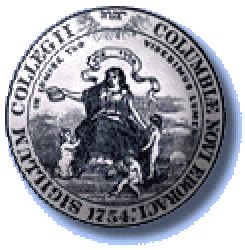

Columbia University

Department of Economics
Discussion Paper Series

MINIMUM DISTANCE ESTIMATORS FOR NONPARAMETRIC MODELS WITH GROUPED DEPENDENT VARIABLES

Mitali Das

Discussion Paper #:0102-41

Department of Economics
Columbia University
New York, NY 10027

March 2002

Columbia University
Department of Economics Discussion Paper No. 0102-41
Minimum Distance Estimators for Nonparametric Models with Grouped Dependent Variables
Mitali Das*
March 2002

Abstract:

This Version: January 2002

This paper develops minimum distance estimators for nonparametric models where the dependent variable is known only to fall in a specified group with observable thresholds, while its true value remains unobserved and possibly censored. Such data arise commonly in major U.S and U.K data sets where, e.g., the thresholds between which earnings fall are observed, but not its level. Under minor regularity conditions identification of such a model is shown to depend on there being at least two thresholds when the model disturbance's distribution is smooth and invertible. Estimators are motivated by conversion of the model into a set of binary choice models, each corresponding to one finite-valued threshold. This conversion illustrates that the difference of any two thresholds from a function that depends on identified components is identically zero; the function of interest is an additive component of this identity. Minimum distance estimators for possibly nonlinear functionals of the model are proposed, and shown to be consistent with a limiting distribution that is Gaussian. Estimators of the covariance matrix are provided. The estimators are applied to estimation of a problem in labor economics.

Keywords: Minimum Distance Estimation, Nonparametric Models, Nonparametric Estimation, Grouped Dependent Variables, Thresholds

---

* Tel/Fax: 617.924.8371, Correspondence: (e-mail) mitali.das@columbia.edu; (post) 420West 118th Street, Suite 1027a, New York, NY 10027. I gratefully acknowledge useful comments from or discussions with A. Vissing-Jørgensen, J. Angrist, P. Dhrymes, and especially W. Newey, whose comments have given added direction to this paper. Ken Chay graciously provided data for the empirical section. I also acknowledge the very warm hospitality of P. Haley and A. Ney during research on this paper. Remaining errors are mine.

1. INTRODUCTION

This paper develops estimators for nonparametric models wherein a dependent variable is known only to fall in a interval, while its true value remains unobserved. Examples of this problem arise in popular microeconometrics data sets where, e.g., earnings or weeks worked are known only to belong to a specified interval. In addition, such data are often censored below and above arbitrary points on the real line. To develop the ideas consider the model

$$y^* = m_0(x) + \varepsilon, \quad E(\varepsilon|x) = 0 \tag{1}$$

where y^* is a latent dependent variable, $x = (x_1, ..., x_K)' \in \mathcal{X} \subset \mathcal{R}^K$ is a K-vector of regressor variables, $m_0(\cdot) : \mathcal{X} \to \mathcal{R}$ is an unknown function, and ε is a stochastic disturbance with unspecified and continuous distribution function π. The latent dependent variable is known to fall into one of J intervals that exhaust the real line with jth interval given by (G_{j-1}, G_j), where J is fixed, finite and each G_j, $(j = 0, ..., J)$ is observed. Arrange the groups to be increasing such that $G_0 < G_1 < ... < G_J$, with $G_0 = -\infty$ and $G_J = \infty$. Let $1\{\cdot\}$ denote the standard indicator function. The observed data in this model are (y, x) where y is a grouped variable with

$$y = 1\{G_{j-1} \leq y^* \leq G_j\} \cdot (G_{j-1}, G_j), \quad j = 1, ..., J \tag{2}$$

where the first and Jth intervals are open ended.

Suppose ε were independent of the regressor variables, π known or specified and $m_0(x)$ dependent on x only through a linear index, e.g. $m_o(x) = x'\beta_0$, then maximum likelihood estimators of β_0 would be $n^{1/2}$ asymptotically normal as shown in Stewart (1983). More generally, suppose m_0 were a linear index, ε independent of x, but π unspecified. Redefine $y = 0$ if $y^* \in (G_0, G_1)$ and

$$y = 1\{y^* \in (G_{j-1}, G_j)\} \cdot G_{j-1}, \quad j = 2, ..., J. \tag{3}$$

Then one may rewrite (1) as the single index model $y = h(y^*) = h(x'\beta_0 + \varepsilon)$ where h is an unspecified, increasing transform. Various $n^{1/2}$ estimators that estimate β_0 up to an unknown scale exist for this index model, e.g. Stoker (1986), Han (1987), Ichimura (1993), Cavanagh and Sherman (1998). As these estimators are consistent for any increasing transform h, continue to define $y = 0$ if $y^* \in (G_0, G_1)$ and replace each G_j with j in (3). This yields an ordered response model with unobserved threshold points G_j, $j = 0, ..., J$. Assuming independence of the disturbance and regressors, Klein and Sherman (2001) develop $n^{1/2}$ consistent scaled estimators of $(\beta_0, G_0, ..., G_J)$ for this model.

This paper is concerned with estimation of a different model and is motivated by different considerations. Unlike an ordered response model the threshold points are observed and therefore, not of interest in estimation. The model is more general than index models in that (1)

allows for a nonparametric relation between the regressors and the latent y^*, but also leaves π unspecified without requiring the independence of the disturbances from the regressors. Thus even if m_0 were a linear index the estimators of Klein and Sherman (2001) are inapplicable because those estimators are defined strictly for independent disturbances. In fact, the weaker restriction on the disturbances is a substantive component of our model because (1) includes certain limited dependent variable models as special cases, where conditionally heteroskedastic errors may arise quite naturally.

It appears that to date, an estimator of (1) is unavailable in the literature. Estimation of (1) with arbitrary assignment of each y in the jth interval some value in (G_{j-1}, G_j) will in general lead to inconsistent estimation of m_0, prompting the search for an alternate estimation strategy for the model. This paper takes a step in that direction, suggesting a method derived from conversion of the model into a set of $J-1$ binary choice models for each of the finite-valued thresholds, i.e., $y^j = 1\{y^* < G_j\}$ $(j = 1, ..., J-1)$. This conversion results in a set of identities that, for arbitrary j, k $(j \neq k)$, identically equate the difference of the jth and kth thresholds to a function that depends only on the conditional means of the jth and kth binary choice models. The function m_0 is an additive component of this identity. A nonparametric two stage estimator is proposed in which the first stage constitutes estimation of the binary choice models, and the second stage consists of a nonparametric generalization of the classical minimum distance estimator that minimizes some measure of distance between the thresholds and the function of interest. Naturally, observability of the thresholds is central to second step estimation, making the estimators here inapplicable to the standard ordered response models.

Estimators of model (1) can have many uses in microeconometrics empirical research, a compelling example being models with earnings or weeks worked as a dependent variable. In major U.K and U.S data sets such variables are both grouped and censored for administrative or confidentiality reasons (e.g., earnings in various US Censuses, the Survey of Consumer Finances, and UK National Training Service data files; weeks worked in the Current Population Survey), giving useful practical motivation for the estimators of m_0 proposed in this paper. The function m_0 is not the only estimand of interest, however. Let $f_0(x)$ represent a distribution function on $\mathcal{X}$, $w(x)$ a scalar weighting function and $(x_a', x_b')'$ represent subvectors of x. Then academic and policy relevant parameters may correspond to estimands:

$$
\begin{aligned}
&\text{(a) } m_0(\bar{x}); \text{ for some } \bar{x} = (\bar{x}_1, ..., \bar{x}_K) \in \mathcal{X} \\
&\text{(b) } \int E_x[w(x)[\partial m_0(x)/\partial x_a']] \text{ for some } a \subset (1, ..., K) \\
&\text{(c) } \int_{\bar{x}_a}^{\mathcal{X}} m_0(x) df_0(x) \\
&\text{(d) } \int_{\underline{x}_a} h(m_0(x_a', \tilde{x}_b')) dx_a \text{ for fixed } \tilde{x}_b \text{ and known } h : \mathcal{R} \to \mathcal{R}
\end{aligned}
\tag{4}
$$

among others. The examples in (4) correspond to a point estimate, weighted average derivative,

unrestricted expectation, and expectation of some possibly nonlinear function of m_0, respectively. One example of (d) is average consumers' surplus in Hausman and Newey (1995) where m_0 is a demand function, x_a price level, x_b income, and $h(\cdot)$ is $\exp\{\cdot\}$. Section 4 will derive large sample distributional results for each of these estimands, including the function m_0. Rather than derive these results separately for each example, the asymptotic distribution theory will be given for a general nonlinear functional of the model that includes each of these, and others.

We begin in Section 2 by analyzing identification of the model. Section 3 describes estimators of a class of nonlinear functionals of the model, and consistent estimators of the asymptotic covariance matrix of the estimators. Some practical/implementation considerations are also discussed. Section 4 states the assumptions underlying consistency and asymptotic normality of the estimators, and gives the principal distribution results. An important aspect of this paper is feasibility of the proposed method in empirical research. To this end, Section 5 presents an empirical study that applies the suggested estimators to study the returns to education in 1964-1971 in an earnings model where earnings are both grouped and censored.

2. IDENTIFICATION

Define $J-1$ binary-valued dependent variables, corresponding to each of the finite-valued thresholds G_j $(j \neq 0, J)$ with jth dependent variable, denoted y^j, given by

$$y^j = 1\{y^* < G_j\}, \qquad (j = 1, ..., J-1). \tag{5}$$

Define $P^j(x) \equiv E(y^j|x)$. By the joint distribution of (y^j, x, ε) identified, $P^j(x)$ is identified for all j $(j = 1, ..., J-1)$, so that for any other $\bar{P}^j(x) = E(y^j|x)$, $\Pr(\bar{P}^j(x) - P^j(x) = 0) = 1$. By equation (1),

$$P^j(x) = \pi(G_j - m_0(x)). \tag{6}$$

The first theorem will illustrate that this identified component suffices to identify $m_0(x)$, up to an additive constant, under few additional conditions. Abbreviate P^k for $P^k(x)$, let $\psi_0(P^k)$ represent a function that satisfies $\int_{\mathsf{R}} \psi_0(P^k)dP^k = 1$ and denote $\mathcal{R}$ for the support of ε.

Theorem 2.1 *In the model (1) if i) π is strictly increasing everywhere on $\mathcal{R}$, ii) $J > 2$; and, iii) $\forall j \neq k$ the support of P^j conditional on P^k is at least a singleton with probability one, then $m_0(x)$ is identified up to an additive constant.*

Proof: By $P^j(x) = \pi(G_j - m_0(x))$ and condition i), $\forall j \ = 1, ..., J-1$

$$m_0(x) \equiv -\pi^{-1}(P^j(x)) + G_j. \tag{7}$$

By $J > 2$ (condition ii), for any $j \neq k$ $(j, k = 1, ..., J-1)$ define $\Delta^{jk} \equiv G_j - G_k$. By equation (7),

$$\Delta^{jk} \equiv \pi^{-1}(P^j(x)) - \pi^{-1}(P^k(x)) = g_0(P^j(x), P^k(x)). \tag{8}$$

By constancy of Δ^{jk}, for any $j \neq k$, $E\{\Delta^{jk}|P^j(x), P^k(x)\} = E\{g_0(P^j(x), P^k(x))|P^j(x), P^k(x)\} = g_0(P^j(x), P^k(x))$, giving identification of g_0 up to an additive constant. Thus, by condition iii)

$$-\int [E\{\Delta^{jk}|P^j(x), P^k(x)\}]\psi_0(P^k)dP^k = m_0(x) + \int \pi^{-1}(P^k(x))\psi_0(P^k)dP^k. \text{ ¥} \tag{9}$$

Equation (9) derives an explicit form for $m_0(x)$ in terms of the identified components $g_0(P^j(x), P^k(x))$ and $P^j(x)$. Thus, it is possible to recover the entire function m_0 by varying x over its support. Theorem 2.1 therefore gives identification of $m_0(x)$ to within an additive constant.

