An Efficient Semiparametric Estimator for Binary Response Models
Author(s): Roger W. Klein and Richard H. Spady
Source: *Econometrica*, Vol. 61, No. 2 (Mar., 1993), pp. 387-421
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/2951556
Accessed: 14-03-2018 14:33 UTC

Econometrica, Vol. 61, No. 2 (March, 1993), 387–421

AN EFFICIENT SEMIPARAMETRIC ESTIMATOR FOR BINARY RESPONSE MODELS

BY ROGER W. KLEIN AND RICHARD H. SPADY[1]

This paper proposes an estimator for discrete choice models that makes no assumption concerning the functional form of the choice probability function, where this function can be characterized by an index. The estimator is shown to be consistent, asymptotically normally distributed, and to achieve the semiparametric efficiency bound. Monte-Carlo evidence indicates that there may be only modest efficiency losses relative to maximum likelihood estimation when the distribution of the disturbances is known, and that the small-sample behavior of the estimator in other cases is good.

KEYWORDS: Binary choice, semiparametric estimation, adaptive kernel density estimation.

1. INTRODUCTION

IN THIS PAPER WE FORMULATE a semiparametric estimator for the discrete choice model that satisfies the classical desiderata for such estimators: consistency, root-N normality, and efficiency. The estimator is semiparametric in that it makes no parametric assumption on the form of the distribution generating the disturbances. We do, however, assume that the choice probability function depends on a parametrically specified index function. A method for computing the efficiency bound in semiparametric contexts was given by Begun, Hall, Huang, and Wellner (1983). General discussions of these bounds and their calculation can be found in Newey (1989, 1990). The actual efficiency bound for the binary choice model was given by Chamberlain (1986) and Cosslett (1987).

To simplify our exposition we formally consider throughout most of this paper only the binary choice model given by

$$(1) \qquad y = \begin{cases} 1 & \text{if } v(x;\theta_0) \geq u_0, \\ 0 & \text{otherwise,} \end{cases}$$

where $v(\cdot;\cdot)$ is a known function, x is a vector of exogenous variables, θ_0 an unknown parameter vector, and u_0 a random disturbance. We assume that observations $\{x_i, y_i\}$ are i.i.d. and that the model satisfies an index restriction: $E(y|x) = E[y|v(x;\theta_0)]$, where E is the indicated conditional expectation and $v(x;\theta_0)$ is an aggregator or index.

The index restriction permits multiplicative heteroscedasticity of a general but known form and heteroscedasticity of an unknown form if it depends only

[1] Earlier versions of this paper were distributed in a session of the Econometric Society Summer Meetings, June, 1987, and presented at the European Econometric Society Meetings, September, 1988. We thank Bob Sherman for helpful discussions and Whitney Newey for numerous comments and technical assistance. In particular, Whitney Newey provided the basic proof showing the equivalence between the asymptotic variance-covariance matrix derived for the estimator here and its lower bound. We would also like to thank L. P. Hansen and anonymous referees for their comments and suggestions. Any errors are the sole responsibility of the authors.

on the index. We characterize such heteroscedasticity by $u_0 \equiv s(x;\beta_0)\circ\varepsilon$, where ε is independent of x. When the (positive) scaling function s is a known and general function of x, the index restriction is satisfied in the transformed model with index $v^* \equiv v(x;\theta_0)/s(x;\beta_0)$. If s is unknown, but depends on x only through the index $v(x;\theta_0)$, then the index restriction will also be satisfied. In general, the index restriction will hold if the model can be written in the form shown in (1) with $u_0 \equiv u[v(x;\theta_0),\varepsilon]$, where ε is independent of x. Notice that we distinguish $u_0 \equiv u[v(x;\theta_0),\varepsilon]$ from $u \equiv u[v(x;\theta),\varepsilon]$. We need not make this distinction when u is independent of x, in which case $u = u_0 = \varepsilon$.

Throughout, for notational simplicity we will define probability functions through their arguments. In this manner, for example, $\Pr[u_0 \leqslant a(x)|b(x)]$ denotes the probability of the event $u_0 \leqslant a(x)$ conditioned on $b(x)$. Formally, with $F_{u_0|x}$ as the distribution function for u_0 conditioned on x and E denoting a conditional expectation taken with respect to x conditioned on $b(x)$:

$$\Pr\left[u_0 \leqslant a(x)|b(x)\right] \equiv E\left[F_{u_0|x}\left[a(x)\right]|b(x)\right]. \tag{2}$$

To simplify notation, we will also for the most part not distinguish random variables from their realizations. Typically, the appropriate interpretation will be clear from the context.

The most common way to estimate θ_0 in (1) is to apply the method of maximum likelihood. When the disturbances are independently distributed according to a known conditional parametric distribution, $F_{u|x}$, the log likelihood is

$$L = \sum_{i=1}^{N}\left[y_i \ln\left(P_i^*\right) + (1-y_i)\ln\left(1-P_i^*\right)\right], \tag{3}$$

$$P^*\left[v(x;\theta);\theta\right] \equiv \Pr\left[u < v(x;\theta)|v(x;\theta)\right] = F_{u|x}\left[v(x;\theta)\right],$$

$$P_i^*(\theta) \equiv P^*\left[v(x_i;\theta);\theta\right].$$

Typically, it is further assumed that u and x are independent, in which case $F_{u|x}$ may be replaced by the unconditional distribution F_u. When F_u is unknown, one approach is to choose F_u itself jointly with θ to maximize the likelihood. This approach is taken in Cosslett (1983), who shows that the resulting estimator of θ_0 is consistent.

In the above approach, once the distribution function is replaced with its maximum likelihood estimate, the resulting concentrated likelihood is not a smooth function of θ. Consequently, it is difficult to establish the asymptotic distribution for the estimator of θ. To circumvent this problem, we propose to select θ so as to maximize a semiparametric likelihood that is a smooth function of θ and that locally (for θ in a neighborhood of θ_0) approximates the corresponding parametric likelihood.

To formulate this semiparametric likelihood, which might more appropriately be termed a quasi-likelihood, we begin by replacing the probability function $P_i^*(\theta)$ in (3) with a tractable function $P_i(\theta)$ that locally approximates it. To

construct $P_i(\theta)$, note that with C as the event $[u < v(x;\theta)]$, P^* in (3) is equivalent to $\Pr[C|v(x;\theta)]$, the probability of the event C conditioned on $v(x;\theta)$. We may characterize this probability function by

$$P^*[v(x;\theta);\theta] \equiv \Pr[C|v(x;\theta)] = \Pr[C]g_{v|C}(v;\theta)/g_v(v;\theta), \tag{4}$$

where $\Pr(C)$ is the unconditional probability of C, $g_{v|C}$ is the density for $v = v(x;\theta)$ conditioned on C, and g_v is the unconditional density for v. Let C_0 be the event C at θ_0: $[u_0 < v(x;\theta_0)]$, which is observable and is equivalent to the event $y = 1$. If we replace C in (4) with C_0, we obtain the probability function $P(v;\theta)$. In other words, $P(v;\theta)$ is the probability of the event C_0 conditioned on $v = v(x;\theta)$:

$$\begin{aligned} P[v(x;\theta);\theta] &\equiv \Pr[C_0|v(x;\theta)] = \Pr[C_0]g_{v|C_0}(v;\theta)/g_v(v;\theta) \\ &= \Pr[y=1]g_{v|y=1}(v;\theta)/g_v(v;\theta), \quad P_i(\theta) \equiv P(v_i;\theta), \end{aligned} \tag{5}$$

where $g_{v|y=1}$ is the density for v conditioned on $y = 1$.

This function has several desirable features. First, at $\theta = \theta_0$ and under the index restriction, it is equivalent to the true choice probability. Namely, with $v_0 \equiv v(x;\theta_0)$,

$$P(v_0;\theta_0) = \Pr[C_0|v(x;\theta_0)] \equiv \Pr[y=1|v(x;\theta_0)] = \Pr[y=1|x]. \tag{6}$$

Subject to a continuity condition, the function P in (6) may therefore be viewed as a local approximation to the corresponding parametric probability function, P^* in (3). Second, although the probability function in (5) is unknown, it can be estimated directly as a smooth function of θ. We can estimate $\Pr[y = 1]$ by the sample proportion of individuals making the choice in question, while both the conditional and unconditional densities can be estimated nonparametrically as smooth functions of the parameter vector θ. Notice that if these densities are estimated by kernel methods (with the same window being employed for conditional and unconditional densities), then the estimate of (5) will be the usual kernel estimate of the expected value of y conditioned on the index. Finally, since v aggregates the information in the possibly high dimensional vector x into a scalar, we need only to be concerned with univariate density estimation.

Using $\hat{P}_i(\theta)$ to denote $P_i(\theta)$ as defined in (5) with its three components replaced by estimates, we propose to estimate θ_0 by choosing θ to maximize an adjusted version of the estimated quasi-likelihood:

$$Q(\hat{P}(\theta)) = \sum_{i=1}^{N} (\hat{\tau}_i/2)\left(\left[y_i \ln\left[\hat{P}_i(\theta)^2\right] + (1-y_i)\ln\left[\left(1-\hat{P}_i(\theta)\right)^2\right]\right]\right), \tag{7}$$

where $\hat{\tau}_i$ is a trimming sequence that is introduced for technical reasons to control the rate at which the estimated densities comprising $\hat{P}_i$ tend to zero. Notice also that for both y_i and $(1 - y_i)$ terms, the argument of the ln function is squared. As will become apparent below, there are two methods for estimating the densities upon which estimated probability functions depend: locally-

smoothed kernels and bias reducing kernels. For the former, estimated probability functions lie between zero and one, in which case (7) reduces to the usual form for a binomial likelihood. For the latter, estimated densities and hence estimated probabilities can be negative. This problem, which can be ignored asymptotically, may occur in any given finite sample. The estimated quasi-likelihood function in (7) remains well-defined in this case. We refer to this function as "estimated" to distinguish (7) from the quasi-likelihood proper which is (7) with $\hat{P}_i(\theta)$ replaced by $P_i(\theta)$.

Our strategy is essentially to show that the estimator $\hat{\theta}(\hat{P})$, which is obtained by maximizing (7), behaves asymptotically like the estimator $\hat{\theta}(P)$ defined for a known probability function P as

$$(8) \qquad \hat{\theta}[P] \equiv \arg\sup \overline{Q}(P) \equiv \arg\sup \sum_{i=1}^{N} [y_i \ln(P_i) + (1-y_i)\ln(1-P_i)]/N.$$

The estimator $\hat{\theta}(P)$ can be analyzed by conventional methods to establish consistency and asymptotic normality. Perhaps not surprisingly, as a likelihood-based estimator, it is also efficient (under the appropriate regularity conditions).

The recent econometric literature includes a variety of different approaches to semiparametric estimation in this context, including the papers of Cosslett (1983), Han (1984, 1988), Horowitz (1992), Ichimura (1986), Manski (1975, 1985, 1988), Powell, Stock, and Stoker (1989), Ruud (1983, 1986), Sherman (1993), and Stoker (1986). Our estimator is similar to Ichimura's, in that both can be interpreted as minimum distance estimators. It also resembles Cosslett's in that it replaces P_i^* in the MLE objective function (3) with a nonparametrically estimated function.

However, the spirit of our approach is perhaps best illustrated by considering it in light of the nonparametric discrimination problem first posed in the classic paper of Fix and Hodges (1951).[2] The univariate version of that problem, in our notation, can be posed as: given a scalar random variable v with samples $v[y=0]$ and $v[y=1]$ of realizations of v from two populations (say, respectively, of the $y=0$ and $y=1$ populations), how should a new realization v of unknown origin be classified? Fix and Hodges' solution was to estimate $g_{v|y}[v]$ nonparametrically and to assign v to the $y=0$ group if $g_{v|y=0}[v]/g_{v|y=1}[v] > c$, where c is predetermined. It is a short step from this setup to the realization that the densities of v among the subpopulations plus a knowledge of the overall subpopulation frequencies allows, via Bayes' rule, the calculation of the probability of group membership conditional upon v. Optimal classification is, with such probabilities known or estimated, a matter of one's loss function. If we select θ so as to minimize the KLIC (Kullback-Leibler information criterion)

[2] Reprinted recently with a commentary by Silverman and Jones (1989). This paper not only clearly formulated the nonparametric discrimination problem, but also originated two methods of nonparametric density estimation (kernel density estimation and nearest neighbors), as well as recognizing optimal smoothing as a bias-variance tradeoff.

discrepancy of $P_i(\theta)$ from $P_i^*(\theta)$ over the sample y_i realizations, then the resulting estimator will be equivalent to that which maximizes the quasi-likelihood function in (8).

In what follows, we begin in Section 2 by first providing an overview of the main assumptions and then discussing conditions required for identification. In Section 3 we prove consistency, and in Section 4, we establish asymptotic normality (at rate $N^{1/2}$) and efficiency. In each of these sections, for expositional purposes we will outline the proofs. Complete proofs are relegated to an Appendix. To illustrate the performance of the estimator in practice, in Section 5 we undertake several Monte-Carlo experiments in which the semiparametric estimator is compared to probit under several model specifications. We find that the semiparametric estimator performs quite well relative to probit, and can, in models sufficiently perturbed from the usual probit specification, substantially dominate the probit estimator. In Section 6 we conclude by indicating several extensions of the proposed estimator and directions for future research.

2. ASSUMPTIONS

2.1. *An Overview*

We will require conditions that serve to define the model, insure that various estimated functions are sufficiently well-behaved, and establish those functions of the parameters that are identified. In stating these assumptions, we need to introduce some notation. Let $F_{u_0|x}$ be the distribution function for u_0 in (1) conditioned on x. Then, with E denoting a conditional expectation taken with respect to x conditioned on $v(x;\theta)$, define the probability function $P[\ ]$ as:

$$P[v(x;\theta);\theta] \equiv E_{x|v(x;\theta)}\{F_{u_0|x}[v(x;\theta_0)]\}. \tag{9}$$

In the above definition, $P[\ ]$ inherits its dependence on θ from the manner in which the distribution of x conditioned on $v(x;\theta)$ depends on θ. In general, this distribution will depend on $v(x;\theta)$ and separately on θ.[3] Finally, let $D_z^r f(z)$ denote the rth order partial of $f(z)$ with respect to z and $D_z^0 f(z) \equiv f(z)$. Then, with c denoting a positive generic constant, we assume the following conditions hold:

(C.1) *The data consist of a random sample* (y_i, x_i), $i = 1, \ldots N$. *The random variable y is binomial with realizations* 1 *and* 0.

(C.2) *The parameter vector* θ *lies in a compact parameter space,* Θ. *The true value of* θ, θ_0, *is in the interior of* Θ.

[3] For example, suppose that x is normal with a nonzero mean vector and that v is linear in the components of x. In this case, the distribution of x conditioned on $v(x;\theta) = x\theta$ will be normal with a mean that depends on $x\theta$ and on θ.

