bound $C(Na^q b)^{-4}$ times

$$\begin{aligned}
&E\left(\sum K_{1i}^2 \sum |V_j|^2 K_{1j}^2\right) \\
&\quad \leqslant C\left[E|V_1|^2 + NE\left\{|V_1|^2 E_1\left(K_{12}^2\right)\right\} + N^2 E\left\{|V_3|^2 E_1\left(K_{12}^2\right)E_1\left(K_{13}^2\right)\right\}\right] \\
&\quad = O(N^2 a^{2q})
\end{aligned}$$

by Lemma 2. After taking expectations over $\{U_i\}$ and applying (A.5),

$$\begin{aligned}
&\left|E\left(\hat{U}_1\hat{U}_2\hat{V}_1'\hat{V}_2 I_1 I_2\right)\right| \\
&\quad = \sigma^2(Na^q)^{-4}\left|E\left\{\left(\sum K_{1i}K_{2i}\right)\left(\sum |V_j|^2 K_{1j}K_{2j}\right)\hat{f}_1^{-1}\hat{f}_2^{-1} I_1 I_2\right\}\right| \\
&\quad \leqslant \sigma^2(Na^q b)^{-4}E\left(\sum |K_{1i}K_{2i}| \sum |V_j|^2 |K_{1j}K_{2j}|\right) \\
&\quad \leqslant C(Na^q b)^{-4}E\left\{\left(|K_{12}| + N|K_{13}K_{23}|\right)\right. \\
&\qquad\qquad \left.\times\left(|V_1|^2|K_{12}| + |V_3|^2|K_{13}K_{23}| + N|V_4|^2|K_{14}K_{24}|\right)\right\}
\end{aligned}$$

of which the dominant term has bound $C(Na^{2q}b^2)^{-2}E\{|V_4|E_4(|K_{14}K_{24}|E_2|K_{23}|)\} = O(N^{-2}a^{-q}b^{-4})$.

PROPOSITION 13: $S_{\hat{U}} = O_p(N^{-1}a^{-q}b^{-2})$.

PROOF: $E(S_{\hat{U}}) = \sigma^2(Na^q b)^{-2}E(\Sigma K_{1i}^2 I_1) = O(N^{-1}a^{-q}b^{-2})$.

PROPOSITION 14: $S_U = \sigma^2 + o_p(1)$.

PROOF: By Khinchine law of large numbers $N^{-1}\Sigma U_i^2 \xrightarrow{p} \sigma^2$, whereas $E|N^{-1}\Sigma U_i^2(1 - I_i)| = \sigma^2 P(\hat{f}_1 < b) \to 0$ by Proposition 4's proof.

PROPOSITION 15: $N^{1/2}S_{UV} \xrightarrow{d} N(0, \sigma^2\Phi)$.

PROOF: By Levy central limit theorem $N^{-1/2}\Sigma U_i V_i \xrightarrow{d} N(0, \sigma^2\Phi)$, whereas

$$E\left|N^{-1/2}\sum U_i V_i(1 - I_i)\right| \leqslant \sigma^2\left\{E|X|^4 P\left(\hat{f}_1 < b\right)\right\}^{1/2} \to 0$$

as before.

APPENDIX B: TECHNICAL LEMMAS

Lemmas 1–3 below are unoriginal, merely versions of results used time after time in the immense kernel estimation literature, but they are presented for ease of reference, while their short proofs will aid the reader unfamiliar with kernel manipulations. Although Lemmas 4 and 5's proofs use techniques familiar in the kernel literature, previous results on effects of higher-order kernels of which we are aware concern bias of estimation at a fixed, rather than random, point, and we were unable to find the results we need. It is inconceivable that Lemma 6 is new, but we failed to locate a reference.

LEMMA 1: *Let* $\sup_u|k(u)| + \int|u^\lambda k(u)|\,du < \infty$, *for some* $\lambda \geq 0$. *Then uniformly in* z

$$\int|y-z|^\lambda|K((y-z)/a)|\,dy = O(a^{q+\lambda}). \tag{B.1}$$

PROOF: The left-hand side is

$$a^{q+\lambda}\int|y|^\lambda|K(y)|\,dy \leq a^{q+\lambda}q^\lambda\int|u^\lambda k(u)|\,du\left(\int|k(u)|\,du\right)^{q-1}.$$

LEMMA 2: *Let* $\sup_z f(z) < \infty$, $\sup_u|k(u)| + \int|k(u)|\,du < \infty$. *Then uniformly in* z

$$E|K((Z-z)/a)| = O(a^q).$$

PROOF: The left-hand side $\leq \sup_z f(z)\int|K((y-z)/a)|\,dy$; then apply Lemma 1.