An important by-product of this theorem is that it implicitly analyzes nonidentification of $m_0(x)$ in a standard binary choice model with threshold zero where $J = 2$, $G_0 = -\infty$, $G_1 = 0$ and $G_2 = \infty$. In that model, $P^1(x) = \pi(m_0(x))$ so that π and m_0 are not separately identified, as is well known in the literature. Thus, Theorem 2.1 highlights the role of having at least two thresholds in identification of the model (1), with one threshold possibly zero. These features will form the basis of the proposed estimators of m_0, that are discussed next.

3. MINIMUM DISTANCE ESTIMATION

By equation (7), estimation of m_0 up to an unknown additive constant is equivalently the estimation of $-\pi^{-1}(P^j(x))$ for some $j \in [1, J-1]$. We will therefore derive an estimator of m_0 by treating it as as an additive component of $-g_0$, $g_0(\cdot, \cdot) \in \mathcal{G}$, $\mathcal{G} : \mathcal{R}^2 \to \mathcal{R}$.

Define $\tilde{J} = \binom{J-1}{2}$, let $\Delta = (\Delta^{12}, \Delta^{13}, ..., \Delta^{\tilde{J}-1,\tilde{J}})'$ denote the $\tilde{J}$ vector consisting of each unique j, k pair of threshold points, and define the corresponding $\tilde{J}$ vector $g_0(P) = (g_0(P^1, P^2), ..., g_0(P^{\tilde{J}-1}, P^{\tilde{J}}))$. By equation (8),

$$\Delta - g_0(P) = 0. \tag{10}$$

The proposed estimator will use equation (10) to estimate g_0 by minimizing some measure of distance between Δ and g_0; this estimator will be implemented in two stages. A first stage will constitute estimation of the $J-1$ conditional distribution functions $P(x) = (P^1(x), ..., P^{J-1}(x))'$ by nonparametric estimation of each binary choice model $y^j = 1\{y^* < G_j\}$ $(j = 1, ..., J-1)$. The approach in the second stage will be to take a linear-in-parameters (i.e., series) approximation to g_0, such that minimization of equation (10) over all j, k pairs will amount to a vectorized, nonparametric generalization of the classical minimum distance estimator. This is one version

of a minimum distance estimator where, for each j,k pair the distance minimized is between a vector of constants and a random vector. Additivity of g_0 will imply that marginal integration of $-g_0(\cdot,\cdot)$ over its second argument will estimate $m_0(\cdot)$ to within an additive constant; see, e.g., Linton and Nielsen (1995).

It is clear from the described estimation strategy that the suggested estimator requires the dependent variable to be categorized into a minimum of three groups (i.e., $J \geq 3$, when $G_0 = -\infty$, $G_J = \infty$) such that there are at least two finite-valued threshold points G_j, G_k $(j \neq k)$. Although not necessary to the estimation strategy, the presence of additional groups will increase the number of unique j,k pairs, and thereby increase the number of identities that can be used in "solving" equation (10). We will give the most general form of the estimators that utilize all $\tilde{J}$ unique pairings, although neither consistency nor the asymptotic distribution theory will depend on this generality, but only on the use of any j,k pair of thresholds from the available $J-1$ finite-valued thresholds. This feature of the model is analogous to the "first-difference" estimators in linear panel models with T data points per cross sectional observation; see Chamberlain (1994). In that model it is well known that the use of all possible first-difference pairs increases the number of moment restrictions and thereby raises the asymptotic efficiency of the estimated parameters, but that consistency obtains from the use of any arbitrary first-difference pair. It is plausible that an analogous efficiency result obtains for this model (e.g., when equation (1) is a linear index), although verifying that conjecture is outside the scope of this paper.

Series estimators of this model will have certain practical benefits, but will not be essential, in the estimation strategy. As in the nonparametric panel model (see, e.g., Porter (1996) and Das (2001)), a linear in parameters expansion of $\pi^{-1}(P^j(x))$ and $\pi^{-1}(P^k(x))$ in the second stage will lead naturally to the restriction of equality of parameters on both approximations, implying that a regression of Δ^{jk} on a first difference of the approximation will yield a simple estimate of g_0. With this smoother, the function m_0 can be recovered by a linear combination of the coefficients obtained from estimation of g, and thus avoid the additional step of marginal integration that could be computationally cumbersome in an application. This aspect will be different from other nonparametric methods for the second stage, e.g., kernel regression. We will also use series estimation of the first step, although the conditional distributions are consistently estimable by a number other nonparametric methods as well. As in other methods (e.g., bandwidth choice in kernel regression) the number of summands in the series approximating basis will depend on subjective or data-dependent choices. We will discuss these considerations below, subsequent to introducing the notation and estimators.

Consider the first step of estimation. Let

$$\{\chi_s(x) : s = 1, 2, ...\}$$

denote a sequence of functions with $\chi_s(\cdot) : \mathcal{X} \to \mathcal{R}$, such that $\chi^S(x)'\beta = \sum_{s=1}^{S} \chi_s(x)\beta_s$ forms

an approximation to $P(x)$ when the basis consists of S finite summands, β is a vector of unknown coefficients, and the approximation improves (in a sense to be made precise below) as $S \to \infty$. As in Andrews (1991), this paper takes S to be nonrandom, imposing certain regularity conditions on the choice of S that restrict its growth with sample size. The use of a common approximating basis for all components of $P(x)$ reflects the common distribution function estimated in each of the binary choice models. Let $i = 1, ..., n$ denote the observations and define $y = (y^1, ..., y^{J-1})$. Estimators of $P(x)$ are obtained by least squares regression of y on $\chi^S(x)$, given by

$$\hat{P}(x_i)' = \chi^S(x_i)'\hat{\beta}, \qquad \hat{\beta} = \left(\sum_{i=1}^{n} \chi^S(x_i)\chi^S(x_i)' \right)^{-1} \sum_{i=1}^{n} \chi^S(x_i)[y_i^1, ..., y_i^{J-1}]. \tag{11}$$

Turning now to estimation of m_0, suppress the dependence of P^j on x and let $\{q_b(P^j); b = 1, 2, ...\}$ represent a sequence of functions where $q_b(\cdot) : \mathcal{R} \to \mathcal{R}$. Also, let $q^B(P^j)'\gamma = \sum_{b=1}^{B} q_b(P^j)\gamma_b$ denote a linear-in-parameters approximation to $\pi^{-1}(P^j)$ for some choice of B that satisfies regularity conditions given below. Let P^{jk} abbreviate (P^j, P^k) and define $\xi^B(P^{jk}) \equiv q^B(P^j) - q^B(P^k)$. By $g_0(P^{jk}) = \pi^{-1}(P^j) - \pi^{-1}(P^k)$, the linear-in-parameters expansion for π^{-1} leads naturally to imposing both, $\xi^B(P^{jk})$ as an approximating basis for g_0 and the equality of coefficients on corresponding elements $q_b^B(P^j)$ and $q_b^B(P^k)$ $(b = 1, ..., B)$. Let this approximation to $g_0(P^{jk})$ be given by

$$\sum_{b=1}^{B} \xi_b^B(P^{jk})'\gamma_b. \tag{12}$$

By $\Delta^{jk} - g_0(P^{jk}) = 0$, equation (12) immediately suggests a natural estimator of $g_0(P^{jk})$ as one that chooses γ to minimize the Euclidean distance between Δ^{jk} and $\xi^B(P^{jk})'\gamma$. To formalize this idea, let $\omega(P^{jk}) = \mathbf{1}(0 < P^j(x) < 1)\mathbf{1}(0 < P^k(x) < 1)$ define a fixed trimming function that restricts estimation to those probabilities that lie strictly below zero and one. Let $\hat{\omega}$ abbreviate $\omega(\hat{P}^{jk})$, $\hat{\xi}^B$ abbreviate $\hat{\omega}\xi^B(\hat{P}^{jk})$, and define the residual

$$\hat{\rho}(\hat{P}_i^{jk}, \gamma; \hat{\xi}^B) = \Delta^{jk} - \left(\sum_{b=1}^{B} \hat{\xi}_b^B(\hat{P}_i^{jk})'\gamma_b \right).$$

Let $\hat{\rho}(\hat{P}_i, \gamma; \hat{\xi}^B) = (\hat{\rho}(\hat{P}_i^{12}, \gamma; \hat{\xi}^B), ..., \hat{\rho}(\hat{P}_i^{J-1,J}, \gamma; \hat{\xi}^B))'$ represent a $\tilde{J} \times 1$ residual vector consisting of all distinct j, k residuals. An estimator of $g_0(P^{jk}) = g_0(P^j(x), P^k(x))$ is obtained by solving for the vector γ that sets the sample residual vector $\rho(\hat{P}_i, \gamma; \hat{\xi}^B)$ closest, in Euclidean distance, to zero:

$$\begin{aligned} \hat{g}(P^{jk}) &= \xi^B(P^{jk})'\hat{\gamma} \\ \hat{\gamma} &= \arg\min_{\gamma} \sum_{i=1}^{n} \hat{\rho}(\hat{P}_i, \gamma; \hat{\xi}^B)\ A(x_i)\ \hat{\rho}(\hat{P}_i, \gamma; \hat{\xi}^B)' \end{aligned} \tag{13}$$

where $A(x_i)$ is a conformable positive definite weighting matrix. Thus, $\hat{\gamma}$ is analogous to a standard mininimum distance estimator of the parameters of a linear model with a fixed number of regressors, e.g., those developed in Chiang (1956).

As suggested above, an estimate $\hat{m}(x)$ can now be obtained by averaging $-\hat{g}(P^j, P^k)$ over its second argument. Because m_0 is an additive component of g_0, this partialling will yield an estimate of m_0 up to an unknown additive constant, for any nonparametric estimate $\hat{g}$. In series estimation linearity of the approximation in equation (12) implies that an estimator of m_0 may alternatively be obtained by a linear combination of $\hat{\gamma}$ and the approximating basis $q^B(P^j)$, which can be thought of as implicitly integrating $-\hat{g}$ over its second argument. This estimator of $m_0(x)$ is

$$\hat{m}(x) = -1 \cdot \left[q^B(P^j(x))'\hat{\gamma} \right]. \tag{14}$$

Further, estimators of functionals of the model such as the examples in (4) can be derived using the estimator $\hat{g}$ from equation (13). To describe these functional estimators, for some g let $\lambda(g) : \mathcal{G} \to \mathcal{R}$ denote a generic scalar estimand where $\lambda(\cdot)$ is a known and possibly nonlinear function. Consider the functional estimator $\lambda(\hat{g})$, with population value given by $\lambda(g_0)$. The function $m_0(\cdot)$ is an immediate example of $\lambda(g_0)$ with

$$m_0(x) = \lambda(g_0); \quad \lambda(g) = -\int g(P^{jk}) dP^k, \tag{15}$$

where the dependence of each P^j on x is suppressed. Each of the estimands in (4) is also a functional $\lambda(g)$ with examples (a), (b), (c) and (d) represented as

$$\begin{aligned}
&\text{(a)} \quad \lambda(g) = m(\bar{x}); \ \ m(x) \text{ as in (15)} \\
&\text{(b)} \quad \lambda(g) = -\int w(x) \left\{ \frac{\partial}{\partial x} \int g(P^{jk}) dP^k(x) \right\} dx \\
&\text{(c)} \quad \lambda(g) = -\int \int g(P^{jk}) dP^k \, dx \\
&\text{(d)} \quad \lambda(g) = \int_{\underline{x}_a}^{\bar{x}_a} h\left[m(x'_a, \tilde{x}'_b) \right] dx_a; \ \ m(x'_a, x'_b) \text{ as in (15)},
\end{aligned} \tag{16}$$

respectively.