(C.3a) *There exists a scalar aggregator,* $a(x;\theta_0)$*, such that for any* x:

$$\Pr[y=1|x] \equiv E(y|x) = E[y|a(x;\theta_0)] \equiv \Pr[y=1|a(x;\theta_0)],$$

(C.3b) *With* x_1 *a continuous element of* x*, there exists a monotonic transformation,* T*, such that under* T

$$v(x;\theta_0) \equiv T[a(x;\theta_0)] = v_1(x_1) + v_2(x_2;\theta_0), \quad |\partial v_1/\partial x_1| > c.$$

Conditions (C.1)–(C.3) describe the manner in which the data are generated. Assumption (C.3a) is especially important because it allows us to reduce the dimension of the conditioning variables. As will become apparent below, assumption (C.3b) is useful in relating properties of the distribution of the index to those of the exogenous variables. This assumption will also be convenient in obtaining conditions under which identification holds. For the linear index case, as is standard in the literature, condition (3b) will hold under a location-scale transformation provided that the continuous variable has a nonzero coefficient. For a nonlinear example, consider the aggregator

$$(10) \qquad a \equiv [x_1]^{\gamma_{10}}\left[[x_2]^{\gamma_{20}} + [x_3]^{\gamma_{30}}\right]^{\gamma_{40}},$$

$a > 0$. Then the index $v \equiv (1/\gamma_{10})\ln(a)$ has the required form for $\gamma_{10} \neq 0$.

(C.4a) *There exist* $\underline{P}, \bar{P}$ *that do not depend on* x *such that (employing* (C.3)*):*

$$0 < \underline{P} \leqslant \Pr[y=1|v(x;\theta_0)] = \Pr(y=1|x) \leqslant \bar{P} < 1.$$

This condition serves to bound the probability function $P[v(x;\theta);\theta]$ away from zero and one. In the notation introduced above,

$$(11) \qquad \Pr[y=1|v(x;\theta_0)] = F_{u_0|x}[v(x;\theta_0)] \quad \text{and}$$

$$P[v(x;\theta);\theta] = E\left(F_{u_0|x}[v(x;\theta_0)]\right).$$

This restriction, which requires that $v(x;\theta_0)$ be a bounded random variable, can be relaxed as we will be downweighting observations for which the densities comprising $P[v(x;\theta);\theta]$ are small.

(C.4b) *With* $H[v(x;\theta_0)] \equiv \Pr[u_0 < v(x;\theta_0)|v(x;\theta_0)]$*, assume that* $H(t)$ *is continuously differentiable in* t *for all* t *and that* $|\partial H(t)/\partial t| < c$.

It should be noted that this condition could be replaced by a somewhat weaker and (notationally) more complicated dominance condition. To interpret this condition, note that when x and u_0 are independent, $H[v(x;\theta_0)] = F_{u_0}[v(x;\theta_0)]$, where F_{u_0} is the distribution function for u_0. Consequently, the above condition holds when u_0 is independent of x and has a bounded and continuous density. We require that it also hold when u_0 is dependent on x under the index restriction in (C.3a)–(C.3b).

(C.5) *The index v is smooth in that for θ in a neighborhood of θ_0 and all t:*

$$\left\{|D_\theta^{[r]} v(t;\theta)|, |\partial\left(D_\theta^{[r]} v(t;\theta)\right)/\partial x|\right\} < c \qquad (r = 0,1,2,3,4).$$

(C.6) *With $v(x;\theta_0) \equiv v_1(x_1) + v_2(x_2;\theta_0)$ from* (C.3b), *let $f_{x_1|\circ}(x_1)$ be the density for the continuous variable x_1 conditioned on the remaining exogenous variables, x_2, and on y, where x_2 is a vector of either discrete and/or continuous random variables. This conditional density is supported on $[a,b]$ and is smooth in that for all x*

$$|D_t^r f_{x_1|\circ}(t)| < c \qquad (r = 1,\ldots,4;\ t \in [a,b]).$$

Conditions (C.5)–(C.6) insure that the densities that underlie $P(v;\theta)$ are sufficiently smooth. Let $g_{x_1|\circ}$ and $g_{v_1|\circ}$ be the densities for x_1 and $v_1 \equiv v_1(x_1)$ respectively, conditioned on x_2 and y. Given the smoothness assumptions on the index, if $g_{x_1|\circ}$ has r derivatives with respect to x_1, $g_{v_1|\circ}$ must also have r derivatives with respect to v_1. It now follows that since

$$g_{v|\circ}(s;\theta) = g_{v_1|\circ}\left[s - v_2(x_2;\theta)\right], \tag{12}$$

and v_2 is smooth in θ (C.5), $g_{v|\circ}$ must have r derivatives in θ. Finally, since

$$g_{v|y}(s;\theta) \equiv E\left[g_{v|\circ}(s;\theta)|y\right], \tag{13}$$

$g_{v|y}(s;\theta)$ must also have r derivatives with respect to θ. For example, from (12)–(13) with $r = 1$:

$$|\partial g_{v|y}(s;\theta)/\partial\theta| \leqslant \sup_t |\partial v_2(t;\theta)/\partial\theta| \sup_s |\partial g_{v_1|y}(s;\theta)/\partial s|. \tag{14}$$

The first term above is bounded by assumption, and the second (from the above discussion) inherits its bound from the upper bound on $|\partial g_{x_1|\circ}(w)/\partial w|$ and the lower bound on $|\partial v_1/\partial x_1|$.

To illustrate this smoothness condition in (C.5), let the density for x_1 conditioned on x_2 and y be given by

$$g_{x_1|\circ}(t) \propto \left[\alpha_1^2 - (t-\alpha_2)^2\right]^{r+1}, \tag{15}$$

which, with $\alpha_2 > \alpha_1 > 0$, has compact support on $[\alpha_2 - \alpha_1, \alpha_2 + \alpha_1]$ and r derivatives on the entire real line. As stated earlier, this smoothness assumption is technically convenient in insuring that the index has a smooth density. As there are important cases in which this assumption does not hold, in the concluding section we will indicate how the argument can be modified. In what follows, the only problematic features of the densities for the index will be that they are permitted to approach zero at an arbitrary rate and at unknown points (that may be either on the support boundaries or in the interior of the support).

(C.7) *With* $h_N \to 0$, *the trimming function employed to downweight observations has the form*

$$\tau(t;\varepsilon) \equiv \left\{1 + \exp\left[\left(h_N^{\varepsilon/5} - t\right)/h_N^{\varepsilon/4}\right]\right\}^{-1},$$

where $\varepsilon > 0$ *and* t *is to be interpreted as a density estimate.*

With t above replaced by an estimated density, this function will be employed to downweight observations for which the corresponding densities are small. As $h_N \to 0$, this trimming function behaves as a smoothed indicator in that it tends exponentially to one for t sufficiently far from zero and to zero for t sufficiently small.

(C.8) *Letting* $g_{yv}(v_i;\theta) \equiv \Pr(y) g_{v|y}(v_i;\theta)$, *define the estimator for* g_{yv} *as*

$$\hat{g}_{yv}\left(v_i;\theta,\hat{\lambda}_y;h_N\right) \equiv \sum_{j\neq i}^{N} \frac{1\{y_j = y\}}{h_N\hat{\lambda}_{yj}} K\left[\frac{v_i - v_j}{h_N\hat{\lambda}_{yj}}\right] \Big/ (N-1) \qquad (y = 0,1).$$

The kernel function, $K(z)$, *is a symmetric function that integrates to one, has bounded second moment, and*

$$\left\{|D_z^r K(z)|, \int |D_z^r K(z)|\, dz\right\} < c \qquad (r = 0,1,2,3,4).$$

The parameter h_N *is a nonstochastic window,* $N^{-1/6} < h_N < N^{-1/8}$. *The kernel is also either*:

(C.8a) *Bias Reducing*: *With* $\hat{\lambda}_{yj} \equiv 1$, $h_N\hat{\lambda}_{yj} = h_N$, $j = 1,\ldots,N$, *and*

$$\int z^2 K(z)\, dz = 0;$$

or

(C.8b) *Adaptive with local smoothing*: *Windows have the form*

$$h_N\hat{\lambda}_{yj} \equiv [h_N]\left[\hat{\sigma}_y(\theta)\right]\left[\hat{L}_{yj}\right].$$

The component, $\hat{\sigma}_y$, *is the sample standard deviation of* $v(x;\theta)$ *conditioned on* y. *For the final component, which reflects local smoothing, let* $\hat{l}_{yj} \equiv \hat{g}_{yv}(v_j;\theta,1;h_{NP})$ *be a preliminary density estimate of the form shown in* (C.8) *with window (without local smoothing)* $h_{NP}[\hat{\sigma}_y(\theta)]$. *For* $0 < \varepsilon' < 1$, *from* (C.7) *define*:

$$\hat{\tau}_{Pj} \equiv \tau\left(\hat{l}_{yj}, h_{NP}^{\varepsilon'}\right).$$

The local smoothing parameter can now be defined by:

$$\hat{L}_{yj} \equiv \left\{\left[\hat{l}_{yj} + (1 - \hat{\tau}_{Pj})\right]/m\right\}^{-1/2},$$

where m is the geometric mean of the $\hat{l}_{yj}$. *For notational simplicity, we will write* $\hat{g}_{yv}(t;\theta) \equiv \hat{g}_{yv}(t;\theta,\hat{\lambda}_y;h_N)$ *to refer to either of the estimates in* (C.8*a*) *or* (C.8*b*).

Assumptions (C.8a) and (C.8b) specify two alternative density estimators whose similar bias properties are required in the normality proof. The estimator in (C.8a) is a bias reducing kernel that is permitted to be negative in the tails. With density derivatives uniformly bounded up to order 4, such a kernel has a uniform bias of order h_N^4. Assumption (C.8b) specifies an adaptive kernel estimator with a variable and data dependent window size. This estimator differs from that in (C.8a) first by adjusting the global window size by the scale of the distribution. We have followed Silverman (1986) in this regard. Such scaling could also be introduced into the bias reducing kernel estimator in (C.8a). Note that the density estimator in (C.8b) is also restricted to be positive.

Finally, these estimators differ in that the window in (C.8a) is constant at each sample size. In contrast, the estimator in (C.8b) specifies variable windows. To select such windows, Abramson (1982) showed that the $[g_{yv}(v_j)/m]^{-1/2}$, $j = 1,\ldots,N$, are optimal invariant local smoothing parameters in a pointwise mean-squared error sense when densities are bounded away from zero. Under known local smoothing and for densities bounded away from zero with uniformly bounded derivatives up to order 4, Silverman obtains a uniform bias of order h_N^4 for the density estimated with local smoothing. By downweighting observations for which densities are small and appropriately modifying the local smoothing parameters (to control the rate at which they can tend to zero), estimated local smoothing parameters can be taken as given. As a result, we will be able to show that under local smoothing the density estimate has a bias that is sufficiently small for our purposes. It should be remarked at this point that both Abramson (1982) and Silverman (1986) report that this two-stage procedure (the first stage being required to obtain estimated local smoothing parameters) performs well in Monte-Carlo studies. Moreover, the second-stage density estimate is reportedly not too sensitive to the first-stage estimate.

(C.9) *The model is identified in that for any* $x \in A$, *a set of positive probability,*

$$P[y = 1|v(x;\theta_0)] = P[y = 1|v(x;\theta_*)] \Rightarrow \theta_0 = \theta_*.$$

The final assumption, which is made for purposes of identification, can be interpreted as either a statement about what parameter restrictions are necessary for identification or what functions of the parameters are identified. To avoid separate notation for "true" parameter values and identifiable functions of them, we will proceed under the first interpretation. In the next subsection, we provide sufficient conditions for identification and discuss several examples.

2.2. *Identification*

Any value of θ that maximizes the limiting quasi-likelihood must satisfy the condition on probabilities in (C.9), which clearly holds if $\theta_* = \theta_0$. The identifi-

cation assumption requires that this solution be unique. There are two classes of models and corresponding restrictions for which such identification holds. When the probability function, $P[y=1|v(x;\theta_0)]$, is monotonic in the index, we have the following result.

THEOREM 1 (Monotonic Models): *Assume that* $P[y=1|v_0]$ *is strictly monotonic in* $v_0 \equiv v(x;\theta_0)$ *and that conditions* (C.3a)–(C.3b) *and* (C.5)–(C.6) *hold. Let* $\mathscr{A}$ *be a set of positive probability for x on which* $\partial P[y=1|v_0]/\partial v_0 \neq 0$. *For* $x \in \mathscr{A}$, *assume that if* $v_2(x_2;\theta_0) = v_2(x_2;\theta_*)$, *then* $\theta_0 = \theta_*$. *It now follows that identification holds as specified in* (C.9).

Cosslett (1983) obtained this result for the linear index model, with the theorem above requiring an appropriate extension to the argument. From monotonicity of the probability function, there exists a function H for which

$$P\{y=1|[v(x;\theta_0)]\} = P\{y=1|[v(x;\theta_*)]\}$$
$$\Rightarrow v_1(x_1) + v(x_2;\theta_0) = H[v_1(x_1) + v_2(x_2;\theta_*)].$$

Differentiating both sides of this equality with respect to x_1 (for $x \in \mathscr{A}$), it follows that H must be the identity, in which case $v_2(x_2;\theta_0) = v_2(x_2;\theta_*)$. Identification then holds if this condition implies that $\theta_0 = \theta_*$ as claimed.

To provide an example, it is convenient to write $\theta \equiv \theta(\gamma)$ and $\theta_0 \equiv \theta(\gamma_0)$, where θ_0 is an identifable parameter vector that depends on some possibly higher dimensional vector, γ_0. Then, employing this notation, suppose that the (untransformed) index in (C.3a) is given by

$$a(x;\theta_0) = (x_1)^{\gamma_{10}}(x_2)^{\gamma_{20}}(x_3)^{\gamma_{30}}, \qquad a>0. \tag{16}$$

From (C.3b), with T as the log transform and x_1 continuous, $v(x;\theta) = (1/\gamma_1)\ln[a(x;\theta_0)]$ for $\gamma_1 \neq 0$. Under Theorem 1, the parameter vector $\theta_0 \equiv \theta(\gamma_0) \equiv (\gamma_{20}/\gamma_{10}, \gamma_{30}/\gamma_{10})$ is identified (the matrix of log observations is assumed to have full rank). This result is as expected since in log form we have the usual linear index model with identification obtained via a scale restriction (there is no intercept).

A somewhat more interesting (nonlinear) example is given by

$$a(x;\theta_0) = (x_1)^{\gamma_{10}}\left((x_2)^{\gamma_{20}} + (x_3)^{\gamma_{30}}\right)^{\gamma_{40}}, \quad a>0. \tag{17}$$

As in the first example, with $\gamma_{10} \neq 0$, $v(x;\theta_0) \equiv (1/\gamma_{10})\ln(a)$. From Theorem 1, we can now identify $\theta_0 \equiv (\gamma_{40}/\gamma_{10}, \gamma_{20}, \gamma_{30})$ provided that

$$\theta_{10}\ln\left((x_2)^{\theta_{20}} + (x_3)^{\theta_{30}}\right) = \theta_{1*}\ln\left((x_2)^{\theta_{2*}} + (x_3)^{\theta_{3*}}\right), \tag{18}$$

only holds when $\theta_0 = \theta_*$. Even if x_2 and x_3 are discrete in this example, in the monotonic index case it is still possible to identify the (appropriately scaled) parameters of the model. For instance, (18) will be uniquely solved at the true parameter values when (x_2, x_3) take on the discrete values: $(1,1)$, $(0,2)$, and $(2,0)$ with positive probability.