LEMMA 3: *Let* $\sup_z f(z) < \infty$, $E|g(Z)| < \infty$, $\sup_u|k(u)| + \int|k(u)|\,du < \infty$. *Then*

$$E|g(Z_1)K_{12}| = O(a^q).$$

PROOF: The left-hand side $\leq E|g(Z_1)E_1|K_{12}|\,| \leq Ca^qE|g(Z)|$, by Lemma 2.

LEMMA 4: *For* λ *satisfying* $l-1 < \lambda \leq l$, *where* $l \geq 1$ *is an integer, let* $f \in \mathscr{G}_\lambda^\infty$, $k \in \mathscr{K}_l$. *Then*

$$E\{a^{-q}E_1(K_{12}) - f(Z_1)\}^2 = O(a^{2\lambda}). \tag{B.2}$$

PROOF: Let $R(y,z) - f(z)$ be defined relative to f in the way $Q(y,z)$ was in relation to g in Definition 2, so it is a homogeneous $(l-1)$-degree polynomial in $y-z$ with coefficients that are bounded functions of z, the remainder term in the Taylor expansion being in $\mathrm{Lip}(\lambda - m + 1)$. By (3.2),

$$\int\{R(y,z) - f(z)\}K((y-z)/a)\,dy \equiv 0,$$

so $E\{K((Z-z)/a)\} - a^qf(z)$ is bounded by

$$\left|\int_{\mathscr{S}_{z\rho}}\{f(y) - R(y,z)\}K\left(\frac{y-z}{a}\right)dy\right| + \left|\int_{\bar{\mathscr{S}}_{z\rho}}\{R(y,z) - f(z)\}K\left(\frac{y-z}{a}\right)dy\right|$$

$$+\left|\int_{\bar{\mathscr{S}}_{z\rho}}\{f(y) - f(z)\}K\left(\frac{y-z}{a}\right)dy\right|,$$

which is bounded by $CL(\lambda)$, where $L(\lambda)$ is the left-hand side of (B.1), because $|y-z| \geq \rho$ for

$y \in \bar{\mathscr{S}}_{z\rho}$ and $\lambda > l - 1$. Now $k(u) = O((1 + |u|^{l+1+\varepsilon})^{-1})$ implies $\int |u^l k(u)|\, du < \infty$. Thus by Lemma 1, not only $E\{a^{-q}K((Z-z)/a) - f(z)\} = O(a^{\lambda})$ for all z, but (B.2) follows by dominated convergence.

LEMMA 5: *For λ, μ satisfying $l - 1 < \lambda \leqslant l$, $m - 1 < \mu \leqslant m$, where $l \geqslant 1$, $m \geqslant 1$ are integers, and for $\alpha \geqslant 1$, let $f \in \mathscr{G}_{\lambda}^{\infty}$, $g \in \mathscr{G}_{\mu}^{\alpha}$, $k \in \mathscr{K}_{l+m-1}$. Then*

$$E\left|E_1\left[\{g(Z_1) - g(Z_2)\}K_{12}\right]\right|^{\alpha} = O(a^{\alpha(q+\eta)}).$$

PROOF: By (3.2), $\int Q(y,z)R(y,z)K((y-z)/a)\, dy \equiv 0$, so $|E[\{g(Z) - g(z)\}K((Z-z)/a)]|$ is bounded by

$$\begin{aligned}
&\left|\int_{\mathscr{S}_{z\rho}} \{g(y) - g(z) - Q(y,z)\}f(y)K\left(\frac{y-z}{a}\right) dy\right| \\
&\quad + \left|\int_{\mathscr{S}_{z\rho}} Q(y,z)\{f(y) - R(y,z)\}K\left(\frac{y-z}{a}\right) dy\right| \\
&\quad + \left|\int_{\bar{\mathscr{S}}_{z\rho}} Q(y,z)R(y,z)K\left(\frac{y-z}{a}\right) dy\right| \\
&\quad + \left|\int_{\bar{\mathscr{S}}_{z\rho}} \{g(y) - g(z)\}f(y)K\left(\frac{y-z}{a}\right) dy\right| \\
&\leqslant Ch(z)L(\mu) + G(z)\sum_{i=1}^{m-1} L(i+\lambda) + H(z)L(\lambda+\mu) \\
&\quad + C\{|g(z)| + E|g(Z)|\}\, a^{q+\eta} \sup_u \{|u|^{q+\eta}|k(u)|^q\},
\end{aligned}$$

where $E\{G(Z)^{\alpha} + H(Z)^{\alpha}\} < \infty$. Then again apply Lemma 1 and dominated convergence, noting that $\min(\mu, \lambda + 1, \lambda + \mu) = \eta \leqslant \min(l + 1, m) \leqslant l + m - 1 < q(l + m - 1 + \varepsilon)$.