As in Andrews (1991) and Newey (1997), the proposed estimator of $\lambda(g_0)$ considered here is $\lambda(\hat{g})$, obtained simply by substituting estimates of g in place of its population value. These are often simple to construct from a preliminary estimate of $\hat{g}$. For example, in equation (15) we can obtain an estimate $\hat{m}(\bar{x}) = n^{-1} \sum_{i=1}^{n} \hat{g}(\hat{P}^j(\bar{x}), \hat{P}^k(x_i))$, where $\hat{g} = \xi^B(\hat{P}^{jk})'\hat{\gamma}$. In example (b), the average derivative of m_0 with respect to x is by definition the average derivative of $-\pi^{-1}(\hat{P}^j(x))$ with respect to x, where $\pi^{-1}(\cdot)$ is approximated by $q^B(\cdot)'\hat{\gamma}$ and $\hat{P}^j$ is approximated

by $\chi^S(x)'\hat{\beta}$, so that an estimator of example (b) can be obtained by applying the chain rule to $q^B(\hat{P}^j)\hat{\gamma}$. For the possibly nonlinear functional in (d), an estimate is $\lambda(\hat{g}) = (\bar{n}\ )^{-1} \sum_{\substack{x_a = \bar{x}_a \\ x_a = \underline{x}_a}} h\left(\hat{m}(x_a', \tilde{x}_b')\right)$, where $\bar{n}$ is the number of observations in $[\underline{x}_a, \bar{x}_a]$.

Thus, a number of useful estimands may be obtained from a preliminary estimate of $\hat{g}$. We next derive consistent estimators of the covariance matrix of these functionals, followed by the large sample distribution theory.

3.2 Covariance Matrix Estimators

Large sample confidence intervals as required for inference will require a consistent estimator of the covariance matrix of the limiting distribution of $\lambda(\hat{g})$. Let the asymptotic covariance matrix of the scalar functional be given by

$$\Omega = \mathrm{Var}(\lambda(\hat{g})),$$

and $\Omega^{-1/2}$ denote the square root of the inverse of Ω. Using the notation $F_g(g; g_0)$ to denote a functional derivative of $\lambda(g)$ at g_0, define the Jacobian matrix

$$\Lambda = (F_g(g; q_1^B), ..., F_g(g; q_B^B)), \tag{17}$$

where $\Omega^{-1/2}$ and Λ will each exist under the regularity conditions given below. As $\hat{g}$ is derived from a linear combination of minimum distance parameters, the variance of the functionals can be obtained as a function of the vectors $\hat{\gamma}$ by applying the "delta method" to a standard parametric minimum distance estimator that depends on generated components.

The following additional notation is required. Let χ_i^S abbreviate $\chi^S(x_i)$. For $y_i = (y_i^1, ..., y_i^{J-1})$ define

$$\begin{aligned} W_\beta &= E(\mathrm{Var}(y_i|x_i) \otimes \chi_i^S \chi_i^{S\prime}), \\ V_\beta &= \{E(I_{J-1} \otimes \chi_i^S \chi_i^{S\prime})\}^{-1} W_\beta \{E(I_{J-1} \otimes \chi_i^S \chi_i^{S\prime})\}^{-1} \end{aligned}$$

where the matrix V_β is analogous to the White (1980) covariance matrix for a least squares estimator of a parametric model, with fixed S, that accommodates possibly heteroskedastic errors. By linearity of the approximation to the vector of conditional probabilities, $(I_{J-1} \otimes \chi_i^{S\prime}) V_\beta (I_{J-1} \otimes \chi_i^S)$ is the asymptotic variance of $\hat{P}_i$. When the minimum distance estimates of γ depends on an estimate of $\hat{P}(x_i)$ from the first stage, the covariance matrix of $\hat{P}(x_i)$ will be one component of the asymptotic variance of $\hat{g}$. The other component will depend on the second moment matrix of the approximating basis ξ^B, that itself depends on the vector $\hat{P}_i'$ $(i = 1, ..., n)$.

To derive this additional term let $\omega\xi^B(P_i^{jk})$ abbreviate $\omega(P_i^{jk})\xi^B(P_i^{jk})$ and define the $B \times \tilde{J}$ matrix of second stage regressors $\xi_{\tilde{J}}^B(P_i) = (\omega\xi^B(P_i^{12}), ..., \omega\xi^B(P_i^{\tilde{J}-1,\tilde{J}}))'$. Then, the asymptotic variance formula for the minimum distance estimates $\hat{\gamma}$ is given by

$$V_\gamma = \{E(\xi_{\tilde{J}}^B(P_i) A(x_i) \xi_{\tilde{J}}^B(P_i)')\}^{-1} W_\gamma \{E(\xi_{\tilde{J}}^B(P_i) A(x_i) \xi_{\tilde{J}}^B(P_i)')\}^{-1} \tag{18}$$

where

$$\begin{aligned} W_\gamma &= E(\xi_J^B(P_i)A(x_i))\{\zeta_{\gamma,\beta}\ V_\beta\ \zeta'_{\gamma,\beta}\}E(\xi_J^B(P_i)A(x_i))' \\ \zeta_{\gamma,\beta} &= -E([\partial g_0(P_i)/\partial P]' \otimes \chi^S(x_i)') \end{aligned}$$

and $A(x_i)$ is a conformable positive definite weighting matrix. Corresponding to $\hat{\gamma}$ an analogue to the classical minimum distance estimator, V_γ is an analogue to the variance formula for the parameters of a parametric minimum distance estimator (with fixed B), e.g., Chiang (1956). Note that when V_β is the asymptotic variance of $\hat{\beta}$, by constancy of Δ and iterated expectations, the matrix $\{\zeta_{\gamma,\beta}\ V_\beta\ \zeta'_{\gamma,\beta})\}$ is the asymptotic variance of $\rho(\hat{P}_i, \gamma; \xi^B)$. The form of the variance matrix in equation (18) implies that setting $A(x_i) = \{\zeta_{\gamma,\beta}\ V_\beta\ \zeta'_{\gamma,\beta}\}^{-1}$ will lead to a simplified variance formula for V_γ, reducing it, in a positive definite sense, to

$$V_\gamma = \{E(\xi_J^B(P_i)W_\gamma^{-1}\xi_J^B(P_i)')\}^{-1}. \tag{19}$$

The asymptotic covariance matrix formula of the linear functional $\lambda(\hat{g})$ follows from equations (17) and (18) as

$$\Omega = \Lambda\{V_\gamma\}\Lambda'.$$

Each of the terms in the covariance matrix of $\lambda(\hat{g})$ is easily computable as a sample average. To describe a consistent estimator of Ω let

$$\hat{\Lambda} = \partial\lambda(\xi_J^{B\prime}\gamma)/\partial\gamma|_{\gamma=\hat{\gamma}}$$

define an estimator of the Jacobian matrix, let $\sum_i$ denote the sum from $i = 1$ to n, and

$$\begin{aligned} \hat{W}_\beta &= n^{-1}\sum_i\{[y_i - \hat{P}(x_i)][y_i - \hat{P}(x_i)]'\} \otimes \chi_i^S\chi_i^S, \\ \hat{V}_\beta &= \left[n^{-1}\sum_i I_{J-1} \otimes \chi_i^S\chi_i^{S\prime}\right]^{-1}\hat{W}_\beta\left[n^{-1}\sum_i I_{J-1} \otimes \chi_i^S\chi_i^{S\prime}\right]^{-1}. \end{aligned}$$

Under our regularity conditions $\mathrm{Var}(y_i|x_i)$ will be bounded, facilitating construction of $\hat{W}_\beta$. For the remaining terms as before let $\hat{\omega}$ abbreviate $\omega(\hat{P}^{jk})$, $\hat{\xi}_J^B(\hat{P}_i) = (\hat{\omega}\xi^B(\hat{P}_i^{12}), ..., \hat{\omega}\xi^B(\hat{P}_i^{\tilde{J}-1,\tilde{J}}))'$, $\hat{A}(x_i)$ represent any consistent estimator of $A(x_i)$ and define

$$\begin{aligned} \hat{\zeta}_{\gamma,\beta} &= -\left[n^{-1}\sum_i \frac{\partial\hat{g}(\hat{P}_i)}{\partial\hat{P}} \otimes \chi^S(x_i)'\right], \\ \hat{W}_\gamma &= \left[n^{-1}\sum_i \hat{\xi}_J^B(\hat{P}_i)\ \hat{A}(x_i)\right]\{\hat{\zeta}_{\gamma,\beta}\ \hat{V}_\beta\ \hat{\zeta}_{\gamma,\beta}\}\left[n^{-1}\sum_i \hat{\xi}_J^B(\hat{P}_i)\ \hat{A}(x_i)\right]'. \end{aligned}$$

Then the asymptotic variance estimator for $n^{1/2}(\lambda(\hat{g}) - \lambda(g_o))$ is given by

$$\begin{aligned} \hat{\Omega} = \hat{\Lambda}&\left[n^{-1}\sum_i \hat{\xi}_J^B(\hat{P}_i)\hat{A}(x_i)\hat{\xi}_J^B(\hat{P}_i)'\right]^{-1}\hat{W}_\gamma \\ &\left[n^{-1}\sum_i \hat{\xi}_J^B(\hat{P}_i)\hat{A}(x_i)\hat{\xi}_J^B(\hat{P}_i)'\right]^{-1}\hat{\Lambda}'. \end{aligned} \tag{20}$$

As in equation (19), this covariance matrix simplifies when $\hat{A}(x_i)$ is set equal to$\{\hat{\zeta}_{\gamma,\beta}\ \hat{V}_\beta\ \hat{\zeta}'_{\gamma,\beta}\}^{-1}$, yielding $\hat{\Omega} = \hat{\Lambda}\{\hat{W}_\gamma^{-1}\}\hat{\Lambda}'$. In the next section, we will state the regularity conditions which $\hat{\Omega}$ is nonsingular and $n^{1/2}\hat{\Omega}^{-1/2}(\lambda(\hat{g}) - \lambda(g_0))$ has a limiting distribution that is standard normal. Prior to that, some practical considerations are discussed. They are discussed ahead of the theory to facilitate discussion in the next section.

3.3. Implementation Issues

We briefly describe some implementation issues here, as pertinent to the estimators of Section 3.2 and in anticipation of the empirical application in Section 5. Series estimators have a long history in methodology, and detailed discussions of series estimators are given in Gallant (1981), Powell (1986), Andrews (1991), Newey (1997), and others. This paper primarily considers regression splines as well as orthogonal power series in a one-to-one bounded transformation of the covariates. Other choices (e.g., trigonometric series, which require the periodicity of m_0 on [0,2Pi])[1] are difficult to motivate in most economics applications. The issues are heuristically described with the approximating basis χ^S as an example.

Approximating Bases: Let θ denote a $K-$vector of nonnegative integers, a multi-index, with norm $|\theta| = \Sigma_{k=1}^K \theta_k$ and $\{\theta(s)\}_{s=1}^\infty$ be a monotonically ordered sequence consisting of distinct multi-indices with degree $|\theta(s)|$ increasing in s, and kth element denoted by $\theta(s)_k$ $(k = 1, ..., K)$. When $\chi^S(x)$ is a power series basis, $\chi_s^S(x)$ will be obtained as the product of powers, i.e, $\chi_s^S(x) = x^{\theta(s)} \equiv \Pi_{k=1}^K x_k^{\theta(s)_k}$. In the second stage, additivity can be accommodated by including only those multi-indices in which the nonzero elements correspond to either P^j or P^k. Generally, when lower power terms are included first the inclusion of higher order terms will correspond to less smooth functions (see Andrews (1991)). As power series may be unfavorably affected by the presence of outliers or discontinuities, it is possible to reduce their sensitivity to outliers by using functions in a one-to-one bounded transform of the original data, e.g., replace x by $1/1 + e^x$ (see, e.g., Newey, 1997).