As an alternative model for which other identifying conditions are required, let the data be generated by:

$$(19)\qquad y=\begin{cases}1 & \text{if } v(x;\theta_0)>s[v(x;\theta_0)]\varepsilon\\ 0 & \text{otherwise,}\end{cases}$$

where s is an unknown function of the index and ε is independent of x. If the function s was a known function of x, then with s bounded away from zero, we could redefine the index as v/s and apply Theorem 1. When s is an unknown function of the index, the probability function need not be monotonic in $v(x;\theta_0)$ and Theorem 1 is no longer applicable. We could generalize this model further by replacing the condition under which y is one above with $f[v(x;\theta_0)+u_0]>0$, where the function f need not be monotonic. For the linear index model, this case is discussed by Manski (1988). Extending this result for nonlinear indices, we have the following theorem.

THEOREM 2 (Nonmonotonic Models): *Assume that $x\equiv(x_1,x_2)$ is a vector whose components are all continuous on a set $\mathscr{A}$ of positive probability. With the index given from* (C.3b) *by $v(x;\theta_0)=v_1(x_1)+v_2(x_2;\theta_0)$, assume that for $x\in\mathscr{A}$, the functions v_1 and v_2 are differentiable in x_1 and x_2 and $\partial v_1/\partial x_1\neq 0$. Let θ_* be a value of θ such that on $\mathscr{A}$,*

$$\partial v_2(x_2;\theta_*)/\partial x_2=\partial v_2(x_2;\theta_0)/\partial x_2.$$

Then, identification holds if $\theta_*=\theta_0$ uniquely solves this equation.

To prove Theorem 2, let $v\equiv v_1(x_1)+v_2(x_2;\theta_0)$ and $v_*\equiv v_1(x_1)+v(x_2;\theta_*)$. Then

$$(20)\qquad G(v)\equiv P[y=1|v_1(x_1)+v_2(x_2;\theta_0)]$$
$$=P[y=1|v_1(x_1)+v_2(x_2;\theta_*)]\equiv H(v_*).$$

Differentiating both sides with respect to x_1, it follows that for $x\in\mathscr{A}$: $\partial G(v)/\partial v=\partial H(v_*)/\partial v_*$. Employing this result, the claim now follows by differentiating both sides of the equality $G(v)=H(v_*)$ with respect to x_2 (where $x\in\mathscr{A}$).

3. THE ESTIMATOR AND CONSISTENCY

3.1. *The Estimator*

Parameter estimates are obtained by maximizing an estimated quasi-likelihood function, which will be defined and discussed in this section. We begin by defining the estimated probability functions, as they are the fundamental components of this estimation method. Recall that the "true" probability function is given as

$$(21)\qquad P_i(\theta)\equiv g_{1v}(v_i;\theta)/g_v(v_i;\theta),\quad g_v\equiv g_{1v}+g_{0v}.$$

It would seem natural to define an estimated probability function by replacing all of these components with the corresponding kernel estimates given in (C8). However, in so doing we would encounter technical difficulties with a uniform convergence argument when estimated densities become too small.

To guard against small estimated densities, let

$$(22) \qquad \hat{\delta}_{yN} \equiv h_N^a\left[e^z/(1+e^z)\right], \quad z \equiv \left[\left(h_N^b - \hat{g}_{yv}(v;\theta)\right)/h_N^c\right]$$

be adjustment factors, where (a, b, c) are parameters that must satisfy certain restrictions to be explained below. Then, with $\hat{g}_v \equiv \hat{g}_{1v} + \hat{g}_{0v}$ and $\hat{\delta}_N \equiv \hat{\delta}_{0N} + \hat{\delta}_{1N}$, we define the estimated probability function as

$$(23) \qquad \hat{P}(v;\theta) \equiv \left[\hat{g}_{1v}(v;\theta,h_N) + \hat{\delta}_{1N}(v;\theta)\right]/\left[\hat{g}_v(v;\theta) + \hat{\delta}_N(v;\theta)\right].$$

The purpose of the adjustment factors in (23) is to control the rate at which numerator and denominator of estimated probability functions tend to zero. For this purpose, we first require that $0 < b < c$ in (21). Under this restriction, for small estimated densities (of smaller order than h_N^b) the adjustment factors behave exponentially like h_N^a. For sufficiently large estimated densities, the adjustment factors tend exponentially to zero as no adjustment is required in this case. In this manner, numerator and denominator of the estimated probability function behave asymptotically like h_N^a for small estimated densities and like the unadjusted form $\hat{g}_{1v}/\hat{g}_v$ otherwise. As a second restriction on the adjustment parameters, we need to select a, which essentially controls the adjustment rate for the estimated probability functions. In this regard, it must be noted that we will require two derivatives of the probability function with respect to θ. As might be expected, it is technically convenient if we can ignore derivatives of the adjustment factors. In this regard, we require that a be sufficiently large, and the restriction: $a > 2c + 2b > 0$ will suffice. For expositional purposes, we parameterize the adjustment factors in a manner that insures that all restrictions on trimming parameters are satisfied. Namely, with $\varepsilon' > 0$, in (21) we select $a \equiv \varepsilon'$, $b = \varepsilon'/5$, and $c = \varepsilon'/4$. Throughout we will refer to these probability adjustments as "probability trimming."

With $\hat{P}_i \equiv \hat{P}(v_i;\theta)$, it can be shown that $|\hat{P}_i(\theta) - P_i(\theta)|$ converges in probability, uniformly in i, θ to zero except for v_i in a region with vanishing probability. Then, one can show that without any further trimming, the quasi-likelihood (i.e. (7) with $\hat{\tau}_i = 1$ for all i) converges to a function that is uniquely maximized at θ_0. The estimator that maximizes this objective function would then be consistent. For the normality argument, we would need to show that the gradient to this objective function was asymptotically distributed as normal at $\theta = \theta_0$. Unfortunately, such an argument requires (at $\theta = \theta_0$) not only that $\hat{P}_i$ converge to P_i, but also that such convergence be at a sufficiently fast rate. Since we cannot establish such a rate in the presence of observations for which densities are "too" small, we will introduce a trimming sequence that (exponentially) downweights such observations.

Let $\hat{\theta}_P$ be a preliminary consistent estimate for θ_0 for which $|\hat{\theta}_P - \theta_0|$ is $O_P(N^{-1/3})$.[4] Then, employing (C.7), define the following "likelihood trimming" function:

$$\hat{\tau}_i \equiv \hat{\tau}_{i0}\tau_{i1}, \quad \hat{\tau}_{iy} \equiv \tau\left(\hat{g}_{yv}\left[v\left(x_i;\hat{\theta}_P\right);\hat{\theta}_P\right],\varepsilon''\right) \quad \text{for} \quad y = 0,1, \tag{24}$$

where the first argument of τ is the kernel estimate in (C.8) for $\theta = \hat{\theta}_P$. For the second argument or trimming rate, $0 < \varepsilon'' < a \equiv \varepsilon'$, where a is given as in (22). The estimated quasi-likelihood is now defined as

$$\hat{Q}_N(\theta;\hat{\tau}) \equiv \sum (\hat{\tau}_i/2)\left[y_i \ln\left[\hat{P}_i(\theta)^2\right] + (1-y_i)\ln\left[\left(1-\hat{P}_i(\theta)\right)^2\right]\right]/N. \tag{25}$$

There are several features of this expression that should be noted. First, as discussed earlier, this function remains well-defined even in the presence of (asymptotically negligible) estimated probability functions that are negative. Second, as to trimming, it should be noted that probability trimming is less severe than likelihood trimming ($\varepsilon'' < \varepsilon'$). As a consequence, we will be able to show that the gradient to $\hat{Q}_N$ in (25) at θ_0 behaves (asymptotically) as if there were no probability trimming. This result will prove useful in the normality argument below. Third, this expression incorporates both likelihood and probability trimming. For reasons discussed earlier, probability trimming (as in (23)) is not in itself sufficient for our purposes. It might appear that likelihood trimming evaluated at θ (instead of at the preliminary estimate, $\hat{\theta}_P$) would suffice as the sole form of trimming. However, in this case the gradient would depend on derivatives of the likelihood trimming function with respect to θ. Since the trimming function approximates an indicator, there must be regions in which such derivatives become arbitrarily large. Such regions would present problems as they are not sufficiently small to ignore in analyzing the gradient (multiplied by $N^{1/2}$).

3.2. *Consistency*

To prove consistency for the estimator maximizing the estimated quasi-likelihood function in (25), we need to establish the uniform (in θ) limit of this function. In this section, we will sketch the argument and provide details of the proof in the Appendix which contains all required lemmas. Based on convergence rates for kernel estimates and their derivatives (Lemma 2), Lemma 4 of the Appendix establishes a uniform (in i,θ) convergence rate for $\hat{P}_i(\theta)$ to $P_i(\theta)$. As a result, we are able to show that for the estimated quasi-likelihood, $\hat{Q}_N(\theta;\hat{\tau})$ in (25), estimated probability functions can be taken as known:

$$\sup_{\theta} |\hat{Q}_N(\theta;\hat{\tau}) - Q_N(\theta;\hat{\tau})| = o_p(1), \tag{26}$$

$$Q_N(\theta,\hat{\tau}) \equiv \sum \hat{\tau}_i\left[y_i \ln\left[P_i(\theta)\right] + (1-y_i)\ln\left[1-P_i(\theta)\right]\right]/N.$$

[4] At the cost of complicating the argument, it is possible to permit a slower convergence rate. For the given convergence rate, the estimators given by Han (1988) (see also Sherman (1992)), Horowitz (1992), and Manksi (1975) would suffice.

From Lemma 3, we may take the trimming function as given in that with $v_0 \equiv v(x;\theta_0)$, $\tau_i \equiv \tau_{i0}\tau_{i1}$, and $\tau_{iy} \equiv \tau[g_{yv_0}(v(x_i;\theta_0);\theta_0);\varepsilon'']$:

$$\sup_{\theta} |Q_N(\theta;\hat{\tau}) - Q_N(\theta;\tau)| = o_p(1), \tag{27}$$

$$Q_N(\theta,\tau) \equiv \sum \tau_i \left[y_i \ln\left[P_i(\theta)\right] + (1-y_i)\ln\left[1-P_i(\theta)\right]\right]/N.$$

Finally, we show that the trimming function in (27), τ_i, may be treated as if it were one (its limiting value) in that with

$$Q_N(\theta) \equiv Q_N(\theta,1)\colon \sup_{\theta} |Q_N(\theta,\tau) - Q_N(\theta)| = o_p(1). \tag{28}$$

This limiting function, $Q_N(\theta)$, converges in probability, uniformly in θ, to its expectation with maximizing value, θ_*, characterized by

$$P[y=1|x] = P[y=1|v(x;\theta_0)] = P[y=1|v(x;\theta_*)]. \tag{29}$$

The first equality follows from the aggregator condition, (C.3a), and the second equality is a necessary condition for θ_* to maximize the limiting quasi-likelihood. From the identification assumption, (C.9), $\theta_* = \theta_0$ is the unique maximizing value. Consistency now follows as summarized in Theorem 3 below.

THEOREM 3 (Consistency): *Under conditions* (C.1)–(C.9):

$$\hat{\theta} \equiv \arg\sup_{\theta}\left(\hat{Q}(\theta;\tau)\right) \xrightarrow{p} \theta_0.$$

Having established consistency, in the next section, we show that $\hat{\theta}$ is distributed asymptotically as (root N) normal with mean zero and minimum variance-covariance matrix (under the appropriate regularity conditions).

4. THE ASYMPTOTIC DISTRIBUTION OF THE ESTIMATOR

4.1. *Normality*

We begin with a Taylor series expansion for the gradient of the quasi-likelihood function,

$$N^{1/2}\left(\hat{\theta}-\theta_0\right) = -\left[\partial^2\hat{Q}[\theta^+;\hat{\tau}]/\partial\theta\,\partial\theta'\right]^{-1} N^{1/2}\partial\hat{Q}[\theta_0;\hat{\tau})]/\partial\theta, \tag{30}$$

where $\theta^+ \in [\hat{\theta},\theta_0]$. Lemma 5 establishes convergence in probability for the Hessian term in (30):

$$-\left[\partial^2\hat{Q}[\theta^+;\hat{\tau}]/\partial\theta\,\partial\theta'\right]^{-1} \xrightarrow{p} E\left\{\left[\frac{\partial P}{\partial\theta}\right]\left[\frac{\partial P}{\partial\theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}^{-1}_{\theta=\theta_0}. \tag{31}$$

Accordingly, it remains to show that the gradient term in (30) is asymptotically distributed as normal with mean zero. As in earlier sections, we will sketch the proof and leave the details to the Appendix.

To analyze the gradient, write it as a weighted sum of residuals:

$$\hat{G}(\theta_0) \equiv \partial \hat{Q}[\theta_0; \hat{\tau}]/\partial\theta \equiv \sum \hat{\tau}_i \hat{r}_i \hat{w}_i / N; \tag{32}$$

$$\hat{r}_i \equiv \left[y_i - \hat{P}_i \right] / \hat{\alpha}_i, \quad \hat{\alpha}_i \equiv \left[\hat{g}_v(v_i; \theta_0) + \hat{\delta}_N(v_i; \theta_0) \right] \left[\hat{P}_i (1 - \hat{P}_i) \right];$$

$$\hat{w}_i \equiv \left[\hat{g}_v(v_i; \theta_0) + \hat{\delta}_N(v_i; \theta_0) \right] \left[\partial \hat{P}(\theta_0)/\partial\theta \right].$$

The decomposition is helpful in that the residuals and weights each have desirable properties that we will exploit below. The residual in (32) estimates

$$r_{iN} \equiv r_N(x_i; \theta_0) \equiv \left[y_i - P_i(\theta_0) \right] / \alpha_i, \tag{33}$$

$$\alpha_i \equiv \left[\left[g_v(v_i; \theta_0) + \delta_N(v_i; \theta_0) \right] P_i(\theta_0) \left[1 - P_i(\theta_0) \right] \right].$$

With E denoting the indicated conditional expectation,

$$E\left[r_N(x; \theta) | v(x; \theta_0) \right] = E\left[r_N(x; \theta) | x \right] = 0, \tag{34}$$

a useful property in establishing asymptotic normality.

With $P_i \equiv \Pr[y = 1 | v(x; \theta) \equiv v(x_i; \theta)]$, the weight function in (32) estimates $w(x_i; \theta_0) \equiv g_v(v_i; \theta_0)[\partial P_i / \partial\theta]_{\theta = \theta_0}$, which has the useful property that

$$E\left[w(x; \theta_0) | v(x; \theta_0) \right] = 0. \tag{35}$$

This property will hold if the conditional expectation of the derivative of the semiparametric probability function, P, is zero. To examine this derivative, we need to be careful to account for the manner in which P depends on θ. Let $F(-|x) \equiv F_{u_0|x}$ be the distribution function for u_0 conditioned on x and $G(-|s,\theta)$ the distribution function for x conditioned on $v(x; \theta) = s$. Then, we define

$$P(s; \theta) \equiv \int F\left[v(x; \theta_0) | x \right] dG(x | s, \theta) = \int H\left[v(x; \theta_0) \right] dG(x | s, \theta), \tag{36}$$

$$F\left[v(x; \theta_0) | x \right] \equiv \Pr\left[u_0 < v(x; \theta_0) | x \right] = \Pr\left[u_0 < v(x; \theta_0) | v(x; \theta_0) \right]$$

$$\equiv H\left[v(x; \theta_0) \right].$$

In (36) the function H, which we introduce for notational convenience, depends only on $v(x; \theta_0)$ under the index restriction.