LEMMA 6: $\lim_{b \to 0} P(f(Z) < b) = 0$.

PROOF:

$$P(f(Z) < b) \leqslant b\int_{|z| \leqslant B} dz + P(|Z| > B) \leqslant (2B)^q b + P(|Z| > B),$$

for all $B > 0$. For any $\varepsilon > 0$, choose B so $P(|Z| > B) < \varepsilon$; then $b < (2B)^{-q}\varepsilon$.

REFERENCES

AMEMIYA, T. (1980): "Selection of Regressors," *International Economic Review*, 21, 331–354.

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

BEGUN, J., W. J. HALL, W. HUANG, AND J. A. WELLNER (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432–452.

BERAN, R. (1977): "Adaptive Estimates for Autoregressive Processes," *Annals of the Institute of Statistical Mathematics*, 28, 77–89.

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

CACOULLOS, T. (1966): "Estimation of a Multivariate Density," *Annals of the Institute of Statistical Mathematics*, 18, 179–189.

CARROLL, R. J. (1982): "Adapting for Heteroscedasticity in Linear Models," *Annals of Statistics*, 10, 1224–1233.

CHAMBERLAIN, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

COLLOMB, G. C. (1985): "Nonparametric Regression: An Up-to-Date Bibliography," *Statistics*, 2, 309–324.

COSSLETT, S. J. (1983): "Distribution-free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51, 765–782.

——— (1984): "Distribution-Free Estimator of a Regression Model with Sample Selectivity," manuscript, University of Florida.

COX, D. D. (1985): "A Penalty Method for Nonparametric Estimation of the Logarithmic Derivative of a Density Function," *Annals of the Institute of Statistical Mathematics*, 37, 271–288.

EICKER, F. (1963): "Asymptotic Normality and Consistency of the Least Squares Estimator for Families for Linear Regressions," *Annals of Mathematical Statistics*, 34, 447–456.

ELBADAWI, I., A. R. GALLANT, AND G. SOUZA (1983): "An Elasticity Can Be Estimated Consistently Without *A Priori* Knowledge of its Functional Form," *Econometrica*, 51, 1731–1751.

ENGLE, R. F., C. W. J. GRANGER, J. RICE, AND A. WEISS (1986): "Semiparametric Estimates of the Relation Between Weather and Electricity Demand," *Journal of the American Statistical Association*, 81, 310–320.

FRIEDMAN, J., AND W. STUETZLE (1981): "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

GALLANT, A. R. (1985): "Identification and Consistency in Seminonparametric Regression," paper presented at the World Congress of the Econometric Society.

HAUSMAN, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

HECKMAN, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

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

KAGAN, A. M., Y. V. LINNIK, AND C. R. RAO (1973): *Characterization Problems in Mathematical Statistics*. New York: Wiley.

LIVIATAN, N. (1963): "Consistent Estimation of Distributed Lags," *International Economic Review*, 4, 44–52.

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

——— (1984): "Adaptive Estimation of Non-Linear Regression Models," (with comment), *Econometric Reviews*, 3, 145–194.

NEWEY, W. K. (1986): "Efficient Estimation of Models with Conditional Moment Restrictions," manuscript, Princeton University.

POWELL, J. L. (1984): "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303–325.

POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1986): "Semiparametric Estimation of Weighted Average Derivatives," manuscript, Massachusetts Institute of Technology.

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

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

ROBINSON, P. M. (1983): "Nonparametric Estimators for Time Series," *Journal of Time Series Analysis*, 4, 185–207.

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

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

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

SCHUSTER, E., AND S. YAKOWITZ (1979): "Contributions to the Theory of Non-parametric Regression, with Application to System Identification," *Annals of Statistics*, 7, 139–149.

SCHILLER, R. J. (1984): "Smoothness Priors and Nonlinear Regression," *Journal of the American Statistical Association*, 72, 420–423.

STOCK, J. H. (1985): "Nonparametric Policy Analysis; An Application to Estimating Hazardous Waste Cleanup Benefits," manuscript.

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

STONE, C. J. (1981): "Admissible Selection of an Accurate and Parsimonious Normal Linear Regression Model," *Annals of Statistics*, 9, 475–485.

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

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

WAHBA, G., (1984): "Partial Spline Models for the Semi-Parametric Estimation of Functions of Several Variables," in *Statistical Analysis of Time Series*. Tokyo: Institute of Statistical Mathematics, 319–329.

——— (1985): "Discussion to 'Projection Pursuit', by P. J. Huber," *Annals of Statistics*, 13, 518–521.

WHITE, H. (1980): "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149–170.

——— (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

ZELLNER, A. (1962): "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57, 348–368.

——— (1970): "Estimation of Regression Relationships Containing Unobservable Variables," *International Economic Review*, 11, 441–454.