A regression spline basis is useful when functions are possibly discontinuous. For a regression spline basis, $\chi_s^S(x)$ $(s = 1, ..., S)$ is obtained by taking products of functions that have the form

$$x_k^{\theta(s)_k}, \quad \theta(s)_k \leq \bar{\theta} \quad \text{or} \quad 1(x_k \geq \mho_{km})(x_k - \mho_{km})^{\bar{\theta}},$$

where $\mho_{km}$ is the mth join point that is placed in the support of x_k such that a linear combination of these functions yields a piecewise polynomial of order $\bar{\theta}$. Relative to power series, the presence of join points will render spline estimates less sensitive to outliers and less oscillatory; see Stone (1985).

Orthogonal Series: One practical consequence of the theory (which will require the dimension of the approximating bases to grow with sample size) is the plausibly high multicollinearity in

[1] It is written "Pi" in order to differentiate from the use of π in the remainder of the paper.

the elements of the approximation; also, choosing arbitrary polynomials in the approximating bases could lead to singularity of the population second moment matrix of the bases. As series estimates are invariant to nonsingular linear transforms of the bases (see e.g., M. Powell (1981)), this concern can be addressed by replacing each element of the sequence $\{\chi_s^S(x)\}$ with the product of polynomials that have the same order as the corresponding multi-index and are orthogonal with respect to some weighting function (such as a density) in the support of x. When $|\theta(s)|$ is monotonically increasing in s the transformed basis functions will be a nonsingular linear combination of the original vector of functions, and will lead to reduced collinearity. Previous empirical applications of series estimators have found such transforms to work quite well; e.g., Das (2001).

Cross Validation: As with other smoothers that depend on some subjective choices (e.g., bandwidth choice in kernel regression), the most important practical consideration is in choosing the number of summands in the approximating bases. Since the number of summands will most likely depend on the data, and therefore differ across applications, a data-based choice of summands seems indispensable in this context. One data based method is delete-one cross validation (CV) in each stage used in varying contexts in Newey (1990), Robinson (1991), Porter (1996) and recently in Ai, Blundell and Chen (2001) among others.

For the first stage, for each S the delete-one CV computes the sum of square predicted residuals, where each prediction is calculated from all observations except that to be predicted. The CV choice of S, $\bar{S}(cv)$, is that which minimizes this criterion over S. For example,

$$cv(S) = n^{-1} \sum_{i=1}^{n} [y_i - \hat{P}_{S,i}(x_i)]'[y_i - \hat{P}_{S,i}(x_i)], \qquad \hat{P}_{S,i}(x_i)' = \chi^S(x_i)'\hat{\beta}_{-i} \tag{21}$$

where $\hat{\beta}_{-i}$ are the least squares coefficients computed using all but the ith observation, and $\bar{S}(cv) = \min[cv(1), cv(2), cv(3), ...]$. It is well known (e.g., Andrews (1991)) that at the CV choice of S the bias and standard deviation approach zero at the same rate, minimizing asymptotic mean square error as $n \to \infty$, but general results for analogous properties for the CV choice of the second-stage approximating basis are to date unavailable in the literature. The distribution theory for the estimator will not prescribe any particular choice for the smoothing parameters as long as S and B satisfy certain rate and regularity conditions, that are given next. While $\bar{S}(cv)$ should automatically satisfy the rate conditions for large n, no similar claim is made for the smoothing parameters in the second stage, although it is anticipated that a CV choice may still lead to better properties than an arbitrary choice. We now turn to the large sample theory.

4. LIMITING DISTRIBUTION THEORY

The limiting distribution of the estimators will be derived similarly, but with specific differences, as for previous series estimators (e.g., Gallant and Souza (1991), Andrews (1991) and

Newey, Powell and Vella (1999)). It will differ from the two-stage weighted estimators of Newey et al (1999) in two subsantive ways, namely the extension to nonlinear functionals, and a faster mean-square convergence rate for the two-step estimator $\hat{g}$ than their two-stage weighted estimator. The latter feature will arise from minimum distance estimation of an identity, and will in turn lead to rate conditions for the nonlinear functional estimators that are no stronger than that for standard series estimators (i.e., those that do not depend on a first stage) of nonlinear functionals.

To derive the theory, additional notation will be required. To characterize the bias of the estimators arising from the approximation with a finite dimensional basis vector we will use the supremum Sobolev norm. Let μ denote a 2-vector of nonnegative integers. For $P_i^{jk} = (P^j(x_i), P^k(x_i))$ define

$$|\mu| = \sum_{j=1}^{2} \mu_j, \quad \partial^\mu g(P_i^{jk}) = \frac{\partial^{|\mu|}}{\partial P_i^{j\mu_1}, \partial P_i^{k\mu_1}} g(P_i^{jk}).$$

Let $\mathcal{P} = \{P^{jk} | \omega(P^{jk}) = 1\}$ and for a nonnegative integer q define

$$|g|_{q,\mathcal{P}} = \max_{|\mu| \leq q} \sup_{\mathcal{P}} |\partial^\mu g(P^{jk})| \tag{22}$$

(and specify $|g|_{q,\mathcal{P}} = \infty$ if $\partial^\mu g(P^{jk})$ does not exist up to order q in $\mathcal{P}$). Also, for any matrix Γ, let $||\cdot||$ denote the matrix norm $||\Gamma|| = trace\{(\Gamma'\Gamma)\}^{1/2}$. Define $\eta_i = [y_i - \hat{P}(x_i)']'$ and let $\chi_{J-1}^S(x_i)$ abbreviate $I_{J-1} \otimes \chi_i^{S\prime}$.

Our first theorem will show that there is a mean zero random variable Z_{in} with $\{E(Z_{in})\}^2 = n^{-1}$ such that $n^{1/2}\Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0)] = n^{-1/2} \sum_{i=1}^{n} Z_{in}$. Under the assumptions below, asymptotic normality of $n^{1/2}[\lambda(\hat{g}) - \lambda(g_0)]$ will then follow from the Lindberg-Feller central limit theorem. For the theory we will require the following regularity conditions and assumptions.

Assumption 1: *The data* $\{(y_1, x_1\}, ..., (y_n, x_n)\}$ $(i = 1, ..., n)$ *are i.i.d;* $E[\varepsilon_i^4 | x_i]$ $E[||\eta_i^4|| \ |x]$ *are bounded;* $\mathrm{Var}(y_i|x_i)$ *is bounded and bounded away from zero.*

The finite fourth conditional moment assumption is irrelevant for deriving convergence rates of the estimators, but required in deriving asymptotic normality. The bounded second conditional moment assumption is consistent with the presence of conditionally heteroskedastic errors, as may be important for the binary choice models in equation (5).

For q as in equation (22) define $\kappa_q(B)$ as the supremum of the norm of derivatives of order q: $\kappa_q(B) = \sup_{|\mu|=q,\mathcal{P}} ||\partial^\mu \xi^B(P_i^{jk})||$. Let $|g|_q$ abbreviate $|g|_{q,\mathcal{P}}$. Recall $F_g(g; g_0)$ denotes a functional derivative of $\lambda(g)$ at g_0 and let $\bar{\mathcal{G}}$ represent all bounded subsets of $\mathcal{G}$.

Assumption 2: *For all* $\tilde{g}$, $\bar{g} \in \mathcal{G}$, $F_g(g; \tilde{g})$ *exists and is linear in* g *such that i)* $\lambda(g)$ *is* $\bar{\mathcal{G}}-$*differentiable in* g *with respect to the norm* $|g|_q$*; ii) for some* $\varsigma > 0$ *and* q *from (22),* $|\tilde{g} - g_0|_q < \varsigma$ *and* $|\bar{g} - g_0|_q < \varsigma$ *implies* $||F_g(g; \tilde{g}) - F_g(g; \bar{g})|| \leq C|g|_q|\tilde{g} - \bar{g}|_q$*; iii)* $\kappa_q(B)^4 S^2/n \to 0$.

Assumption 2, which imposes certain smoothness and regularity conditions on the functionals, is equivalent to requiring Fréchet differentiability of $\lambda(g)$ in g with respect to the $|\cdot|_q$ norm, when $\bar{\mathcal{G}}$ represents all bounded subsets of $\mathcal{G}$. That is, for some constant c and any $\tilde{g}$ in the neighborhood of g_0, part i) will hold if $||\lambda(g) - \lambda(\tilde{g}) - F_g(g-\tilde{g};\tilde{g})|| \leq c(|g-\tilde{g}|_q)^2$. Either this condition or part i) will ensure that for all $\tilde{g}$ sufficiently close to g_0, $\lambda(g)$ can be approximated by a linear functional. The rate conditions of part iii) is required to show convergence of the Jacobian terms in the $||\cdot||$ norm.

Assumption 3: i) $\lambda(g)$ is scalar; ii) if $\lambda(g)$ is linear in g then $|\lambda(g)| < |g|_q$, otherwise $|F_g(g;g_0)| < |g|_q$; iii) there exists γ_B and a sequence of continuous functions $\{g_B = \xi^B(P^{jk})'\gamma_B\}_{B=1}^{\infty}$ such that as $B \to \infty$, $\lambda(g_B)$ is bounded away from zero but $E[g_B(P_i^{jk})^2] \to 0$.

Assumption 3, requiring $F_g(g;g_0)$ to be continuous in the supremum norm but not mean-square continuous, is a useful regularity condition for bounding the square root of Ω away from zero and also required in other unrelated parts of the proofs. Also, by Assumption 3ii) the functional derivatives are uniformly bounded, a condition that is often simply verified. For instance, for the estimands in (4), (a), (b), (c) and (d) are satisfied with q=0, q=1, $q = 0$ and q=0, respectively.

Assumption 4: i) $g_0(P^{jk})$ is Lipschitz and continuously differentiable of order v on $\mathcal{P}$; $P(x)$ is continuously differentiable of order v_1 on $\mathcal{X}$; ii) P^{jk} is continuously distributed with density bounded away from zero on $\mathcal{P}$; $\mathcal{X}$ is a Cartesian product of compact, connected intervals on which x is continuously distributed with density bounded away from zero.

The first part of Assumption 4 is a primitive condition on smoothness of the functions that will determine the bias of the estimators, e.g., v is a smoothness parameter that will specify the rate of approximation to g_0 by ξ^B. When g_0 is sufficiently smooth, then by Theorem 8 in Lorentz (1986) Assumption 4 implies $\exists\ \bar{\gamma}_B = (\bar{\gamma}_{1B}, ..., \bar{\gamma}_{BB})' \in \mathcal{R}^B$ with

$$\left\| g_0 - \sum_{b=1}^{B} \xi_b^B(\cdot)'\bar{\gamma}_{bB} \right\|_0 \to O(1/B^{v/2}),\ B \to \infty. \tag{23}$$

This implies that the slower the rate at which B increases, the greater the dimension-normalized smoothness of g_0 will have to be to satisfy equation (23) (when the approximating basis is ordered so that higher order terms correspond to less smooth functions). Consequently, the slower the growth of B, the faster will be the rate of decay of the series coefficients $\{\bar{\gamma}_{bB}\}$ in the series expansion of g_0. Analogously, the approximation rate of P in the first stage will be $O(S^{-v_1/K})$, and will improve to $O(S^{-v_1/\tilde{K}})$ if $m_0(x)$ is partially linear, where $\tilde{K}$ is the dimension of the elements of x in the nonparametric component of m_0.