In what follows, we want to differentiate P when $s \equiv v(t; \theta)$ for fixed t. To this end, let $\alpha(x; \theta) \equiv v(x; \theta_0) - v(x; \theta)$ and with $v(t; \theta)$ substituted for s in

(36), rewrite P as

$$(37)\qquad P[v(t,\theta);\theta] = \int H[\alpha(x;\theta) + v(x;\theta)]\, dG[x|v(t;\theta),\theta].$$

Then, from the chain rule with $P[v(t;\theta);\theta]$ given as in (37),[5]

$$(38)\qquad \partial P/\partial\theta|_{\theta=\theta_0} = D_1(t,\theta_0)\colon \frac{\partial}{\partial\theta}\left[\int H[v(x;\theta) + \alpha(x;\theta_0)]\, dG[x|v(t;\theta),\theta]\right]_{\theta=\theta_0} + D_2(t,\theta_0)\colon \frac{\partial}{\partial\theta}\left[\int H[v(x;\theta_0) + \alpha(x;\theta)]\, dG[x|v(t;\theta_0),\theta_0]\right]_{\theta=\theta_0}.$$

To explain the two terms in (38), for simplicity consider $f[r(\theta), s(\theta)]$ as any function of θ. The derivative of this function at $\theta = \theta_0$ is

$$(39)\qquad \partial f/\partial\theta|_{\theta=\theta_0} = [[\partial f/\partial r][\partial r/\partial\theta] + [\partial f/\partial s][\partial s/\partial\theta]]_{\theta=\theta_0} = [\partial f[r(\theta), s(\theta_0)]/\partial\theta + \partial f[r(\theta_0), s(\theta)]/\partial\theta]_{\theta=\theta_0}.$$

Notice that in the first term in which r varies and s is held fixed, we may evaluate s at θ_0 before we differentiate. An analogous result holds for the term in which s varies and r is held fixed. Employing this argument in (38), in D_1 we consider variations in θ with α held fixed. As this term will be evaluated at $\theta = \theta_0$, in α we are entitled to replace θ with θ_0 before we differentiate. In an analogous manner, we obtain D_2 by letting α vary and holding all other functions of θ as fixed.

With $D_i(t,\theta_0)$ given in (38), $i = 1,2$, let $D(x,\theta_0) \equiv D_1(x,\theta_0) + D_2(x,\theta_0)$. Then, (35) will follow if $E[D(x;\theta_0)|v(x;\theta_0)] = 0$. To establish this result, we will proceed by simplifying each of the terms in (38). Under (C4.b)–(C.5), H in (38) is differentiable in a neighborhood of θ_0. Then, since $\alpha(x;\theta_0) = 0$,

$$(40)\qquad \begin{aligned} D_1(t;\theta_0) &= \frac{\partial}{\partial\theta}\left[\int H[v(x;\theta)]\, dG[x|v(t;\theta),\theta]\right]_{\theta=\theta_0} \\ &= \frac{\partial}{\partial\theta}\left[\int H[v(t;\theta)]\, dG[x|v(t;\theta),\theta]\right]_{\theta=\theta_0} \\ &= \frac{\partial}{\partial\theta}\left[H[v(t;\theta)]\int dG[x|v(t;\theta),\theta]\right]_{\theta=\theta_0} \\ &= \frac{\partial}{\partial\theta}H[v(t;\theta)]|_{\theta=\theta_0}. \end{aligned}$$

For D_2, since $\alpha(x;\theta_0) = 0$ and $\partial\alpha/\partial\theta = -\partial v(x;\theta)/\partial\theta$, from (C.4b)–(C.5) we may differentiate within the integral to obtain

$$(41)\qquad \begin{aligned} D_2(t;\theta_0) &= \left[\int \frac{\partial}{\partial\theta}\left(H[v(x;\theta_0) + \alpha(x;\theta)]\right) dG(x|v(t;\theta_0),\theta_0)\right]_{\theta=\theta_0} \\ &= -E[D_1(x;\theta_0)|v(x;\theta_0) = v(t;\theta_0)], \end{aligned}$$

[5] We are grateful to Whitney Newey for suggesting this "chain-rule" approach for the case in which x and u are independent. A similar argument holds for models satisfying an index restriction.

where the above expectation is taken with respect to the distribution of x conditioned on $v(x;\theta_0) = v(t;\theta_0)$. From (38)–(41),

$$E\big[D(x;\theta_0)|v(x;\theta_0)\big] = E\big[D_1(x;\theta_0)|v(x;\theta_0)\big] - E\big[D_1(x;\theta_0)|v(x;\theta_0)\big] = 0. \tag{42}$$

Proceeding with the normality argument, the strategy is to show that all estimated components of the gradient can be replaced by the corresponding true components. If this substitution is permissible, then asymptotic normality will readily follow as the new gradient will conform to a standard central limit theorem.

Let $G_N(\theta_0)$ be the "true" gradient obtained by replacing $\hat{\tau}_i$, $\hat{r}_i$, and $\hat{w}_i$ in $\hat{G}(\theta_0)$ with τ_i, r_{iN}, and w_i. In Lemma 6 of the Appendix, we prove $N^{1/2}[\hat{G}(\theta_0) - G_N(\theta_0)] = o_p(1)$. To briefly sketch out the argument here, from (32)–(35) decompose this difference as follows:

$$N^{1/2}\big[\hat{G}(\theta_0) - G(\theta_0)\big] \equiv N^{-1/2}\sum \hat{\tau}_i\hat{r}_i\hat{w}_i - N^{-1/2}\sum \tau_i r_{iN} w_i \tag{43}$$

$$= \quad A\colon N^{-1/2}\sum \tau_i\big(\hat{r}_i\hat{w}_i - r_{iN}w_i\big) \tag{43a}$$

$$+ B\colon N^{-1/2}\sum \big(\hat{\tau}_i - \tau_i\big) r_{iN} w_i \tag{43b}$$

$$+ C\colon N^{-1/2}\sum \big(\hat{\tau}_i - \tau_i\big)\big(\hat{r}_i\hat{w}_i - r_{iN}w_i\big). \tag{43c}$$

In the Appendix, we show that each of the above components converges to zero in probability. As the arguments for terms B and C are similar to and simpler than those for A, here we will sketch out the proof for A and leave other details to the Appendix. To analyze A in (43a), write it as

$$N^{-1/2}\sum \tau_i\big(\hat{r}_i - r_{iN}\big)w_i + N^{-1/2}\sum \tau_i\big(\hat{r}_i - r_{iN}\big)\big(\hat{w}_i - w_i\big) + N^{-1/2}\sum \tau_i r_{iN}\big(\hat{w}_i - w_i\big). \tag{44}$$

A_1 A_2 A_3

For A_1, note from (32) that the estimated residual is a ratio of estimated functions: $\hat{r}_i \equiv (y_i - \hat{P}_i)/\hat{\alpha}_i$. As estimated denominators will pose a technical problem to the (mean-square convergence) argument that we seek to employ, expand $\hat{r}_i$ as a function of $\hat{\eta}_i \equiv 1/\hat{\alpha}_i$ about $\eta_i \equiv 1/\alpha_i$:

$$\tau_i\hat{r}_i = \tau_i\hat{r}_i^* + o_p(N^{-1/2}), \quad \hat{r}_i^* \equiv \big[\big(y_i - \hat{P}_i\big)\eta_i\big]\big[1 - \big(\hat{\alpha}_i - \alpha_i\big)\eta_i\big], \tag{45}$$

where the remainder term above is uniformly (in i) of order $N^{-1/2}$. Then,

$$A_1^* \equiv N^{-1/2}\sum \tau_i\big(\hat{r}_i^* - r_{iN}\big)w_i = A_1 + o_p(1). \tag{46}$$

Proceeding with A_1^*,

$$E\big(A_1^*\big)^2 = E\Big(\sum \tau_i^2\big(\hat{r}_i^* - r_{iN}\big)^2 w_i^2/N\Big). \tag{47}$$

Notice that in obtaining the above expectation, we have ignored all cross product terms. By taking an iterated expectation, conditioning first on x_i, $i = 1,\ldots,N$, and then on v_i, $i = 1,\ldots,N$, such terms vanish from the property of

the weight function in (35) (see Lemma 6 of Appendix A). Each remaining term within the expectations operator in (47) converges to zero in probability and can be shown to satisfy an integrability condition. Therefore, the expectation in (47) does converge to zero as claimed.

Turning to A_2, it converges in probability to zero because from Lemma 4 one can show that

$$(48) \qquad \sup_i |\tau_i^{1/2}(\hat{r}_i - r_{iN})| \sup_i |\tau_i^{1/2}(\hat{w}_i - w_i)| = o_p(N^{-1/2}).$$

For the final "A-term", A_3, the argument is somewhat similar to that for A_1. Namely, by making use of the fact that the "true" residual has zero conditional expectation (see (34)), this term converges to zero in mean-square and is therefore $o_p(1)$.

From the above analysis of the gradient and (30)–(33):

$$(49) \qquad N^{1/2}\left(\hat{\theta} - \theta_0\right) = E\left\{\left[\frac{\partial P}{\partial \theta}\right]\left[\frac{\partial P}{\partial \theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}_{\theta=\theta_0}^{-1}\left[N^{-1/2}\sum \tau_i r_{iN} w_i\right] + o_P(1).$$

Since this expression consists of a sum of i.i.d. terms, each with zero mean and bounded pth absolute moment, $p > 2$, from Serfling (1980, Corollary, p. 32) we have the following result.

THEOREM 4 (Asymptotic Normality): *Under conditions* (C.1)–(C.9), *the asymptotic distribution of* $N^{1/2}(\hat{\theta} - \theta_0)$ *is* $N(0, \Sigma)$,

$$\Sigma \equiv E\left\{\left[\frac{\partial P}{\partial \theta}\right]\left[\frac{\partial P}{\partial \theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}_{\theta=\theta_0}^{-1}.$$

It should be remarked that the covariance matrix above can be estimated in the usual way when employing this distributional result for inference purposes. The informational equality holds here between the negative of the Hessian and the expected outer product gradient matrices. Consequently, one may employ the standard estimated Hessian, outer product gradient, or White's (1982) estimator to estimate the covariance matrix. As a result, with the quasi-likelihood treated as if it were the true likelihood, standard errors given from a conventional likelihood routine will be correct provided that the functional form for the value function is correct. Indeed, for purposes of inference, the estimated quasi-likelihood may be treated as if it were a likelihood function. In this manner, for example, one can conduct the usual likelihood ratio tests.

4.2. *Asymptotic Efficiency*

Having established asymptotic normality, we now turn to the issue of efficiency. In the next theorem, we show that under an independence assumption the estimator attains the efficiency bound as given in Chamberlain (1986) and Cosslett (1987).

THEOREM 5 (Asymptotic Efficiency): *For the index model given in equation* (1), *assume that x and u are independent and denote* Σ *as the covariance matrix given in Theorem* 4. *Then* Σ *attains the efficiency bound as specified in Chamberlain* (1986) *and Cosslett* (1987).

To prove Theorem 5, note that the inverse efficiency bound for this problem as given in Cosslett (1987, p. 587) is

$$(50) \qquad E\left[\frac{\operatorname{Var}\left[\partial v(x;\theta_0)/\partial\theta | v(x;\theta_0)\right] f_{u_0}^2\left[v(x;\theta_0)\right]}{P(y=1|x)\left[1-P(y=1|x)\right]}\right],$$

where f_{u_0} is the density for u_0. From Theorem 4,

$$(51) \qquad \Sigma^{-1} = E\left\{\left[\frac{\partial P}{\partial\theta}\right]\left[\frac{\partial P}{\partial\theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}_{\theta=\theta_0}.$$

When u_0 is independent of x, the theorem readily follows from the expression for $\partial P(\theta)/\partial\theta|_{\theta=\theta_0}$ in (38)–(42).

5. MONTE-CARLO RESULTS

The usefulness of the proposed estimator depends at least partly on its performance at moderate sample sizes, where perhaps the leading issues are efficiency loss relative to a correct fully parametric specification and effectiveness in situations where the correct parametric specification is unusual. To give an indication of the performance of the estimator in both these regards, we have performed a series of simulations, of which we report three here.

For all three simulations, the true model is given by $y_i^* = \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$ for $i = 1,\ldots,100$ and $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ otherwise. The x's are independently and identically distributed. For the compact support case, in the first two simulations, x_1 is a chi-squared variate with 3 degrees of freedom truncated at 6 and standardized to have zero mean and unit variance; x_2 is a standard normal variate truncated at ± 2 and similarly standardized. Notice that in these first two simulations we are comparing estimators without assuming that one of the x's has a density that is differentiable everywhere. For technical convenience, we formulated the theory under the assumption that there was one x with a density that was everywhere differentiable. It is possible to relax this assumption, and for purposes of evaluating the semiparametric estimator, we wanted to present some results that did not impose such smoothness. The third simulation has the x's constructed in the same manner except that they are not truncated.[6] In Design 1, the u_i's are standard normal; in Designs 2 and 3 they are normal with mean zero and variance $.25(1+v_i^2)^2$ where v_i is $\beta_1 x_{i1} + \beta_2 x_{i2}$. In all designs $\beta_1 = \beta_2 = 1$, and the u_i's are independently distributed.

[6] In the first two designs, x_1 is standardized by subtracting 2.348 and dividing by 1.511; x_2 is divided by .8796. Thus the range of x_1 is $(-1.551, 2.420)$ and x_2 is $(-2.274, 2.274)$. The third design is included to show a case where probit is markedly biased.

TABLE I

STANDARDIZED CUMULANTS OF SIMULATED ESTIMATES OF β
(Jackknife Estimates of Standard Errors in Parentheses.)

Design 1	κ_1	κ_2	$\kappa_3/\kappa_2^{3/2}$	κ_4/κ_2^2
Untrimmed	0.99958	0.01532	0.01400	0.82470
Semiparametric	(0.00392)	(0.00082)	(0.13875)	(0.38237)
Probit	0.99987	0.01196	−0.07293	0.21619
MLE	(0.00346)	(0.00056)	(0.09369)	(0.21359)
Design 2	κ_1	κ_2	$\kappa_3/\kappa_2^{3/2}$	κ_4/κ_2^2
Untrimmed	0.99758	0.01773	−0.08483	3.99493
Semiparametric	(0.00421)	(0.00138)	(0.31315)	(1.04304)
Probit	1.00508	0.04767	0.11877	0.82421
MLE	(0.00691)	(0.00254)	(0.12233)	(0.30838)
Design 3	κ_1	κ_2	$\kappa_3/\kappa_2^{3/2}$	κ_4/κ_2^2
Untrimmed	0.99195	0.02195	−0.23636	2.81254
Semiparametric	(0.00469)	(0.00153)	(0.25983)	(0.87196)
Probit	0.91525	0.06146	−0.15995	0.53852
MLE	(0.00784)	(0.00310)	(0.10444)	(0.23182)

To facilitate comparison between probit and the semiparametric estimator, we adopt the normalization $|\hat{\beta}_1| + |\hat{\beta}_2| = 2$. In our simulations, we compute probit (with an intercept) in the standard manner and impose the normalization on the result; for the semiparametric estimator the normalization is imposed in the estimation process. Thus we need to report results concerning only one number ($\hat{\beta}_1$).