Part ii) of the Assumption 4 is analogous to the full rank identification condition for consistent estimation of parameters in linear regression models. The density of P^{jk} being bounded away from zero will imply there is no degeneracy in the joint distribution of P^{jk} on $\mathcal{P}$, and therefore no

functional relationship between P^j and P^k on $\mathcal{P}$. When the distribution function π is invertible, this will suffice for marginal integration of $-g_0$ over P^k to identify m_0 to within an additive constant. Similarly, the use of orthonormal polynomials for $\{\chi_s(x)\}_{s=1}^S$ in the first step when $\{x\}$ has compact support and density bounded away from zero will be a sufficient condition for the population second moment matrix $E[\chi^S(x)\chi^S(x)']$ to be bounded away from singularity.

Assumption 5: *i) Each of* $B^{-v/2}$ *and* $S^{-v_1/K}$ *are* $o(n^{-1/2})$; *ii) for a power series basis* $(B^9 + S^9) = o(n)$; *iii) for a regression spline basis* $(B^6 + S^6) = o(n)$.

The main rate conditions are given in Assumption 5 which restrict the growth of the dimension of the approximating bases. Suppose S and B grow at the same rate, then Assumption 5 requires that these approximating sequences grow no faster than $n^{1/9}$ when the bases consists of power series, and $n^{1/6}$ when the basis consists of splines. This assumption also imposes conditions on the magnitude of the dimension-normalized smoothness of the functions requiring, for example, that $v_1/K > 9/2$ when χ^S consists of power series and $v_1/K > 3$ for regression splines.

These rates give wide bounds on the choice of B and S that might satisfy the assumption. Let $\varsigma = v_1/K$. Although a complete discussion is outside the scope of this work, it can be heuristically argued that when $P(x)$ is approximated at the rate $S^{-\varsigma}$, the cross validation choice $\bar{S}(cv)$ of equation (21) will satisfy Assumption 5 while giving the best convergence rate for $\hat{P}$. Suppose χ^S is a power series approximation and define $S^*(n)$ as the supremum of the set of S that satisfies the rate condition for given n. By choosing S proportional to $n^{1/(1+2\varsigma)}$, $\bar{S}(cv)$ approximately minimizes mean square error for large n, as alluded to in Section 3.2.2. When $v_1/K > 9/2$ (as required by the conditions of the assumption), then $\bar{S}(cv)$ will satisfy the rate conditions of Assumption 5 because $1/9 > 1/(1+2v_1/K)$. We note that in a practical application n might need to be fairly large for $\bar{S}(cv)$ to behave approximately like the optimal S.

With these assumptions we state the first distributional result.

Theorem 4.1

If Assumptions 1-5 are satisfied and $Z_{in} = n^{-1/2}(\Omega^{-1/2}\Lambda[\xi_j^B(P_i)A(x_i)\{\zeta_{\gamma,\beta}\chi_{J-1}^S(x_i)'\eta_i\}])$, *then*

$$n^{1/2}\ \Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0)] = \sum_{i=1}^{n} Z_{in} \xrightarrow{d} N(0,1), \quad n \to \infty.$$

Remark For linear functionals, $\lambda(g)$ is linear in g by definition, so conditions i) and ii) of Assumption 2 will be satisfied trivially; i.e., Assumption 2 is a relevant regularity condition only for the distribution theory for nonlinear functionals, and an inspection of the proofs will show that the conclusion of Theorem 4.1 will hold for all linear functionals even when the rate requirements of Assumption 2 (iii) do not hold and only Assumptions 1,3-5 are satisfied.

To facilitate asymptotic inference with an estimator of the covariance matrix, an additional condition is required.

Assumption 6: *i) If a spline basis is used,* $\xi^B(P^{jk})$ *is of order* $\bar{\theta} \geq 2$*, and* $B^{(q-s)} = o(n^{-1/2})$; *ii) for a power series basis* $(B^{10} + S^{10}) = o(n)$ and *for a spline basis* $(B^6 + S^6) = o(n)$.

While (23) gives uniform rates when $q = 0$, general results for arbitrary $q > 0$ are available in the literature for only a few specific cases, and consistent estimation of Ω will require rates for $q = 1$. If $g_0(x)$ is analytic then it is known that $|\xi^{B\prime}\gamma_B - g_0|_{1,\mathcal{P}} = O(1/B^\mu)\ \forall q,\ \forall \mu > 0$ when $\xi^B(\cdot)$ is a power series basis; when P^j and P^k are each univariate (as in this model), $|\xi^{B\prime}\gamma_B - g_0|_{1,\mathcal{P}} = O(B^{(q-v)})$ (if ξ^B is either a spline or a power series basis) giving Assumption 6(i). The slower growth conditions of B and S in ii) relative to Assumption 5 are required because additional conditions have to be satisfied in estimation of $\hat{\zeta}_{\beta,\gamma}$.

Theorem 4.2

If Assumptions 1-6 are satisfied then

$$n^{1/2}\ \hat{\Omega}^{-1/2}[\lambda(\hat{g}) - \lambda(g_0)] \xrightarrow{d} N(0,1), \qquad n \to \infty.$$

Remarks

1. When the functional is linear, the rate restrictions of Assumption 2(iii) are irrelevant, so that the conclusions of Theorems 4.1 (respectively, 4.2) will depend only on satisfying the rate conditions in Assumptions 5 (respectively, 6(ii)). For nonlinear functionals, whether satisfying Assumption 2(iii) will imply that the rate restrictions of Assumptions 5 or 6(ii) are satisfied, however, will depend on the derivative order. For $q > 0$ (as required for Theorem 4.2), the rate conditions of Assumption 2(iii) will dominate, while for $q = 0$ (in which case the conclusion of Theorem 4.1, but not 4.2, can hold), the rate conditions of Assumption 5 will dominate.

2. For $n^{1/2}$ asymptotic normality of $\lambda(\hat{g})$ the functional derivative $F_g(g; g_0)$ (or $\lambda(g)$ in the linear case) must be continuous in $(E[g(x)^2])^{1/2}$, which in general rules out simultaneous satisfaction of Assumption 3(iii). Therefore, for $n^{1/2}$ asymptotic normality different conditions must hold, to which we turn next.

Let $\bar{\omega} = E[\omega(P^{jk})]$ and $||\cdot||^\omega = \{E[\omega(P^{jk})(\cdot)/\bar{\omega}]\}^{1/2}$. Let ϱ denote the set of functions additive in $P^j(x)$ and $P^k(x)$ $(\forall j = 1, ..., J-1;\ k \neq j)$ that are approximable by $\xi^B(\cdot)$ in the $||\cdot||^\omega$ norm as $B \to \infty$. Define $\mathcal{S}$ as the mean-square limit of $\chi^S(x)'\beta$ over all possible χ^S as $S \to \infty$. Let $\tau(P^{jk})$ denote a function in ϱ that is mean square integrable and $d(x)$ denote the matrix of projections of $\{\omega(P^{jk})\tau(P^{jk})[\partial g_0(P^{jk})/\partial P^{jk}]\}$ on $\mathcal{S}$. Suppose that $\alpha(S) \geq \sup_{\mathcal{X}} ||\chi^S(x_i)||$. Note that, while satisfying Assumption 4(ii), χ^Scan be replaced by the nonsingular linear transform $E(\chi_i^S\chi_i^S)^{-1/2}\chi_i^S$ such that the second moment matrix of the transformed basis is I_S, satisfying

$\tilde{\alpha}(S) \geq c\alpha(S) = \sup_{\mathcal{X}} ||E(\chi_i^S \chi_i^S)^{-1/2} \chi^S(x_i)||$ for some constant $c > 0$. Let $\bar{\Omega}$ denote the variance in the $n^{1/2}$ case. Then,

$$\bar{\Omega} = E[d(x_i)\mathrm{Var}(y_i|x_i)d(x_i)']$$

Asymptotic normality in the $n^{1/2}$ case will require the following assumption.

Assumption 7: $\exists\ \tau(P^{jk}) \in \varrho$ *and* $\bar{\gamma}_B$ *such that for i)* $E(\omega(P^{jk})[\tau(P^{jk})\tau(P^{jk})']) < \infty$*; ii) for all* b, B*,* $F_g(g; \xi_b^B) = E(\omega(P^{jk})\tau(P^{jk})\xi_b^B(P^{jk}))$*; iii) either* $\lambda(g)$ *is linear in* g *and* $\lambda(g_0) = E(\omega(P^{jk})\tau(P^{jk})g_0(P^{jk}))$ *or* $F_g(g; g_0) = E(\omega(P^{jk})\tau(P^{jk})g_0(P^{jk}))$*; iv)* $E[\omega(P^{jk})||\tau(P^{jk}) - \xi^B(P^{jk})'\bar{\gamma}_B||^2] \to 0$ *as* $B \to \infty$.

Theorem 4.2 ($n^{1/2}$ consistency)**:**

If Assumptions $1, 2, 4-7$ *are satisfied then* $\lambda(\hat{g})$ *is asymptotically linear with influence function* $\vartheta(x_i) = \Lambda[\xi_J^B(P_i)A(x_i)\{\zeta_{\beta,\gamma}\chi_{J-1}^S(x_i)'\eta_i\}]$ *and*

$$n^{1/2}[\lambda(\hat{g}) - \lambda(g_0)] = n^{-1/2} \sum_{i=1}^{n} \vartheta(x_i) \xrightarrow{d} N(0, \bar{\Omega}), \qquad \hat{\Omega} \xrightarrow{p} \bar{\Omega}.$$

One example in which this theorem is particularly useful is when model (1) is partially linear e.g., $m_0(x) = m_{10}(x_a) + x_b'\varkappa_b$, where Theorem 4.2 will give the distribution theory for the coefficient vector $\varkappa_b$. To see that $\varkappa_b$ is a $n^{1/2}$-consistent functional, let $f(P^j, P^k)$ represent a joint density function that is bounded away from zero on $\mathcal{A} = \{P^{jk} : \omega(P^{jk}(x)) = 1\}$, and let $f_j(P^j)$ and $f_k(P^k)$ denote the marginal density functions. Also, suppressing the dependence of $P^{jk}(x)$ on x define $u(x) = x_b - E(x_b|x_a)$, $q(x) = \{E(\omega(P^{jk})u(x)u(x)')\}^{-1}u(x)$ and $\tau(P^{jk}) = -q(x)\{f_j(P^k)\ f_k(P^j)\}f(P^j, P^k)^{-1}$. Then, if $q(x)$ is nonsingular, by $m_0(x) = -\pi^{-1}(P^j(x)) + G_j$ and $\pi^{-1}(P^j(x)) + \bar{c} = \int g(P^{jk})f_k(P^k)dP^k$ for some constant $\bar{c}$,

$$\begin{aligned}
\lambda(g_0) &= E[\tau(P^{jk})g_0(P^{jk})] = \int\int \tau(P^{jk})g_0(P^{jk})f(P^j, P^k)dP^j dP^k \\
&= -\int q(x) \int g_0(P^{jk})f_k(P^k)dP^k \ f_j(P^j)dP^j \\
&= -E[q(x)\{\pi^{-1}(P^j(x)) + \bar{c}\}] \\
&= \{E(\omega(P^{jk})u(x)u(x)')\}^{-1}\{E(u(x)m_{10}(x_a)) + E(u(x)x_b')\varkappa_b = \varkappa_b.
\end{aligned}$$

This partially linear model can be estimated as suggested in Section 3, where ξ^B depends on functions of x_a but is linear in elements of x_b. Such partially linear models are especially useful for dimension reduction when x consists of many distinct elements making precise estimation of m_0 difficult due to the well known curse of dimensionality. The empirical application will be one example of such a partially linear model, where we allow binary-valued covariates to enter linearly in the model.