In what follows, results are presented for the untrimmed semiparametric estimator and the probit MLE. There is a wide range of trimming specifications that have almost no effect on the estimates. Moreover, the estimate obtained without any trimming performed quite similar to that under the trimming that we employed. Accordingly, we report results for the semiparametric estimator obtained without probability or likelihood trimming (see the discussion in Section 3.1). Throughout, we estimated densities with adaptive local smoothing as defined in (C.8b), with local smoothing parameters defined without any trimming. The nonstochastic window component, h_N, we set at $N^{-1/(6.02)}$ to satisfy the restrictions placed on it.[7]

Table I presents estimates of the first and second cumulants and the third and fourth standardized cumulants of $\hat{\beta}_1$ for both designs derived from a simulation of size 1,000. The second row of the results for each estimator is the square root of the jackknife estimate of the variance of each cumulant, and thus provides a rough estimate of the standard error of the cumulant estimates.

[7] The corresponding pilot window component, h_{NP}, was set equal to h_N. All theoretical results can be shown to hold for this case. It should be noted, however, that after completing this study, we discovered that the results under local smoothing are much easier to establish when the pilot window component is set wider than h_N.

In examining Table I, the following facts are notable. First, the efficiency loss in Design 1 from using the semiparametric estimator is quite tolerable: the relative efficiency is 78%, which is comparable to the loss from using least absolute deviations regression rather than ordinary least squares when the disturbance is i.i.d. normal. Second, there is quite a considerable efficiency gain in using the semiparametric estimator in Design 2; a similar efficiency gain is found in Design 3 where probit is quite biased. Notice that by not truncating the distributions of the x's, the distribution of $x\beta_0$ is much more asymmetric in Design 3 than in Design 2. Third, the semiparametric estimator shows considerably more kurtosis than the probit estimator.

Despite the fact that the simulation results concerning the probit estimator of the normalized coefficients in Design 2 are consistent with that estimator being unbiased, the probit model cannot give an accurate picture of the behavior of the choice probability, $p(y=1|x)$. Figure 1.A plots $p(y=1|v)$ for the two designs. In Design 2, $p(y=1|v)$ is not monotonic in v, with reversals coming at $v=\pm 1$. Probit of course assumes that it is possible to construct a projection of x into v such that $p(y=1|v)$ is monotonic in v. In the current example this is not possible. In Figures 1.B and 1.C we have plotted true probabilities on the horizontal axis and predicted probabilities on the vertical axis for a typical run with 1,000 observations of Design 2.[8] The semiparametric probabilities track the 45° line reasonably well, and experiments with other sample sizes indicate a convergence to the 45° line as required by theory. In contrast, the probit probabilities are inaccurate and Figure 1.C is an accurate representation of their asymptotic behavior.

A further illustration of the behavior of the two estimators in this example is given by the perspective plots of Figure 2. The height of the surface is $\text{prob}(y=1|x)$; the x and y planes correspond to a rescaling of x_1 and x_2 (respectively) to the interval $(-1,1)$; the viewpoint is $(-6,-8,5)$. The semiparametric perspective plot of Figure 2.B reflects the features of the true probability perspective except in the extreme foreground (in which the surface curls slightly up due to sampling error at the extreme corner of the sample range) and in the far background (where the estimate is somewhat flatter than the truth). In contrast, the probit perspective plot is obviously too flat.

Clearly the results presented here are only indicative of the performance of the estimator in a few cases. But apparently there are some cases where the estimator is vastly superior to probit (even when probit is virtually unbiased) and some indication that the efficiency loss relative to a correctly specified parametric model can be relatively small. In addition, these exercises suggest that semiparametric estimation can serve as a specification test of parametric models, a topic we intend to pursue in future research.

[8] For our estimator the fitted probabilities converge to the true probabilities at a rate proportional to $N^{1/2}h_N \cong N^{1/3}$. Thus 1000 observations are required to get about the same precision in fitted probabilities as is provided by 100 observations for parameter estimation. At 100 observations, there is not much difference between plots analogous to Figures 1.B and 1.C.

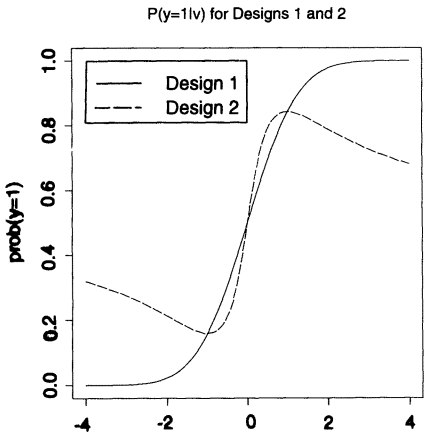

FIGURE 1.A

v

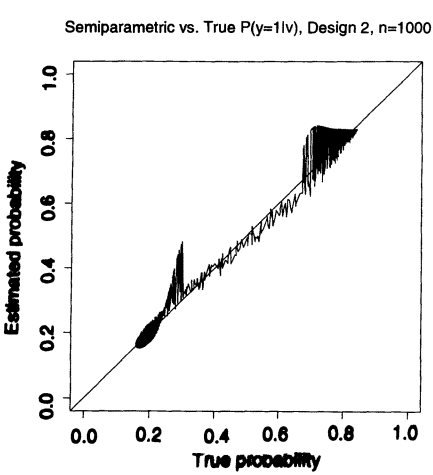

FIGURE 1.B

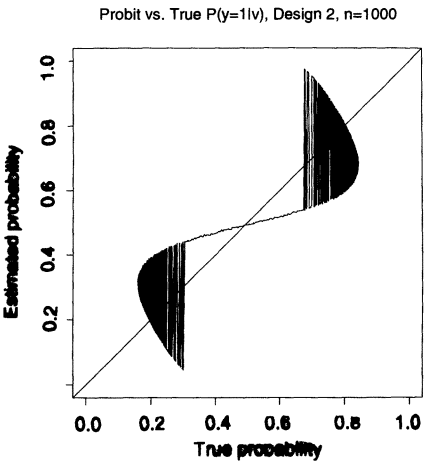

FIGURE 1.C

6. CONCLUSIONS

As discussed above, we have formulated a "likelihood based" estimator for the binary discrete choice model under minimal distributional assumptions. We have shown that in large samples this estimator is consistent, $N^{1/2}$ normally distributed, and that it attains an asymptotic efficiency bound.

We also evaluated the estimator in a Monte-Carlo study by comparing it with the probit estimator for several model specifications. In terms of estimated parameters, we found that the semiparametric estimator had a small efficiency loss relative to probit when the probit model was correct. For a "substantial" perturbation of the probit model, the semiparametric estimator strongly dominated the probit model. Although this paper has focused on parameter esti-

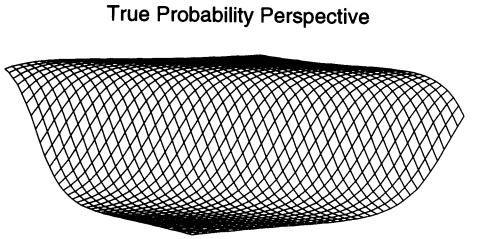

FIGURE 2.A

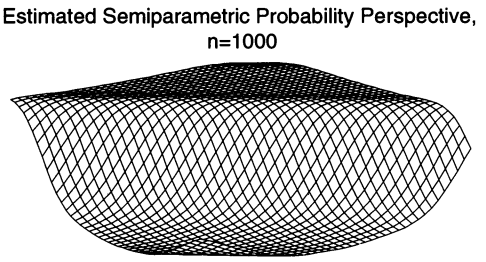

FIGURE 2.B

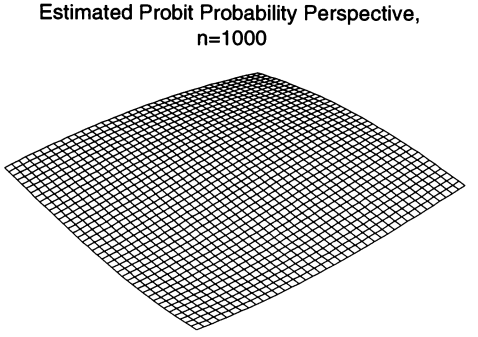

FIGURE 2.C

mates, it is interesting to note that in many situations choice probabilities will be inconsistently estimated by probit, but consistently estimated by a semiparametric estimator. In several heteroscedastic designs reported here, we found that the parametric probit model provided substantially worse probability estimates than the semiparametric procedure.

Restricting attention to the binary response model, it is possible to extend the above results in several respects. First, there can be a finite number of points at which the distribution of the index is not "smooth." For example, if the index is a linear combination of independent uniform random variables, then the density for the index will have kink points at which derivatives do not exist. These points can be detected by comparing left and right numerical derivatives.

Employing the same type of trimming function (τ) used to detect small densities, we can then downweight observations for which numerical right and left derivatives are "sufficiently" far apart. All asymptotic results then hold.[9]

As a second extension, it should be noted that the results above were obtained under a compact support assumption for the density of the index. To control for problematic small densities, in the above proof we downweighted those observations at which the index densities became small. In so doing, the trimming functions did not depend on whether or not the index had a compact support. Accordingly, under such density trimming, one would expect that the above arguments would, with suitable modifications, hold when the index has an infinite support.

In obtaining the results outlined above, the parametric estimator was examined for the binary discrete choice model. However, it extends to the multinomial case. For example, consider the trinary choice problem. Let alternative A_k, $k = 1, 2, 3$, have indirect utility U_k to a given individual. Let u_k be the corresponding measure of average utility, and define the average utility differences: $v_{12} \equiv u_1 - u_2$ and $v_{13} \equiv u_1 - u_3$. Then with these utility differences serving as aggregators, we may write choice probabilities as

$$(52) \qquad P(A_1|x) = P(U_1 > U_2, U_1 > U_3|x) = P(U_1 > U_2, U_1 > U_3|v_{[1]})$$

$$= P(A_1)\, g_1(v_{[1]}|A_1)/g(v_{[1]}), \quad v_{[1]} \equiv (v_{12}, v_{13}).$$

It is now possible to construct a multinomial quasilikelihood as above in terms of estimated probability functions. Notice that in this example we would require bivariate density estimation. In general, as the number of alternatives increases, the estimation problem becomes more difficult, and the theoretical argument becomes more delicate (see, for example, Lee (1989)).

The estimator obtained here can also be extended to ordered single index models. Here, unlike the mlutiple alternatives case (which involves multiple indices), the extension is direct. For each interval of the dependent variable, we may estimate probabilities as above under a single index. Then form a quasilikelihood for the ordered case that is analogous to that for the ordered parametric model. Given the results obtained here, it is relatively straightforward to establish consistency and asymptotic normality for the quasi-maximum likelihood estimator.

Bellcore, 445 South Street, P.O. Box 1910, Morristown, NJ 07962, U.S.A.
and
Nuffield College, Oxford, OX1 1NF, U.K.

Manuscript received March, 1988; final version received September, 1992.

[9] In general, suppose that there are a finite number of "problematic" points at which the maintained assumptions do not hold. Then, provided that one can detect such points, the theory will hold if these points are appropriately downweighted.

APPENDIX A: BIAS REDUCING KERNELS

In Appendix A, we provide the asymptotic results for bias reducing kernels. In Appendix B, we outline those additional arguments required to obtain the same asymptotic results under locally-smoothed kernels. Throughout, we let $g_{v|y}(v_i;\theta)$ be the conditional density for $v \equiv v(x;\theta)$ conditioned on the discrete variable, y, and evaluated at $v_i \equiv v(x_i;\theta)$. It is convenient to state convergence results in terms of the estimator for $\Pr(y)g_{v|y}(v_i;\theta) \equiv g_{yv}(v_i;\theta)$. From (C.8) define this estimator as

$$\text{(A1)} \qquad \hat{g}_{yv}(v_i;\theta) \equiv \sum_{j\neq 1}^{N} \frac{1[y_j = y]}{h_N} K\left[\frac{v_i - v_j}{h_N}\right] \Big/ (N-1), \quad y = 0,1.$$

We will require higher order derivatives of these estimates throughout the proofs below. Notationally, define the rth order derivative of any function g with respect to z by

$$\text{(A2)} \qquad D_z^r[g] \equiv \begin{cases} g, & r = 0, \\ \partial^r g/(\partial z)^r, & r = 1,2,\ldots. \end{cases}$$

When z is a vector, interpret the derivative as the matrix of all rth order partials and cross-partials. We will also denote the above derivative by g^r when the argument of g with respect to which we take the derivative is clear from the context.

Throughout, we assume the basic conditions (C.1)–(C.9) given in Section 2 of the text and denote c as a positive generic constant. Given the convergence rates established in Lemmas 1–2 and known trimming (Lemma 3), consistency and asymptotic normality will follow for the proposed estimator. We begin in Lemma 1, the proof of which is based on Bhattacharya's (1967) method, by determining a uniform convergence rate of an estimate to its expectation.

LEMMA 1 (Convergence of Estimates to Expectations): *Let w be a K dimensional vector and assume that $m(w)$ is a sample average of terms $m(w;z_i)$, $i = 1,\ldots,N$, where $\{z_i\}$ are i.i.d. Assume that with $h_N \to 0$ we have uniformly over N:*

$$h_N^{r+1}|m(w;z_i)| < c, \quad r+1 > 0, \quad \textit{and} \quad h_N^s|\partial m(w;z_i)/\partial w| < c, \quad s > 0.$$

Let $E[m(w)]$ be the expectation of $m(w)$ taken over the distribution of z_i, $i = 1,\ldots,N$. Then with $w \in \mathscr{W}$ a compact set, and for any $\alpha > 0$,

$$N^{(1-\alpha)/2}h_N^{r+1} \sup_w |m(w) - E[m(w)]| \to 0 \quad a.s.$$

PROOF OF LEMMA 1: With w_1 and w_2 any two values of w and $w^+ \in [w_1, w_2]$,

$$m(w_1) - m(w_2) = [\partial m(w^+)/\partial w](w_1 - w_2).$$

Since $h_N^s|[\partial m(w^+)/\partial w]| \leqslant c$, with w_{1t} and w_{2t} as elements of w_1 and w_2,

$$|m(w_1) - m(w_2)| \leqslant c \sum_{k=1}^{K} |w_{1k} - w_{2k}|/h_N^s.$$

Therefore, with $\varepsilon_N > 0$ and $\delta_N \equiv \varepsilon_N h_N^s/(3cK)$,

$$|w_{1k} - w_{2k}| < \delta_N \Rightarrow |m(w_1) - m(w_2)| < \varepsilon_N/3.$$

Since $w \in \mathscr{W}$, a compact space, without loss of generality let $\mathscr{W}$ be contained in a hypercube whose sides each have "length" one. Then, following Bhattacharya (1967), cover $\mathscr{W}$ with sets $\mathscr{W}_{jN}$, $j = 1,\ldots,b_N$, where $(w_1, w_2) \in \mathscr{W}_{jN} \Rightarrow |w_{1k} - w_{2k}| < \delta_N$, $i = 1,\ldots,K$. Here, with $\text{int}(t)$ as the least

integer upper bound on t, $b_N = \text{int}[1/(\delta_N)^K]$. Then, for $w_j \in \mathscr{W}_{jN}$,

$$\begin{aligned}
&\sup|m(w) - E[m(w)]| \\
&\quad \leqslant \max_{j=1,\ldots,b_N}\left[\sup_{w\in\mathscr{W}_{jN}}\left|[m(w) - m(w_j)]\right.\right. \\
&\qquad\qquad \left.\left. + [E[m(w_j) - m(w)]] + [m(w_j) - E[m(w_j)]]\right|\right] \\
&\quad \leqslant (2/3)\varepsilon_N + \max_{j=1,\ldots,b_N}\left|[m(w_j) - E[m(w_j)]]\right|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&p\left[\sup_w|m(w) - E[m(w)]| > \varepsilon_N\right] \\
&\quad \leqslant P\left[\max_j\left|[m(w_j) - E[m(w_j)]]\right| > \varepsilon_N/3\right] \\
&\quad \leqslant \sum_{j=1}^{b_N} P\left[\left|[m^*(w_j) - E[m^*(w_j)]]\right| > h_N^{r+1}\varepsilon_N/3\right], \\
&m^*(w_j) \equiv h_N^{r+1}m(w_j).
\end{aligned}$$

As noted by Bhattacharya, from Hoeffding (1963; Theorem 1, p. 15 and its extension, p. 25) it follows that since m^* is an average of i.i.d. bounded random variables,

$$P\left[\left|[m^*(w_j) - E[m^*(w_j)]]\right| > h_N^{r+1}\varepsilon_N/3\right] \leqslant 2\exp\left[-c'Nh_N^{2r+2}\varepsilon_N^2\right],$$

where c' is a positive constant. Therefore,

$$P\left[\sup_w|m(w) - E[m(w)]| > \varepsilon_N\right] \leqslant 2b_N\exp\left[-c'Nh_N^{2r+2}\varepsilon_N^2\right].$$

With ε_N replaced by $[N^{(1-\alpha)/2}h_N^{r+1}]^{-1}\varepsilon$, where ε is a finite constant that does not depend on N, the sum of the last term over $N = 1,\ldots,\infty$ is finite; the result now follows. *Q.E.D.*

In proving that an estimate converges uniformly to the truth, the strategy is to use Lemma 1 to show that the estimate converges uniformly to its expectation and then that its expectation converges uniformly to the truth. In this manner, Lemma 2 obtains rates for estimated densities and derivatives.