5. Empirical Application

To illustrate the use of the estimators and the practicability of the estimators of the covariance matrix, an empirical study on the estimation of an earnings model was implemented. This study is closely motivated by Chay (1995) and Chay and Powell (2001) who have used censored earnings data in 1964-1971 to examine the change in the average returns to education, and other topics.

We replicate this study using the same variables and years, when the earnings data are not only censored but grouped as well. In doing so we also extend these studies' linear index specification to a more general one, allowing for a nonparametric relationship between education, age and earnings. Permitting a more flexible relation may be useful if the labor market values graduation from one level of schooling such as high school or college more than the preceding years, such that marginal returns of an additional year of schooling vary with the level of schooling (e.g., Schultz, 1997). Furthermore, if experience is valued by employers then marginal returns to education should be increasing in age (a commonly used proxy for experience). Recently, Card (2001) has suggested that if a wage model assumes additive separability of education and a measure of experience, the returns to education will be understated at higher levels of education because the marginal return to education is plausibly increasing in work experience. In addition, Chay and Powell (2001) find strong evidence that in these data the error distributions are characterized by fatter tails than a normal distribution's.

Motivated by each of these considerations, we consider the following partially linear model:

$$y^* = m_0(x) + \varepsilon = m_{10}(x_a) + x_b'\varkappa_b + \varepsilon$$

where y^* is the natural logarithm of annual taxable earnings, $x = (x_a', x_b')'$, $x_a = (x_1, x_2)$ consists of age and education, x_b includes binary indicators for race and married, $m_{10}(\cdot)$ is an unknown function and $\varkappa_b$ are unknown coefficients. This is a partially linear version of our model, where allowing the categorical variables to enter linearly is similar to other applications; e.g. Hausman and Newey (1995).

The data, collected jointly by the U.S Census Bureau and the Social Security Administration, are for a random sample of men living in the southern states. We use data corresponding to the years 1964 and 1971. For each of these years we implement the proposed estimator on the data where the level of annual taxable earnings is grouped into categories: < 1000, $[1000, 5000)$, $[5000, 10000)$, $[10000, 15000]$, > 15000 (in 1984 dollars) with $J = 5$.[2] The data consist of 2863 and 2932 observations for 1964 and 1971 respectively.

We will compare the results obtained from using the proposed minimum distance (MD) estimator against i) ordinary least squares, ii) the maximum likelihood (ML) estimator proposed

[2]Note that y^* is defined as the natural logarithm of annual taxable earnings while the grouped variable is the level of annual taxable earnings. Therefore we use the natural logarithm of the given thresholds in implementing the estimator, e.g., $G_0 = -\infty$, $G_1 = \ln(1000)$, etc.

in Stewart (1983) for grouped dependent variable models assuming normality of errors, and iii) a nonparametric estimator that arbitrarily assigns the dependent variable some value in the jth interval for all men whose income falls in that interval. We will also compare these results to Chay and Powell (1995)'s study with continuous dependent variable data and "top-coded" censoring. We take the nonparametric MD estimators as the benchmark as they are consistent when the dependent variable is grouped and censored, and the model is nonlinear or linear, for arbitrary continuous distribution function π.

Our first step is conversion of the model into the set of binary choice models as described in Section 3, i.e.,

$$y^j = 1(m_{10}(x_a) + x_b'\varkappa + \varepsilon < G_j),\ (j = 1, ..., 4) \tag{24}$$

with G_1, G_2, G_3 and G_4 respectively ln(1000), ln(5000), ln(10000) and ln(15000). Let x_1 represent education. Our principal goal is to employ the proposed method in estimating the returns to education in 1964 and 1971 via the average derivative $E[\partial m_{10}(x_a)/\partial x_1]$, accounting for the grouped nature of the dependent variable, as discussed subsequent to equation (16). We begin by choosing the series approximations for the two stages.

Cross Validation Selection of Summands: For each stage the series approximations are chosen by minimizing the delete-one cross-validation (CV) criterion, described in Section 3.2.2. For example, in first stage estimation of $P^j(x)$ we specify several polynomial spline approximations in x_a with evenly spaced joint points in the sampling support of x_a, and use the CV criterion to determine which approximation to select. The series selected by CV is "overfitted" by adding an additional term to that determined by minimizing the CV criterion. This is because although the CV criterion leads to a mean-squared error minimizing set of approximating functions, the theory requires overfitting in order to produce a bias that is smaller than the variance asymptotically, as implied by Assumption 5; see Das, Newey and Vella (2001) for a similar use and discussion of this CV algorithm. Therefore, upon determining the number of terms that minimizes the CV criterion an additional join point is added to the specification.

Let x_1 and x_2 respectively represent education and age. Table 1 reports the CV values for a subset of the specifications estimated for the first and second stage approximations, for both years. In the top panel of Table 1 we find that in the 1964 data the CV minimizing specificationfor the first stage estimation of $P(x)$ is a fourth order term that depends on a cubic in x_1, an interaction of the quadratics in x_1 and x_2 plus two join points that are placed equidistant in the sampling support of x_1 and x_2. For 1971, the CV minimizing specification is that of 1964 plus an additional interaction between a cubic in x_1 and a quadratic in x_2. These findings reject linearity in the distribution function of $\varepsilon|x$, and support the presence of a joint effect of education and age in this conditional distribution function. As per the above discussion, an additional knot is added to each CV-minimizing specification prior to our next step. We overfit by placing the additional knot at x_1=12 which corresponds to the completion

of high school. Using these CV results we obtain estimates of $\hat{\beta}$, and thus $\hat{P}$, by least squares regression of $y = (y^1, ..., y^4)$ on the overfitted CV specification; these estimates are given in Table 2.

In the second stage we specify a regression spline approximation to $\pi^{-1}(\hat{P}^j(x))$, imposing the same choice for $\pi^{-1}(\hat{P}^k(x))$ $\forall j \neq k$. We take differences of each element in the series as discussed prior to equation (12), and cross validate the differenced series to find the CV choice for g. These results are given in the middle panel of Table 1. Note that we use only a subset of the six unique (j,k) pairs in estimation, namely $\{(j,k)\} = \{(1,2),(1,3),(3,4)\}$; as discussed in Section 3, the proposed estimators are not dependent on the use of every unique pair, while giving some computational benefit with the use of a smaller subset. In accordance with the theory (in particular, the assumptions underlying Theorem 4.2), we use quadratic splines, imposing them at equidistant points in the trimmed support of $\hat{P}$.

We find that the CV minimizing specification for both years of data depend on the difference of a quartic in $\hat{P}^j$ and $\hat{P}^k$, with one join point in the support of each regressor. As with the first stage another term is added to that minimizing the CV criterion; we do so by placing an additional knot in the support of $\hat{P}^j$. Noting that the CV statistics for $B = 3$ and $B = 4$ (in 1964) are fairly indistinguishable we will consider both of the corresponding specifications in computing the average derivative estimates. Using these results we obtain estimates of $\hat{\gamma}$ by minimizing the sum of the squared residual vector $\hat{\rho}(\hat{P}_i, \gamma; \hat{\xi}^B)$ as given in equation (13), with $B = 3$ and the weighting matrix to the identity matrix. These estimates are given in Table 2. Recall that our fixed trimming function $\omega(\hat{P}^j, \hat{P}^k)$ excluded estimated probabilities larger (smaller) than one (zero). In our application, this trimming resulted in excluding approximately 4 percent of all observations, mostly those that were negative.

Finally, as a comparison we wish to consider the estimates that would be obtained from an ad hoc procedure, by assigning every individual in an income group an arbitrary number from the income interval they are classified in. For this, we use the cross validation criterion to discriminate between specifications in estimation of model (24), when y^* is replaced with the arithmetic midpoint of the interval it falls in if $j = 1, 2, 3, 4$, and set at the natural logarithm of 500 and 15500 for G_0 and G_J respectively. CV results for this exercise are given in the bottom panel of Table 1, where we denote the number of summands as L. We find that the CV minimizing specification is a fourth order polynomial that includes a cubic in x_1, the interaction of the quadratics in x_1 and x_2, and two join points in the 1964 data. For the 1971 data, the CV minimizing specification is a fifth order polynomial, with an additional join point. Using these CV statistics, we now turn to computation of the estimands of interest.

Comparison of Estimates: We implement various estimators and report the relevant estimates in Table 2. Column (1) gives OLS estimates that ignore both nonlinearities and the grouped nature of the dependent variable, using the specification in Chay and Powell (2001): $m_0(x) = x_1\varkappa_1 + x_2\varkappa_2 + x_2^2\varkappa_3 + x_b'\varkappa_b$, and the dependent variable equal to the ad hoc assignment;

column (2) contains the ML estimates of this linear model assuming normally distributed errors. In each of these, the average derivative estimate is constant at all levels of education while for the nonparametric estimators we evaluate the average derivative in various intervals holding age (x_2) and each function of age fixed at their sample means (the sample mean of age is 39.23). Column (3) gives nonparametric estimates that permits nonlinearities with the ad hoc assignment for the dependent variable (corresponding to the bottom panel of Table 1b), and column (4) gives the nonparametric MD estimates.[3,4]

The average derivative estimates of the returns to education in Panel (B) indicate that both the nonlinearities and grouped nature of the dependent variable are important in estimation. Comparing columns (1)-(2) with column (4), the average returns to education are quite different for different ranges; and comparing columns (3) and (4), within any range the estimates are quite different when the grouped nature of the dependent variable is disregarded.

Consider the MD average derivative estimate for workers with high school or less education ($x_1 \leq 12$), which is computed by averaging over all values of x_1 less than or equal to 12, with equal weights at each x_1, and age held at its sample mean, 39.23. We find that in this interval, in 1964, the average derivative is estimated to quite precisely at .0517 with a standard error of .0142. The reported standard errors account correctly for the variability of the estimated $\hat{P}$, using the asymptotic variance formula from Theorem 4.2 and equation (20). This estimate is larger than the corresponding estimate for OLS and lower than that of ML, and does not lie within either estimate's 99 percent confidence intervals. Qualitatively similar results are observed for 1971 as well.

A first concern is whether these results are sensitive to the particular choice of series used. To address this concern, we consider altering the specifications used in generating the average derivative. First, we exclude all join points in the estimation of $\hat{g}$ and find that this alteration does not change the CV minimizing specification for g, but results in less precisely estimated coefficients, and consequently less precisely estimated average derivatives; e.g., for the interval $x_1 \leq 12$ this specification change leads to an average derivative of .0492 with a standard error of .0153. The addition of a higher order term to the CV minimizing specification of g from Table 1 also does not appreciably change the obtained estimates, e.g., including the difference of quartics in $\hat{P}^j$ and $\hat{P}^k$ ($B = 4$) results in changing the average derivative by less than .0013 in absolute value, although the resulting estimate is imprecisely estimated and statistically insignificant at

[3]For brevity, the first stage estimates are reported for only one of the four binary choice regressions in Table 1, namely parameter estimates from the y^2 regression (with G_j = ln(5000)) in both of the years. There are some differences in the parameter estimates across the 4 regressions. For example, the coefficient on x_3 changes sign from positive to negative going from y^1 to the y^4 regressions.

[4]Notice that the coefficient on x_3 in column (4) is $\partial Ey^2/\partial x_3$, which is not the Black-White wage gap. That wage gap is given by $\partial m_\theta/\partial x_3$ which, as discussed subsequent to equation (16), can be estimated as the partial derivative of $q^B(\hat{P}^j(x))'\hat{\gamma}$ with respect to x_3 by applying the chain rule since q^B is a known function of $\hat{P}^j$.

the .1 error level.