LEMMA 2 (Uniform Convergence): *Uniformly on* $v(t;\theta)$ *and* θ,

$$\left|D_\theta^r[\hat{g}_{yv}[v(t;\theta);\theta]] - E\{D_\theta^r[\hat{g}_{yv}[v(t;\theta);\theta]]\}\right| = O_p\left(N^{-1/2}h_N^{-(r+1)}\right), \quad r = 0,1,2; \tag{2.1}$$

$$\left|E\{D_\theta^r[\hat{g}_{yv}[v(t;\theta);\theta]]\} - D_\theta^r[g_{yv}[v(t;\theta);\theta]]\right| = O_p\left(h_N^2\right), \quad r = 1,2; \tag{2.2}$$

$$\left|E[\hat{g}_{yv}[v(t;\theta);\theta]] - g_{yv}[v(t;\theta);\theta]\right| = O_p\left(h_N^4\right). \tag{2.3}$$

PROOF OF LEMMA 2: The proof of (2.1) immediately follows from Lemma 1. To establish (2.2), with $z \equiv [v(x;\theta) - v(t;\theta)]/h_N$ for fixed t,

$$\begin{aligned}
E\{D_\theta^r[\hat{g}_{yv}[v(t;\theta);\theta]]\} &= D_\theta^r E\{[\hat{g}_{yv}[v(t;\theta);\theta]]\} \\
&= \Pr(y)D_\theta^r\int_{-\infty}^{\infty}K(z)g_{v|y}[v(t;\theta) + h_N z;\theta]/dz \\
&= \Pr(y)\int_{-\infty}^{\infty}K(z)\,D_\theta^r g_{v|y}[v(t;\theta) + h_N z;\theta]\,dz.^{10}
\end{aligned}$$

[10] Note that in this representation we have moved the derivative operator in and out of the expectations operator. In each case, the function being integrated is differentiable in a neighborhood of θ_0 for almost all x and satisfies an appropriate dominance condition.

Next, expand the integrand above as a function of h_N and terminate the expansion at the h_N^2 term. From the symmetry of the kernel, the h_N term vanishes. The result now follows from the regularity conditions on the kernel and density derivatives.

The proof of (2.3) is also based on a Taylor series expansion in h_N, where the expansion is taken up to the h_N^4 term. From symmetry of the kernel, the h_N and h_N^3 terms vanish. Since $\int z^2 K(z) = 0$ (the bias-reducing kernel assumption), the h_N^2 term also vanishes. The result now follows from the assumed regularity conditions on the kernel and density derivatives. *Q.E.D.*

Apart from the trimming function, Lemma 2 provides the asymptotic behavior of the components of the (estimated) quasi-likelihood. Lemma 3 below characterizes the large-sample behavior of the trimming function.

LEMMA 3 (Estimated vs. Known Trimming): *Let* $\hat{g}_{yi}(\hat{\theta}_P) \equiv \hat{g}_{yv}[v(x_i;\hat{\theta}_P);\hat{\theta}_P]$, *where* $\hat{\theta}_P$ *is a preliminary or pilot estimator for which* $(\hat{\theta}_P - \theta_0) = O_P(N^{-1/3})$.[11] *Define a smooth trimming function as*

$$\tau(t) \equiv 1/[1+e^{z(t)}], \quad z(t) \equiv \left(h_N^b - t\right)/h_N^c,$$

where $0 < b < c < 1/3$ *and* $N^{-1/6} < h_N < N^{-1/8}$. *Let* $\hat{\tau}_{iy} \equiv \tau[\hat{g}_{yi}(\hat{\theta}_P);\varepsilon]$ *and* $\tau_{iy} \equiv \tau[g_{yi};\varepsilon]$, $g_{yi} \equiv g_{yi}(v_i;\theta_0)$. *Then,* τ_{iy} *has the following representation*:

$$\begin{aligned}(a)\qquad \hat{\tau}_{iy} \equiv{}& \tau_{iy} + h_N^{-c}\left(\hat{g}_{yi}(\theta_0) - g_{yi}\right)(1-\tau_{iy})\tau_{iy} + h_N^{-2c}\left(\hat{g}_{yi}(\theta_0) - g_{yi}\right)^2 C_{iy}(1-\tau_{iy})\tau_{iy}\\ &+ h_N^{-c}\left(\hat{\theta}_P - \theta_0\right) D_\theta^1 \hat{g}_{yi}(\theta_0)(1-\tau_{iy})\tau_{iy} + h_N^{-c}\left(\hat{\theta}_P - \theta_0\right)^2 D_\theta^2 \hat{g}_{yi}(\theta^+)(1-\tau_{iy})\tau_{iy}\\ &+ h_N^{-2c}\left(\hat{\theta}_P - \theta_0\right) 2\left[\hat{g}_{yi}(\theta_0) - g_{yi}\right] D_\theta^1 \hat{g}_{yi} C_{iy}(1-\tau_{iy})\tau_{iy} + o_P(N^{-1/2}),\end{aligned}$$

where $C_{iy} = O(1)$, *uniformly in* i, y *and where the remainder term is* $o_P(N^{-1/2})$ *uniformly in* i. *With* $\hat{\tau}_{iy}(\theta) \equiv \tau[\hat{g}_{yi}(\theta);\varepsilon]$, $g_{yi}(\theta) \equiv g_{yi}(v(x_i;\theta))$, *and* $\tau_{iy}(\theta) \equiv \tau[g_{yi}(\theta);\varepsilon]$:

$$(b)\qquad \hat{\tau}_{iy}(\theta) \equiv \tau_{iy}(\theta) + h_N^{-c}\left[\left(\hat{g}_{yi}(\theta) - g_{yi}(\theta)\right)\right]\left[1-\tau_{iy}(\theta)\right]\tau_{iy}(\theta) + o_P\left(N^{-1/2}h_N^{-2}\right),$$

where the order of the remainder term is uniform in i, y, *and* θ.

PROOF OF LEMMA 3: For (a), the proof follows by first expanding $\hat{\tau}_{iy}$ as a function of $\hat{g}_{yi}(\hat{\theta}_P)$ about g_{yi} and then expanding $\hat{g}_{iy}(\hat{\theta}_P)$ as a function of $\hat{\theta}_P$ about θ_0. For (b), the proof follows an analogous argument. *Q.E.D.*

Employing Lemmas 1–3, we can now determine convergence rates for estimated probabilities and their derivatives. In obtaining these results, it is convenient to do so in terms of adjusted or trimmed probability functions. With $0 < \varepsilon' < 1$, $v \equiv v(x;\theta)$, and $g_{yv} \equiv g_{yv}(v;\theta)$, let adjustment factors, which satisfy the restrictions in (22) of subsection 3.1, be given as

$$(A3)\qquad \delta_{yN} \equiv h_N^{\varepsilon'}[e^z/(1+e^z)], \quad z \equiv \left(h_N^{(\varepsilon'/5)} - g_{yv}\right)/h_N^{(\varepsilon'/4)}, \quad \text{and} \quad \delta_N \equiv \delta_{0N} + \delta_{1N}.$$

With $\hat{g}_{yv}(v;\theta)$ replacing $g_{yv}(v;\theta)$, define $\hat{\delta}(v;\theta)$ and $\hat{\delta}_N(v;\theta)$ analogously. Then, with $\hat{g}_v(v;\theta) \equiv \hat{g}_{0v}(v;\theta) + g_{1v}(v;\theta)$ as the estimated marginal density for v, define the following probability functions:

$$(A4)\qquad \text{Estimate:}\quad \hat{P}(v;\theta) \equiv \left[\hat{g}_{1v}(v;\theta) + \hat{\delta}_{1N}(v;\theta)\right]/\left[\hat{g}_v(v;\theta) + \hat{\delta}_N(v;\theta)\right];$$

$$(A5)\qquad \text{Trimmed:}\quad P_N(v;\theta) \equiv \left[g_{1v}(v;\theta) + \delta_{1N}(v;\theta)\right]/\left[g_v(v;\theta) + \delta_N(v;\theta)\right];$$

$$(A6)\qquad \text{True:}\quad P(v;\theta) \equiv g_{1v}(v;\theta)/g_v(v;\theta).$$

Lemma 4 now obtains results in terms of these three functions.

[11] This assumption can be relaxed at the expositional expense of incorporating higher order terms into the series representation of Lemma 4. Manski's (1975) maximum score estimator, Horowitz's (1992) smooth version of this estimator, and Han's rank estimator (see Sherman (1992)) all converge to the truth at least as fast as $O_P(N^{-1/3})$.

LEMMA 4 (Convergence of Estimated Probabilities and their Derivatives): *Let* ε', $0<\varepsilon'<1$, *be the adjustment parameter in* (A3) *employed to control probability functions in* (A4)–(A5). *Then, for* $N^{-1/(6+2\varepsilon')}<h_N<N^{-1/8}$, *and under the conditions of Lemmas* 1–2, *uniformly in* $v\in\mathscr{V}$ *and* $\theta\in\Theta$

$$(4.1a)\qquad h_N^{\varepsilon'}\left|D_\theta^r\left\{\hat{P}[v(t,\theta);\theta]\right\}-D_\theta^r\left\{P_N[v(t;\theta);\theta]\right\}\right|=O_P\left(N^{-1/2}h_N^{-(r+1)}\right),\quad r=0,1,2;$$

$$(4.1b)\qquad h_N^{\varepsilon'}\left|\hat{P}[v(t;\theta);\theta]^{-1}-P_N[v(t;\theta);\theta]^{-1}\right|=O_P\left(N^{-1/2}h_N^{-1}\right).$$

Let $\mathscr{V}_N\equiv\{v\colon g_{0v}(v;\theta)>h_N^{\varepsilon},\ g_{1v}(v;\theta)>h_N^{\varepsilon}\}$, *where with* ε' *as given in* (A3), $0<\varepsilon<(\varepsilon'/5)$. *Note that under this condition, the probability adjustment factors,* δ_{0N} *and* δ_{1N} *vanish exponentially. Then, under the above conditions, uniformly on* $\mathscr{V}_N,\Theta$:

$$(4.2a)\qquad h_N^{\varepsilon}\left|D_\theta^r\left\{\hat{P}[v(t;\theta);\theta]\right\}-D_\theta^r\left\{P[v(t;\theta);\theta]\right\}\right|=O_P\left(N^{-1/2}h_N^{-(r+1)}\right),\quad r=0,1,2,$$

$$(4.2b)\qquad h_N^{\varepsilon}\left|\hat{P}[v(t;\theta);\theta]^{-1}-P[v(t;\theta);\theta]^{-1}\right|=O_P\left(N^{-1/2}h_N^{-1}\right).$$

PROOF OF LEMMA 4: Convergence rates for estimated probability functions and their derivatives are similar to those for estimated densities and their derivatives. The only difference is that for the former rates must be decreased by the rate at which the relevant denominators tend to zero. We illustrate the argument for $r=2$ in (4.1a); other results are proved analogously.[12] A typical term from $D_\theta^r\{\hat{P}[v(t;\theta);\theta]\}-D_\theta^r\{P_N[v(t;\theta);\theta]\}$ is given by[13]

$$T\equiv\left[D_\theta^2\left\{\hat{g}_{1v}+\hat{\delta}_{1N}\right\}\right]/\left[\hat{g}_v+\hat{\delta}_n\right]-\left[D_\theta^2\{g_{1v}+\delta_{1N}\}\right]/[g_v+\delta_N].$$

With $\Delta_N\equiv[(g_v+\delta_N)-(\hat{g}_v+\hat{\delta}_N)]/[g_v+\delta_N]$, from the geometric expansion of an inverse,

$$1/\left(\hat{g}_v+\hat{\delta}_N\right)=[1/(g_v+\delta_N)][1-\Delta_N]^{-1}=[1/(g_v+\delta_N)]\left[1+\Delta_N+\Delta_N^2(1-\Delta_N)^{-1}\right].$$

From Lemmas 2–3 and the fact that $(g_v+\delta_N)^{-1}=O(h_N^{-\varepsilon'})$, substituting this expansion into T yields

$$T=\left[D_\theta^2\left(\hat{g}_{1v}+\hat{\delta}_{1N}\right)-D_\theta^2(g_{1v}+\delta_{1N})\right]/[g_v+\delta_N]+o_P\left(h_N^{-\varepsilon'}N^{-1/2}h_N^{-3}\right).$$

The result now follows from the convergence rates for the derivatives in Lemma 2 and the bound on $g_v+\delta_N$. *Q.E.D.*

From Lemmas 3–4 consistency now follows by showing that the quasi-likelihood converges uniformly in θ to a function uniquely maximized at θ_0. In obtaining this result, we make all assumptions in Lemmas 3–4. In particular, with ε', $0<\varepsilon'<1$, as the probability trimming parameter, the window, h_N, is selected such that $N^{-1/(6+2\varepsilon')}<h_N<N^{-1/8}$.