Turning next to the interval $x_1 \in [13, 16]$ (completion of high school and at most graduation of college), we find the same pattern where the MD estimates are larger than least squares, but smaller then the ML estimates. In this interval, the MD average derivatives are found to lie within 90 percent confidence intervals of the ML estimates in each year. The MD estimates suggest that the returns to education in this interval were actually higher in 1964 than in 1971; and, averaging over the two years, that the wage premium for some college education was approximately .0152 log points relative to just high school education. The MD point estimates in this interval is itself comparable in sign and magnitude with some estimates in the returns to education literature, and in particular to Chay and Powell (2001)'s censored least absolute deviations and symmetrically censored least square estimates using these data. To consider how these estimates might change for another series approximation, we exclude the join points in estimation of $\hat{P}$ (i.e., in first step estimation) and find that the average derivative decreases slightly to .0697 with a standard error of .0245. This is a reasonable finding as the excluded join point was placed at $x_1 = 12$, presumably reflecting the positive earnings premium obtained by high school graduates. The MD estimates in the interval $x_1 \geq 17$ are larger in magnitude than both the LS and ML estimates, but statistically insignificant at conventional error levels.

In comparing the two nonparametric estimators in Panel (B), we find that the average derivative estimates are attenuated in column (3) relative to column (4), although the degree of attenuation varies across the intervals. Relative to the MD estimates, they are also more precisely estimated except for the $(x_1 \geq 17)$ interval in 1964. This is not unexpected, however, as the variance estimator in column (4) must account for the variability of $\hat{P}$ from the prior step. In all but the $(x_1 \geq 17)$ interval, the column (3) estimates are also smaller in magnitude than the OLS estimates in column (1). In fact, although both OLS and the nonparametric estimator of column (3) are inconsistent in the presence of a grouped dependent variable, the linear OLS estimate is in general closer to the MD estimates, suggesting that taking nonlinear functions of the data while ignoring the grouped nature of the dependent variable leads to magnifying the bias that emerges from simple linear methods that ignore the grouping problem. However, this feature may not arise generally, and simply be peculiar to this empirical application.

We also calculate the average derivative over the entire range of values that x_1 takes in the data and find that for the nonparametric MD estimator, this estimate is .0677 in 1964 and .0635 in 1971 with estimated standard errors of .0376 and .0332 respectively. These estimates are larger than the corresponding estimates for least squares and nonparametric estimation with ad hoc assignment to the dependent variable, but smaller in magnitude than ML. These average MD estimates are quite smaller in magnitude than the leading examples surveyed in Card (2001), but may not be directly comparable because of the earlier time period being studied. These estimates are somewhat higher than the average returns to education found in Chay and Powell (2001) for their respective years. However, using our average derivative estimates over the entire

range of x_1, the implied change in the average returns to education between 1964-1971 is .0041 log points, which is somewhat comparable to the .003 log points estimated change in Chay and Powell (2001).

In summary, we find that the suggested estimator is practicable in an empirical application where the dependent variable is both grouped and censored. The difference in the estimates between the nonparametric estimator with ad hoc assignment and the MD estimates indicate that accounting for the grouping of the dependent variable is important for inference. Furthermore, the difference between these and the ML estimate is indicative that accounting for grouping alone is inadequate, because of biases that can arise jointly from the possible non-normality of errors and the nonlinear relation between earnings, education and age.[5] We find that our estimands are fairly precisely estimated, and robust to small changes in specification, indicating that the suggested estimators are practicable for similar empirical applications.

CONCLUSION

This paper considers estimation of a nonparametric model in which the dependent variable is latent and possibly censored, but grouped into categories with observable thresholds. Standard nonparametric regression techniques are inapplicable due to the latency of the dependent variable. The disturbances are assumed to be drawn from an invertible distribution and satisfy a conditional mean zero restriction, but even under stronger restrictions on the disturbances such as independence from the covariates, existing semiparametric estimators are inapplicable for estimation of this model.

Estimators suggested in this paper are derived by conversion of the model into a series of binary choice regression models, corresponding to each of the finite-valued thresholds. This conversion yields a series of identities, in which the function of an interest is an additive component. A nonparametric two stage estimator is proposed in which the first stage constitutes estimation of a component of the identity, and the second stage consists of a nonparametric generalization of the classical minimum distance estimator. Nonlinear scalar functionals of the model are shown to be asymptotically normal (and $n^{1/2}$ asymptotically normal under other regularity conditions) and consistent estimators of the covariance matrix are provided. An empirical application, which examines the change in the average returns to earnings illustrates that the estimators perform quite well, and are practicable for inference.

[5] Although we cannot infer that the difference between column (2) and (4) is entirely due to the non-normality of the errors since column (4) also considers a nonparametaric function $m_{10}(\cdot)$, Chay and Powell (2001) have formally rejected non-normality of the errors in these data.

A Appendix

Let $\bar{A}(x_i)$ denote a symmetric square root of $A(x_i)$, $\bar{A}(x_i)\bar{A}(x_i) = A(x_i)$. Consider a weighted approximating basis and relabel, so that $\xi^B_{\tilde{J}}(P(x_i)) = \xi^B_{\tilde{J}}(P(x_i))\bar{A}(x_i)$. Define

$$G = E(\xi^B_{\tilde{J}}(P_i)\xi^B_{\tilde{J}}(P_i)'), \qquad \hat{G} = n^{-1}\sum_i \hat{\xi}^B_{\tilde{J}}(\hat{P}_i)\hat{\xi}^B_{\tilde{J}}(\hat{P}_i)', \quad \tilde{G} = n^{-1}\sum_i \xi^B_{\tilde{J}}(P_i)\xi^B_{\tilde{J}}(P_i)'.$$

Replace the ξ^B basis by the nonsingular linear transform $G^{-1/2}\xi^B$ to which the estimators are invariant, satisfying $\tilde{\kappa}_q(B) \geq c\kappa_q(B) = \sup_{|\mu|=q,\mathcal{P}} ||\partial^\mu G^{-1/2}\ \xi^B||$ where c will denote a positive constant whose value could vary in different parts of the proof. Similarly let $\alpha(S) \geq \sup_{\mathcal{X}} ||\chi^S(x)||$, $G_1 = E(\chi^S(x_i)\chi^S(x_i)')$ and consider the transformed basis $G_1^{-1/2}\chi^S$ satisfying $\tilde{\alpha}(S) \geq c\ \alpha(S) = \sup_{\mathcal{X}} ||G_1^{-1/2}\chi^S(x)||$.

Under these transforms the second moment matrix of the transformed bases is the identity. Thus, without loss of generality we will set $G = I_B$ and $G_1 = I_S$ for the remainder of the proofs.

As a preliminary step we will state a lemma that will be useful in proving Theorems 4.1-4.3.

LEMMA A1: *If Assumptions 1,2, 4-6 are satisfied,* $\phi_g = (n^{-1}S)^{1/2} + (B^{-v/2} + S^{-v_1/K})$ *and* $\phi_P = (n^{-1}S)^{1/2} + S^{-v_1/K}$ *then the following are each* $o_p(1)$:

1. $n^{1/2}[\kappa_q(B)^2\phi_g^2] = \{\kappa_q(B)^4[n^{-1}S^2]]^{1/2} + (n^{1/2}[B^{-v/2} + S^{-v_1/K}])^2[n^{-1}\kappa_q(B)^4]^{1/2}\}$ $(\forall q)$;
2. $\kappa_q(B)^2\phi_g = [\kappa_q(B)^4[n^{-1}S]]^{1/2} + [n^{-1}\kappa_q(B)^4]^{1/2}n^{1/2}[B^{-v/2} + S^{-v_1/K}]$ $(\forall q)$.

And, for power series,

$$\begin{aligned}
\kappa_0(B)^2 S\kappa_1(B)^2\phi_g^2 &\leq c\ n^{-1}B^8S^2 \\
\{B^{1/2}\kappa_1(B) + \kappa_0(B)^2\alpha(S)\}\phi_P &\leq c\ n^{-1}\{B^7S + B^4S^3\} \\
\kappa_o(B)^2B/n &\leq cn^{-1}B^3 \\
\alpha(S)^2S/n &\leq cn^{-1}S^3 \\
\{S\kappa_1(B)^2 + \kappa_0(B)^2\alpha(S)^4\}\phi_P^2 + n^{-1}B\alpha(S)^2 &\leq cn^{-1}(B^6S^2 + B^2S^5 + BS^2)
\end{aligned}$$

and for splines,

$$\begin{aligned}
\kappa_0(B)^2 S\kappa_1(B)^2\phi_g^2 &\leq cn^{-1}B^4S^2 \\
\{B\kappa_1(B)^2 + \kappa_0(B)^2\alpha(S)\}\phi_P &\leq c\ \ n^{-1}(B^4S + B^2S^3) \\
\kappa_o(B)^2B/n &\leq cn^{-1}B^2 \\
\alpha_o(S)^2S/n &\leq cn^{-1}S^2 \\
\{S\kappa_1(B)^2 + \kappa_0(B)^2\alpha(S)^4\}\phi_P^2 + n^{-1}B\alpha(S)^2 &\leq cn^{-1}(B^3S^2 + BS^3 + BS)
\end{aligned}$$

PROOF OF LEMMA A1: Note that $\phi_P = (S^{1/2}/n^{1/2})(1 + S^{-1/2}n^{1/2}S^{-v_1/K}) = O(n^{-1/2}S^{1/2})$. Similarly, $\phi_g = O(S^{1/2}/n^{1/2})$. By the inequalities derived in Newey (1997; Theorems 4, 7) for any nonnegative integer q, $\kappa_q(B) \leq CB^{1+2q}$ and $\alpha(S) \leq cS$ for power series bases and $\kappa_q(B) \leq cB^{0.5+q}$ and $\alpha(S) \leq cS^{0.5}$ for spline bases. Applying these inequalities, i) and ii) are $o_p(1)$ by Assumption 2(iii) and Assumption 5, and the remaining are $o_p(1)$ by Assumptions 1, 2(iii), 4-6.¥

COROLLARY OF LEMMA A1: *If Lemma A1 holds, then* $||\hat{G}_1 - I_S||$ *and* $||\hat{G} - I_B||$ *are each* $o_p(1)$.
PROOF: By $E[||\hat{G}_1 - I_S||^2] \leq cn^{-1}E[\sum_{k=1}^{S}\chi_k^S(x_i)\chi_k^S(x_i)' \sum_{j=1}^{S}\chi_j^S(x_i)\chi_j^S(x_i)'] = cn^{-1}\alpha_o(S)^2 S$ (Theorem 1 Newey, 1997), it follows by an application of Lemma A1 and the Markov inequality that $||\hat{G}_1 - I_S|| = o_p(1)$. By Lemma A1 (Newey, et al (1999)), $||\hat{G} - I_B|| = O_p(\phi_P[B^{1/2}\kappa_1(B) + \kappa_0(B)^2\alpha(L)] + n^{-1/2}\kappa_o(B)B^{1/2})$, which is $o_p(1)$ by an application of the rates derived in Lemma A1. ¥

LEMMA A2: *If Assumptions 1-3 are satisfied, then*

$$|\hat{g} - g_0|_q = O_p(\kappa_q(B)\phi_g).$$

PROOF OF LEMMA A2: This lemma uses Lemma A1 in Newey, Powell and Vella (1999) (henceforth, NPV), modifying it to address the presence of an identity in the second stage, which will raise the uniform convergence rate of the estimator $\hat{g}$ from that of a conventional two-stage estimator.