PROOF OF THEOREM 3 (Consistency): Let the estimated probabilities, $\hat{P}_i$, be defined in Lemma 4 and the trimming functions, τ_{iy} and $\hat{\tau}_{iy}$, as in Lemma 3. Let $\tau_i\equiv\tau_{i0}\tau_{i1}$ and $\hat{\tau}_i\equiv\hat{\tau}_{i0}\hat{\tau}_{i1}$. Then, the

[12] In the argument for probability functions, it is convenient to first write

$$\hat{P}-P_N=\left[\left[\left(\hat{g}_{1v}+\hat{\delta}_{1N}\right)-(g_{1v}+\delta_{1N})\right]+P_N\left[\left(\hat{g}_v+\hat{\delta}_N\right)-(g_v+\delta_N)\right]\right]\Big/\left[\hat{g}_v+\hat{\delta}_N\right].$$

For reciprocals, with $\Delta_N\equiv(P_N-\hat{P})/P_N$, write

$$1/\hat{P}=(1/P_N)[1-\Delta_N]^{-1}=(1/P_N)\left[1+\Delta_N+\Delta_N^2(1-\Delta_N)^{-1}\right].$$

If P is bounded from below by $\underline{P}>0$, then P_N is also bounded from below and the convergence rate of $1/\hat{P}$ to $1/P_N$ will be the same as that for $\hat{P}$ to P_N. If P is not bounded away from zero, the convergence rate will be slowed by an additional factor of order $h_N^{\varepsilon'}$.

[13] There will be other terms with squared densities in the denominator. For example, one such typical term is given by

$$\left[D_\theta^1\left(\hat{g}_{1v}+\hat{\delta}_{1N}\right)\right]^2/\left[\hat{g}_v+\hat{\delta}_N\right]^2-\left[D_\theta^1(g_{1v}+\delta_{1N})\right]^2/\left[g_v+\delta_N^2\right].$$

The convergence rate increases when first rather than second derivatives appear in the numerator, and this increase more than compensates for the additional density term in the denominator.

quasi-likelihood is given as

$$\hat{Q}(\theta) \equiv \sum (\hat{\tau}_i/2)\Big[y_i \ln\big[\hat{P}_i(\theta)^2\big] + (1-y_i)\ln\Big[\big(1-\hat{P}_i(\theta)\big)^2\Big]\Big]/N.$$

$$\hat{Q}_1(\theta) \qquad\qquad \hat{Q}_0(\theta)$$

Viewing $\hat{Q}_1(\theta)$ as a function of $\hat{P}_i(\theta)$, from a Taylor series expansion about $P_N(v_i;\theta)$ as given in (A5) and from Lemmas 2–4

$$\sup_\theta \big|\hat{Q}_1(\theta) - Q_{1N}(\theta)\big| \xrightarrow{p} 0, \quad Q_{1N}(\theta) \equiv \sum \tau_i\big[y_i \ln\big[P_N(v_i;\theta)\big]\big]/N.$$

In obtaining this result, note that since $P_N > 0$, $(1/2)\ln[P_N^2] = \ln[P_N]$. With $q_{1N}(v_i;\theta) \equiv y_i \ln[P_N(v_i;\theta)]$ and $b_i(\theta) \equiv \{1 - \tau[g_{1v}(v_i;\theta);\varepsilon'']\}$ for $\varepsilon'' < \varepsilon'$ and τ as defined in (C.7), write

$$Q_{1N}(\theta) \equiv \sum b_i(\theta)\tau_i q_{1N}(v_i;\theta)/N + \sum \big[1 - b_i(\theta)\big]\tau_i q_{1N}(v_i;\theta)/N.$$

For the b_i-terms, note that uniformly in (i,θ), $\sup|q_{1N}| = O(1)$ if P and $1-P$ are uniformly bounded away from zero.[14] Thus,

$$\sup_\theta \Big| \sum b_i\tau_i q_{1N}(v_i;\theta)/N \Big| \leqslant \sup_{i,\theta}|q_{1N}| \sup_\theta \sum b_i/N \xrightarrow{p} 0.$$

For the $(1-b_i)$-terms, we may take $g_v \geqslant g_{1v} \geqslant h_N^{\varepsilon}$ for $\varepsilon < \varepsilon'/5$, as any terms for which this condition does not hold will vanish exponentially. Then, from Lemma 4,

$$\sum (1-b_i)\tau_i q_{1N}(v_i;\theta)/N - \sum (1-b_i)\tau_i q_{1i}(\theta)/N \xrightarrow{p} 0 \quad \text{uniformly in } \theta,$$

where q_1 is obtained from q_{1N} by replacing $P_N(v_i;\theta)$ with $P(v_i;\theta)$, the "true" probability function in (A6). Then, since

$$\sum (1-b_i)\tau_i q_{1i}(\theta)/N - \sum q_{1i}/N \xrightarrow{p} 0 \quad \text{uniformly in } \theta,$$

$$\sup_\theta \big|\hat{Q}_1(\theta) - Q_1(\theta)\big| \xrightarrow{p} 0, \quad Q_1 \equiv \sum q_{1i}/N.$$

By an analogous argument, with $\hat{Q}_0(\theta)$ containing the $y_i = 0$ terms of the quasi-likelihood and $Q_0(\theta)$ as the corresponding terms evaluated at the true probability functions, $\hat{Q}_0(\theta)$ converges in probability, uniformly in θ to $Q_0(\theta)$. By standard arguments, $Q_1 + Q_0$ converges in probability, uniformly in θ, to its expectation, with maximum, θ_*, satisfying for almost all x

$$P(y=1|x) = P\big(y=1|v(x;\theta_0)\big) \equiv P_i(\theta_0) = P_i(\theta_*),$$

[14] For P and $1-P$ uniformly bounded away from zero, we must show that P_N and $1-P_N$ are bounded away from zero. Recalling that $g_v = g_{1v}/P$, write P_N as

$$P_N = [g_{1v} + \delta_{1N}]/[g_v + \delta_N] = P[g_{1v} + \delta_{1N}]/[g_{1v} + P(\delta_{1N} + \delta_N)]$$
$$> P\big[g_{1v} + h_N^{\varepsilon'}(1-\tau_1)\big]/\big[g_{1v} + h_N^{\varepsilon'}(1-\tau_0) + h_N^{\varepsilon'}(1-\tau_1)\big] \equiv P_N^*.$$

As the above expression is an increasing function of g_{1v}, for $g_{1v} > .5h_N^{(\varepsilon'/5)}$,

$$P_N^* > P\big[.5h_N^{\varepsilon'/5}\big]/\big[.5h_N^{\varepsilon'/5} + h_N^{\varepsilon'}(1-\tau_0) + h_N^{\varepsilon'}(1-\tau_1)\big]$$
$$> P\big[1/\big(1 + 4h_N^{4\varepsilon'/5}\big) > P/5.$$

For $g_{1v} < .5h_N^{(\varepsilon'/5)}$, as P_N^* is increasing in g_{1v},

$$P_N^* > P(1-\tau_1)/\big[(1-\tau_0) + (1-\tau_1)\big] > P(1-\tau_1)/2.$$

As $(1-\tau_1)$ converges exponentially to one in this region, the result follows. The argument for $(1-P_N)$ is analogous. It should be remarked that the consistency will still hold under a more delicate argument if P is permitted to approach zero and/or one. In this case, for the b-terms, one must show that $\ln(P_N)$ behaves like $O(h_N^{\alpha})$, which is dominated by the rate at which the average of the b_i-terms tends to zero. For the $(1-b)$-terms, a tail assumption on P is required.

where the first equality holds because $v(x_i;\theta_0)$ aggregates information and the second equality characterizes a maximum. From the identification assumption, $\theta_* = \theta_0$; consistency then follows. *Q.E.D.*

To prove asymptotic normality, let $\hat{H}(\theta)$ and $\hat{G}(\theta)$ denote the Hessian and gradient respectively for the quasi-likelihood of Theorem 3. From a standard Taylor series expansion for the gradient:

$$\text{(A7)} \qquad N^{1/2}\left(\hat{\theta}-\theta_0\right) = -\left[\hat{H}(\theta^+)\right]^{-1}\left[N^{1/2}\hat{G}(\theta_0)\right],$$

where $\theta^+ \in [\hat{\theta}, \theta_0]$. Asymptotic normality will follow from a convergence result for the Hessian (Lemma 5) and normality for the normalized gradient (Lemma 6).

LEMMA 5 (Convergence of the Hessian): *With ε' as the probability adjustment parameter in* (A3), *assume $N^{-1/(6+2\varepsilon')} < h_N < N^{-1/8}$. Then, under the conditions of Lemmas* 3–4 *and with $\hat{\theta}^+$ as any value for θ such that $\hat{\theta}^+ \in [\hat{\theta}, \theta_0]$,*

$$-\hat{H}(\theta^+) \xrightarrow{p} E\left\{\left[\frac{\partial P}{\partial \theta}\right]\left[\frac{\partial P}{\partial \theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}_{\theta=\theta_0}.$$

PROOF OF LEMMA 5: As in the consistency argument, from Lemmas 2–4 we may replace (in probability, uniformly in θ) the estimated probability function and its derivatives with $P_N(v_i;\theta)$ in (A5) and its derivatives. Following the consistency argument, we may also replace the likelihood trimming function, $\hat{\tau}_i$, with the corresponding true function, τ_i. Denote $H_N(P_N,\theta)$ as the resulting Hessian matrix. We will show that $H_N(P_N,\theta)$ converges uniformly in θ to its limiting expectation (Lemma 1), which we will show is finite. With $\bar{H}(\theta)$ denoting this limiting expectation, it will then follow that

$$H_N(P_N,\theta^+) \xrightarrow{p} \bar{H}(\theta_0).$$

At θ_0, τ_i keeps densities sufficiently far from zero to insure that $\bar{H}(\theta_0)$ has the form required by the lemma.

Proceeding, with $v_i \equiv v(x_i;\theta)$ and $w_{iN} \equiv \{P_N(v_i;\theta)[1-P_N(v_i;\theta)]\}^{-1}$,

$$H_N(P_N,\theta) = A + B,$$

$$A \equiv -\sum_i \tau_{iN} w_{iN}\left[D_\theta^1 P_N(v_i;\theta)\right]\left[D_\theta^1 P_N(v_i;\theta)\right]'/N,$$

$$B \equiv \sum_i \tau_{iN}\left[y_i - P_N(v_i;\theta)\right] D_\theta^1\left\{w_{iN} D_\theta^1 P_N(v_i;\theta)\right\}/N.$$

For the A-term, since $[D_\theta^1 P_N(v_i;\theta)][D_\theta^1 P_N(v_i;\theta)]'$ is $O(h_N^{-2})$, one can show from Lemma 1 that $A(\theta) - EA(\theta) \to 0$ uniformly in θ, where $EA(\theta)$ is finite.[15] Therefore, since $\hat{\theta}$ is consistent, the

[15] To demonstrate the argument, it suffices to consider one of the terms comprising this expectation. In so doing, to simplify the notation, we will consider the scalar case. Proceeding with one of these terms,

$$E\left[\left(\left[\partial g_{1v}(v)/\partial\theta + \partial\delta_N(v;\theta)/\partial\theta\right]/\left[g_v(v;\theta)+\delta_N(v;\theta)\right]\right)^2\right]$$

$$\leq 2E\left[\left(\left[\partial g_{1v}(v)/\partial\theta\right]^2 + \left[\partial\delta_{1N}(v;\theta)/\partial\theta\right]^2\right)/\left(g_v(v;\theta)+\delta_N(v;\theta)\right)^2\right].$$

For the second term, $\partial\delta_{1N}/\partial\theta = h_N^{(\varepsilon'-\varepsilon'/4)}\tau(1-\tau)$, which vanishes exponentially unless $g_{1v} = O(h_N^{(\varepsilon'/5)})$, in which case $[\partial\delta_{1N}(v;\theta)/\partial\theta]^2/[g_v(v;\theta)+\delta_N(v;\theta)]^2$ is $o(1)$. An upper bound for the first term above is

$$\int\left[E\left(\left[\partial g_{1v}(v;\theta)/\partial\theta\right]^2 | v\right)/\left[g_v(v;\theta)\right]\right] dv.$$

As shown in (14) of Section 2, $[\partial g_{1v}(\bar{v};\theta)/\partial\theta]^2 = O([\partial g_{1v}(\bar{v};\theta)/\partial v]^2)$. Since g_{1v} has four bounded derivatives with respect to v, when g_{1v} tends to zero, then so must these derivatives. One can show that the squared first derivative converges to zero faster than does the density itself (see e.g. the example in (15), Section 2). The claim now follows.

uniform convergence arguments above insure that for $\theta^+ \in [\theta_0, \hat{\theta}]$

$$A(\theta^+) \to \lim\left[E\left[A(\theta_0)\right]\right] = E\left\{\left[\frac{\partial P}{\partial \theta}\right]\left[\frac{\partial P}{\partial \theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}_{\theta=\theta_0},$$

where the latter claim follows at $\theta = \theta_0$ because τ_i converges exponentially to zero unless both conditional densities are appropriately bounded away from zero.

Employing a similar argument, the B-term also converges uniformly in θ to its expectation, which at θ_0 is zero. *Q.E.D.*

To analyze the gradient, write it as a weighted sum of residuals:

$$\text{(A8)} \qquad \hat{G}(\theta_0) \equiv \sum \hat{\tau}_i \hat{r}_i \hat{w}_i / N,$$

$$\hat{r}_i \equiv \left[y_i - \hat{P}_i\right]/\hat{\alpha}_i, \qquad \hat{\alpha}_i \equiv \hat{g}_v(v_i;\theta_0)\left[\hat{P}_i\left(1-\hat{P}_i\right)\right], \quad \hat{w}_i \equiv \hat{g}_v(v_i;\theta_0)\left[\partial \hat{P}_i/\partial\theta\right]_{\theta=\theta_0}.$$

Under this decomposition, with $\alpha_{iN} \equiv [g_v(v_i;\theta_0) + \delta_N(v_i;\theta_0)][P_i(1-P_i)]$, $r_{iN} \equiv [y_i - P_i]/\alpha_{iN}$, and $w_i \equiv \partial P_i/\partial\theta|_{\theta=\theta_0}$ denoting the true functions to which $\hat{\alpha}_i$, $\hat{r}_i$, and $\hat{w}_i$ "tend" (e.g. $|\hat{\alpha}_i - \alpha_{iN}| \to 0$), and with $v_{i0} \equiv v(x_i;\theta_0)$:

$$\text{(A9)} \qquad E\left(r_{iN}|x_i\right) = E\left(r_{iN}|v_{i0}\right) = E\left(w_i|v_{i0}\right) = 0.$$

Employing (A7)–(A8), Lemma 6 shows that all estimated functions defining the gradient in (A7) may be replaced by their limits:

LEMMA 6: *With ε'' as the likelihood trimming parameter and ε' as the probability trimming parameter, assume that $0 < \varepsilon'' < \varepsilon'$, $0 < \varepsilon' < 1$. Select the window, h_N, such that $N^{-1/(6+2\varepsilon')} < h_N < N^{-1/8}$. Then, under the assumptions of Lemmas 3–4,*

$$N^{-1/2}\sum \hat{\tau}_i \hat{r}_i \hat{w}_i - N^{-1/2}\sum \tau_i r_{iN} w_i \equiv N^{1/2}\hat{G}(\theta_0) - N^{1/2}G_N(\theta_0) = o_p(1).$$

PROOF OF LEMMA 6: With $\mu_{iN} \equiv r_{iN} w_i$ and $\hat{\mu}_i \equiv \hat{r}_i \hat{w}_i$, write $N^{1/2}\hat{G} - N^{1/2}G_N$ as

$$\underbrace{N^{-1/2}\sum \tau_i(\hat{\mu}_i - \mu_{iN})}_{\boldsymbol{A}} + \underbrace{N^{-1/2}\sum (\hat{\tau}_i - \tau_i)\mu_{iN}}_{\boldsymbol{B}}$$

$$\text{(A.10)} \qquad + \underbrace{N^{-1/2}\sum (\hat{\tau}_i - \tau_i)(\hat{\mu}_i - \mu_{iN})}_{\boldsymbol{C}}.$$

In analyzing these terms, it is important to note that τ_i essentially restricts all densities to be no smaller than $O(h_N^{\varepsilon''/5})$, $0 < \varepsilon'' < \varepsilon' < 1$. As defined in Lemma 4, we denote $\mathcal{V}_N$ as the set of $v(x_i;\theta_0)$ whose densities are at least this large. Proceeding with $\boldsymbol{A}$,

$$\boldsymbol{A} = \underbrace{N^{-1/2}\sum \tau_i(\hat{r}_i - r_{iN})w_i}_{\boldsymbol{A_1}} + \underbrace{N^{-1/2}\sum \tau_i(\hat{r}_i - r_{iN})(\hat{w}_i - w_i)}_{\boldsymbol{A_2}}$$

$$+ \underbrace{N^{-1/2}\sum \tau_i r_{iN}(\hat{w}_i - w_i)}_{\boldsymbol{A_3}}.$$

For $\boldsymbol{A_1}$, expand $\hat{r}_i$ in a Taylor series about α_{iN}:

$$\tau_i \hat{r}_i = \tau_i \hat{r}_i^* + o_p(N^{-1/2}), \qquad \hat{r}_i^* \equiv \left[\left(y_i - \hat{P}_i\right)/\alpha_{iN}\right]\left[1 - (\hat{\alpha}_i - \alpha_{iN})/\alpha_{iN}\right],$$

where the remainder term above uniformly has the stated order. Letting $\boldsymbol{A_1^*}$ be obtained from $\boldsymbol{A_1}$

by replacing $\hat{r}_i$ with $\hat{r}_i^*$, it suffices to show that A_1^* converges to zero in mean-square:

$$E\left([A_1^*]^2\right) = E\left[\sum \tau_i^2 w_i^2 (\hat{r}_i^* - r_{iN})^2/N\right] + E\left[\sum_{i\neq j} (\hat{r}_i^* - r_{iN})(\hat{r}_j^* - r_{jN})\tau_i\tau_j w_i w_j\right]\Big/N.$$

From Lemma 4 and a uniform integrability condition,[16] the first component above converges to zero. For the second component, letting $X \equiv \{x_k, k = 1, \ldots, N\}$ and $V = \{v(x_k; \theta_0), k = 1, \ldots, N\}$,

$$\begin{aligned} &E\left((\hat{r}_i^* - r_{iN})(\hat{r}_j^* - r_{jN})\tau_i\tau_j w_i w_j\right) \\ &\quad = E\left(E\left[(\hat{r}_i^* - r_{iN})(\hat{r}_j^* - r_{jN})\tau_i\tau_j | X\right] w_i w_j\right) \\ &\quad = E\left(E\left[(\hat{r}_i^* - r_{iN})(\hat{r}_j^* - r_{jN})\tau_i\tau_j | V\right] w_i w_j\right) \\ &\quad = E\left(E\left[(\hat{r}_i^* - r_{iN})(\hat{r}_j^* - r_{jN})\tau_i\tau_j | V\right] E\left[w_i w_j | V\right]\right) = 0. \end{aligned}$$

For A_2, since τ_i essentially restricts v_i to $\mathcal{V}_N$ and $N^{-1/(6+2\varepsilon')} < h_N < N^{-1/8}$, $0 < \varepsilon' < 1$, from Lemma 4

$$|A_2| \leqslant N^{1/2} \sup|\tau_i(\hat{w}_i - w_i)| \sup|\tau_i(\hat{r}_i - r_{iN})| = o_p(1).$$

Turning finally to A_3, the argument parallels that for A_1:

$$E\left[(A_3)^2\right] = \sum_i E\left[\tau_i r_{iN}^2 (\hat{w}_i - w_i)^2\right]/N + E\sum_{i\neq j} r_{iN} r_{jN} \tau_i \tau_j (\hat{w}_i - w_i)(\hat{w}_j - w_j)/N.$$

From the uniform integrability condition employed to analyze A_1, as the first term is $o_p(1)$, it also converges to zero in expectation. For the second term, recall that $\hat{w}_k$ does not depend on y_k and that w_k depends only on $X \equiv \{x_i, i = 1, \ldots, N\}$. Therefore, taking an iterated expectation in which we first condition on X, the second term above simplifies to

$$E\sum_{i\neq j} r_{iN} r_{jN} \tau_i \tau_j \hat{w}_i \hat{w}_j/N \equiv E\sum_{i\neq j} r_{iN} r_{jN} \tau_i \tau_j \left[\hat{w}_{i[J]} + \hat{w}_{ij}\right]\left[\hat{w}_{j[I]} + \hat{w}_{ji}\right]/N,^{17}$$

where $\hat{w}_{i[J]}$ is the component of $\hat{w}_i$ containing all terms that depend neither on y_i nor on y_j, and $\hat{w}_{j[i]}$ is defined analogously. Then, since neither of these two components contribute to the above expectation and since $\hat{w}_{ij}$ and $\hat{w}_{ji}$ are each $O[(Nh^2)^{-1}]$, the expectation of the cross-product terms converges to zero as required.

Turning to the term B in (A10) above, under Lemmas 2–3 and the same mean-square convergence argument used to analyze A_3, $B = o_p(1)$. For the last term in (A10), C, replace $\hat{\tau}_i$ with the representation shown in Lemma 3. Since each term in this representation contains τ_i, in every term v_i will be restricted to $\mathcal{V}_N$. Every term can then be written as a weighted sum of trimmed residuals. Accordingly, we may employ the same argument used to show that $A = o_p(1)$ to show that $C = o_p(1)$. *Q.E.D.*

From Lemma 6, the gradient is equivalent to a random variable to which a standard central limit theorem applies. Therefore, we have the following theorem.

THEOREM 4 (Asymptotic Normality): *Under the Lemmas 6–7 above, $N^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically distributed as $N(0, \Sigma)$,*

$$\Sigma \equiv E\left\{\left[\frac{\partial P}{\partial \theta}\right]\left[\frac{\partial P}{\partial \theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}^{-1}.$$

[16] From Chung (1974, Thm. 4.5.2), if $[z_N - z]^2 \xrightarrow{p} 0$ and $\sup_N E[|z_N - z|^{2+\varepsilon}] = O(1)$, then $\lim[E[z_N - z]^2] = 0$.

[17] For example, with independence over observations and $E(r_{iN}) = 0$,

$$E\left[r_{iN} r_{jN} \tau_i \tau_j \hat{w}_i w_j\right] = E\left[E(r_{iN}|X) E\left(r_{jN} \tau_i \tau_j \hat{w}_i w_j | X\right)\right] = 0.$$

PROOF OF THEOREM 4: From Lemmas 6–7,

$$N^{1/2}(\hat{\theta}-\theta_0)-E\left\{\left[\frac{\partial P}{\partial \theta}\right]\left[\frac{\partial P}{\partial \theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}^{-1}N^{-1/2}\sum \tau_i r_{iN} w_i \xrightarrow{p} 0.$$

From the definitions of r_{iN} and w_i and Serfling (1980, Corollary, p. 32),

$$N^{-1/2}\sum \tau_i r_{iN} w_i - N^{-1/2}\sum r_{iN} w_i \xrightarrow{d} 0.$$

The theorem now follows because $N^{-1/2}\Sigma r_{iN} w_i$ is asymptotically distributed as

$$N\left[0, E\left\{\left[\frac{\partial P}{\partial \theta}\right]\left[\frac{\partial P}{\partial \theta}\right]'\left[\frac{1}{P(1-P)}\right]\right\}\right]. \qquad \textit{Q.E.D.}$$

APPENDIX B: LOCAL SMOOTHING

As the arguments under local smoothing are quite similar to those of Appendix A, here we will briefly outline the argument. From (C8.b), recall that the kernel estimate of $g_{yv}(v_i;\theta)$ under local smoothing is

$$\text{(B1)} \qquad \hat{g}_{yv}(v_i;\theta,\hat{\lambda}_y;h_N)\equiv \sum_{j\neq i}^{N}\frac{1\{y_j=y\}}{h_N\hat{\lambda}_{yj}}K\left[\frac{v_i-v_j}{h_N\hat{\lambda}_{yj}}\right]\Big/(N-1).$$

Abstracting from scaling considerations (see (C8.b)),

$$\text{(B2)} \qquad \hat{\lambda}_{yj}\equiv \lambda_N\left[\hat{g}_{yv}(v_j;\theta,h_{NP})\right],$$

where λ_N is a function that depends on $\hat{g}_{yv}$, a pilot or first-stage kernel density estimate of g_{yv} with pilot window h_{NP}. The function λ_N also depends on N through a trimming parameter that controls the rate at which $\hat{\lambda}_{yj}$ can approach zero (see (C8.b)).

Before proceeding, it should be noted that bias calculations will depend on estimated pilot density derivatives. To deal with this bias, it is helpful to set wider pilot windows (h_{NP}) than second stage windows (h_N). In so doing, the proof simplifies because it becomes possible to take the estimated local smoothing parameters as given (see e.g. Klein (1991)).

In Appendix A, we obtained convergence results in Lemma 2 for estimated densities and derivatives under bias reducing kernels. Once we have obtained analogous results under locally-smoothed kernels, the proofs of consistency and normality will be very similar to those in Appendix A. We begin by providing these results for pilot densities (the local smoothing parameters).

LEMMA 7 (Pilot Estimates): *In the notation of* (B1)–(B2), *for the pilot density estimates*

$$\text{(a)} \qquad \sup_{v,\theta}\left|D_\theta^r\left[\hat{g}_{yv}[v(t;\theta);\theta,h_{NP}]\right]-E\left\{D_\theta^r\left[\hat{g}_{yv}[v(t;\theta);\theta,h_{NP}]\right]\right\}\right|=O_p\left(N^{-1/2}h_{NP}^{-(r+1)}\right),$$

$$\text{(b)} \qquad \sup_{v,\theta}\left|E\left\{D_\theta^r\left[\hat{g}_{yv}[v(t;\theta);\theta,h_{NP}]\right]\right\}-D_\theta^r\left[g_{yv}(v(t;\theta);\theta)\right]\right|=O_p\left(h_{NP}^2\right).$$

PROOF OF LEMMA 7: From Lemma 1 of Appendix A, (a) holds. The proof of (b) follows along the same lines as the proof of Lemma 2 in Appendix A. *Q.E.D.*

Employing Lemma 7, we can now obtain results analogous to Lemma 2 of Appendix A. To this end, from (B2) it is convenient to define an "average" local smoothing parameter:

$$\text{(B3)} \qquad \bar{\lambda}_{yj}\equiv \lambda_N\left[E\left(\hat{g}_{yv}(v_j;\theta,h_{NP})\right)\right].$$

From (B1), denote $\hat{g}_{yv}(v_i;\theta,\bar{\lambda}_y;h_N)$ as the kernel estimate obtained by replacing $\hat{\lambda}_{yj}$ with $\bar{\lambda}_{yj}$. From a standard Taylor series argument and Lemma 7, one can establish convergence rates

(uniform in v, θ) to zero for

$$\text{(B4)} \qquad \left| D_\theta^r\left(\hat{g}_{yv}\left(v_i;\theta,\hat{\lambda}_y;h_N\right)\right) - D_\theta^r\left(\hat{g}_{yv}\left(v_i;\theta,\bar{\lambda}_y;h_N\right)\right)\right|.$$

From Lemma 1, we can now obtain a uniform convergence rate to zero for the latter term in (B4) to its expectation that is analogous to that in Lemma 2 of Appendix A. Proceeding as in Lemma 2, from a Taylor series expansion in h_N about zero, we obtain a convergence rate for the expectation of this term to the truth. When the local smoothing parameters are known and bounded away from zero, one can show that the expected density estimate converges to the truth at a rate (see Silverman (1986)) of h_N^4. When local smoothing parameters are not bounded away from zero, uniform convergence rates are decreased by the proximity of the density to zero (the bias depends on the reciprocal of the density raised to a power). Under the trimming in (C.8b), one can show that the convergence rate exceeds $N^{-1/2}h_N^{-1}$ with estimated local smoothing parameters. This rate suffices for our purposes (see Lemma 6, the analysis of term $\boldsymbol{A}$).

One can now prove convergence results for estimated probabilities and derivatives that are analogous to Lemma 4 of Appendix A. Employing Lemma 3 of Appendix A, which characterizes the large-sample behavior of the trimming functions, τ, we can then establish consistency and asymptotic normality in essentially the same manner as in Appendix A.

REFERENCES

ABRAMSON, I. S. (1982): "On Bandwidth Variation in Kernel Estimates—A Square Root Law," *The Annals of Statistics*, 10, 1217–1223.

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

BHATTACHARYA, P. K. (1967): "Estimation of a Probability Density Function and its Derivatives," *The Indian Journal of Statistics: Series A*, 373–383.

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

CHUNG, K. L. (1974): *A Course in Probability Theory*. New York: Academic Press.

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

——— (1987): "Efficiency Bounds for Distribution-free Estimators of the Binary Choice and Censored Regression Models," *Econometrica*, 55, 559–586.

FIX, E., AND J. L. HODGES (1951): "Discriminate Analysis, Nonparametric Discrimination: Consistency Properties," Technical Report #4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.

HAN, A. K. (1984): "The Maximum Rank Correlation Estimator in Classical, Censored, and Discrete Regression Models," Technical Report No. 412, Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, CA.

——— (1988): "Large Sample Properties of the Maximum Rank Correlation Estimator in Generalized Regression Models," Working Paper, Harvard Univ.

HOEFFDING, H. (1963): "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58, 13–30.

HOROWITZ, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

ICHIMURA, H. (1986): "Estimation of Single Index Models," Unpublished manuscript.

KLEIN, R. W. (1991): "Density Estimation with Estimated Local Smoothing Parameters," Technical Memo #TM-ARH-07103, Bellcore.

LEE, LUNG-FEI (1989): "Semiparametric Maximum Profile Likelihood Estimation of Polytomous and Sequential Choice Models," University of Minnesota, Discussion Paper No. 253.

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

——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

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

NEWEY, W. K. (1989): "The Asymptotic Variance of Semiparametric Estimators," Princeton University, Econometric Research Program Memo. No. 346.

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

POWELL, J., J. STOCK, AND T. M. STOKER (1989): "Semiparametric Estimation of Weighted Average Derivatives," *Econometrica*, 57, 1403–1430.

RUUD, P. A. (1983): "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models," *Econometrica*, 51, 225–228.

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

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

SHERMAN, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137.

SILVERMAN, P. W. (1986): *Density Estimation*. New York: Chapman and Hall.

SILVERMAN, P. W., AND M. C. JONES (1989): "E. Fix and J. L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation," *International Statistical Review*, 57, 233–247.

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

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