Let $\hat{g}_i = \hat{\omega}_i\hat{g}(\hat{P}_i)$, $\tilde{h}_i = \hat{\omega}_i g_0(P_i)$, $g_i = \omega_i g_0(P_i)$ each be $1 \times \tilde{J}$, and $\hat{g}$, $\tilde{g}$ and g represent the corresponding vectors (e.g., $\hat{g} = (\hat{g}_1', ..., \hat{g}_n')'$). By lemma 1 (Newey et al, 1999) $||\hat{\gamma} - \gamma|| = ||n^{-1}\hat{G}^{-1}\xi^{B\prime}\{(\Delta - g) + (g - \xi^{B\prime}\gamma) + (g - \tilde{g})|| = O_p(n^{-1/2}(B^{1/2} + S^{1/2}) + B^{-v/2} + S^{-v_1/K})$, so that

$$\begin{aligned}(\Delta - g_0) &= 0 \ \ w.p1, \\ &\Rightarrow \ ||\hat{\gamma} - \gamma|| = O_p(n^{-1/2}S^{1/2} + B^{-v/2} + S^{-v_1/K}) = O_p(\phi_g).\end{aligned}$$

It follows by Assumption 3 that

$$\begin{aligned}|\hat{g} - g_0|_q &\leq |\xi^{B\prime}(\hat{\gamma} - \gamma)|_q + |\xi^{B\prime}\gamma - g_0|_q \\ &= \kappa_q(B)||\hat{\gamma} - \gamma|| + O(B^{-v/2}) = O_p(\kappa_q(B)\phi_g)). \quad ¥\end{aligned}$$

PROOF OF THEOREM 4.1: This proof extends to nonlinear functionals Lemma A2 of Newey, Powell and Vella (1999).

Let $\Omega^{-1/2}$ denote a symmetric square root of Ω^{-1}. By $\mathrm{Var}(y_i|x_i)$ bounded below and $G_1 = I$, $V_\beta - cI$ is positive semidefinite. Notice that $E(\xi_J^B(P_i)\zeta_{\gamma,\beta})G_1^{-1}$ is the population sum of squares from the multivariate regression of $\xi_J^B(P_i)[\partial g(P_i)/\partial P]$ on χ_i^S, so that $||E(\xi_J^B(P_i)\zeta_{\gamma,\beta})\{E(\chi_i^S\chi_i^S)\}^{-1}|| \leq ||\xi_J^B(P_i)[\partial g/\partial P]|| \leq cI_B$ by boundedness of $\partial g/\partial P$ (Assumption 4) so that $V_\gamma - cI$ is positive semidefinite. Then, by $\Omega = \Lambda[V_\gamma]\Lambda' \geq \Lambda\Lambda'$,

$$(25) \qquad \begin{aligned} ||\Omega^{-1/2}\Lambda|| &= \{tr[\Omega^{-1/2}\Lambda\Lambda'\Omega^{-1/2}\Lambda]\}^{1/2} \\ &\leq \{tr[c\Omega^{-1/2}\Lambda V_\gamma'\Lambda\Omega^{-1/2\prime}]\}^{1/2} \leq tr\{c\Omega^{-1/2}\Lambda V_\gamma'\Lambda'\Omega^{-1/2}\}^{1/2} = c. \end{aligned}$$

Let $g_B = \xi^B\bar{\gamma}_B$. By Assumption 3 $|\lambda(g_J)| > 0$, while $(E(g_J(P_i)^2]) \to 0$. However, since $|\lambda(g_J)| = |\Lambda\bar{\gamma}_B| \leq ||\Lambda||\ ||\bar{\gamma}_B|| = ||\Lambda||(E(g_J(\cdot)^2])^{1/2}$, this implies that $||\Lambda|| \to \infty$. Therefore, $\Omega \geq \Lambda[V_\gamma]\Lambda' \geq c\ ||\Lambda||$, so that $\Omega^{-1/2}$ is bounded.

Now note that $n^{1/2}\Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0)]$ can be decomposed into the following sum of terms:

$$\begin{aligned} &n^{1/2}\Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0)] = n^{-1/2}\Omega^{-1/2}\Lambda\hat{G}^{-1}\hat{\xi}^{B\prime}[g_0 - \tilde{g}] \\ &+n^{-1/2}\Omega^{-1/2}\lambda(\xi^{B\prime}\bar{\gamma}_B - g_0) + n^{-1/2}\Omega^{-1/2}\Lambda\hat{G}^{-1}\hat{\xi}^{B\prime}(\tilde{g} - \hat{\xi}^{B\prime}\bar{\gamma}_B) \\ &+n^{1/2}\Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0) - F_g(g;\hat{g}) + F_g(g;g_0)] + o_p(1) \end{aligned}$$

where $\bar{\gamma}_B$ as in Assumption D(i). Let $\chi_{J-1}^S(x_i)$ abbreviate $I_{J-1} \otimes \chi_i^{S\prime}$. By lemma A2 of NPV (1999) all but the first and last term are $o_p(1)$ and $n^{-1/2}\Omega^{-1/2}\Lambda\hat{G}^{-1}\hat{\xi}^{B\prime}[g_0 - \tilde{g}] = n^{-1/2}\Omega^{-1/2}\Lambda[\xi_J^B\zeta_{\gamma,\beta}\{\chi_{J-1}^{S\prime}\eta\}]$. The last term, which is not present in their derivation, accounts correctly for the presence of a nonlinear functional, by giving rates for approximation of the functional derivative. By equation (25), the linearity of F_g in g, Assumption B(ii), and Lemma A1,

$$n^{1/2}\Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0) - F_g(g;\tilde{g}) - F_g(g;g_0)] \leq cn^{1/2}[|\hat{g} - g_0|_q]^2 O_p(n^{1/2}[\kappa_q(B)\phi_g]) = o_p(1).$$

It then follows by (A.27) in NPV and the above equation that

$$n^{1/2}\Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0)] = n^{-1/2}\Omega^{-1/2}\Lambda[\xi_J^B\zeta_{\gamma,\beta}\ \{\chi_{J-1}^{S\prime}\eta\}] + o_p(1).$$

For a vector υ such that $||\upsilon|| = 1$, define $\psi^S(w_i) = \upsilon'(n^{-1/2}\Omega^{-1/2}\Lambda[\xi_J^B(P_i)\ \{\zeta_{\gamma,\beta}\ \chi_{J-1}^S(x_i)'\eta_i\}])$, where $w_i = (\xi_J^B(P_i)', \chi^S(x_i)')'$ and $\eta_i = (\eta_{1i}, ..., \eta_{J-1i})'$ Notice that $\psi^S(w_i)$ are i.i.d random variables across i, with $\sum_i \psi^S(w_i) = n^{-1/2}\Omega^{-1/2}\ \Lambda[\xi_J^B(P_i)\zeta_{\gamma,\beta}\{\chi_{J-1}^{S\prime}\eta_i\}])$, $E[\psi^S(w_i)] = 0$ and $Var(\psi^S(w_i)) = n^{-1}$. Furthermore, by equation (25) $||\Omega^{-1/2}\Lambda|| \leq c$, and $cI_B - (\xi_J^B\zeta_{\gamma,\beta})(\xi_J^B\zeta_{\gamma,\beta})'$ positive semidefinite (shown prior to equation (25)) $||\Omega^{-1/2}\Lambda[\xi_J^B\zeta_{\gamma,\beta}]|| \leq c||\Omega^{-1/2}\Lambda||$. The remainder of lemma A2 of NPV holds as is, giving the proof of Theorem 4.1. ¥

To prove asymptotic normality with an estimate of the asymptotic covariance matrix in Theorem 4.2 we first give the following preliminary result.

LEMMA A3: *If Assumptions 1-6 are satisfied then* $||\hat{\Lambda} - \Lambda|| = o_p(1)$.

PROOF OF LEMMA A3: In the case of linear functionals $\hat{\Lambda} = \Lambda$; for the nonlinear case, by Lemma A2 and $F_g(g; \tilde{g})$ linear in $\tilde{g}$ (Assumption 2) it follows that

$$\begin{aligned} ||\hat{\Lambda} - \Lambda||^2 &= |F_g(\hat{\Lambda} - \Lambda)'\xi^B; \hat{g}) - F_g(\hat{\Lambda} - \Lambda)'\xi^B; g_0)| \leq c|(\hat{\Lambda} - \Lambda)'\xi^B|_q(\kappa(B)|\hat{g} - g_0|_q) \\ &= O_p(\kappa_q(B)\phi_g) = o_p(1) \text{ (by lemma A1).} \end{aligned}$$

PROOF OF THEOREM 4.2: Asymptotic normality with an estimator of the variance will primarily require showing that $\Omega^{-1/2}\hat{\Omega}\Omega^{-1/2\prime} \xrightarrow{p} 1$. To apply lemma A2 of NPV requires a multivariate generalization to show that $||\hat{V}_\beta - V_\beta|| = o_p(1)$. As before let $\chi^S_{J-1}(x_i)$ abbreviate $I_{J-1} \otimes \chi^{S\prime}_i$, $\hat{\eta}_i = (\hat{\eta}_{1i}, ..., \hat{\eta}_{J-1i})'$, and define

$$\hat{V}_\beta = \sum_{i=1}^{n} \chi^S_{J-1}(x_i)'\hat{\eta}_i\hat{\eta}_i'\chi^S_{J-1}(x_i), \quad \tilde{V}_\beta = \sum_{i=1}^{n} \chi^S_{J-1}(x_i)'\eta_i\eta_i'\chi^S_{J-1}(x_i). \tag{26}$$

Also define $\delta_P(x_i) = P(x_i) - \hat{P}(x_i)$. By Lemma A1, $\max_{i\leq n} \delta_P(x_i) = o_p(1)$. Note that $(\hat{\eta}_i\hat{\eta}_i' - \eta_i\eta_i') = \eta_i\delta_P(x_i)' + \delta_P(x_i)\eta_i' + \delta_P(x_i) + \delta_P(x_i)'$. Let e_L denote a $L \times 1$ unit vector. Define $\hat{\Sigma} = \frac{1}{n}\sum_i \chi^S_{J-1}(x_i)'(|\eta_i|e_{J-1}' + e_{J-1}|\eta_i|')\chi^S_{J-1}(x_i)$ and $\Sigma = E(\chi^S_{J-1}(x_i)'(|\eta_i|e_{J-1}' + e_{J-1}|\eta_i|')\chi^S_{J-1}(x_i))$ and note that by the boundedness of η (Assumption 1), $\Sigma \leq cE(\chi^S_{J-1}(x_i)'\chi^S_{J-1}(x_i)) = cI$. It then follows analogously to the argument showing $||\hat{G}_1 - I|| \xrightarrow{p} 0$ (Corollary A1) that $||\hat{\Sigma} - \Sigma|| \leq c||\hat{G}_1 - I|| \xrightarrow{p} 0$. Now for any vector ς with $||\varsigma|| \leq c$,

$$\begin{aligned} |\varsigma(\hat{V}_\beta - \tilde{V}_\beta)\varsigma'| &\leq |\varsigma n^{-1}(\sum_i \chi^S_{J-1}(x_i)'(\eta_i\delta_P(x_i)' + \delta_P(x_i)\eta_i' + \delta_P(x_i) + \delta_P(x_i)')\chi^S_{J-1}(x_i))\varsigma'| \\ &\leq \max_{i\leq n}|\delta_P(x_i)| \max_{i\leq n}|\delta_P(x_i)|\ \varsigma n^{-1}(\sum_i \chi^S_{J-1}(x_i)'(|\eta_i|e_{J-1}' + e_{J-1}|\eta_i|')\chi^S_{J-1}(x_i)') + \\ &\quad + \max_{i\leq n}|\delta_P(x_i)| \max_{i\leq n}|\delta_P(x_i)'|\ \varsigma n^{-1}(\sum_i \chi^S_{J-1}(x_i)'\chi^S_{J-1}(x_i)) \qquad (27) \\ &\leq o_p(1)|\varsigma(\hat{\Sigma} + \hat{G}_1)\varsigma'| \leq o_p(1)|\varsigma(\hat{\Sigma} + \hat{G}_1 - \Sigma - I)\varsigma'| + o_p(1)|\varsigma(\Sigma + I)\varsigma'| \\ &\leq o_p(1)||\varsigma||^2(||\hat{\Sigma} - \Sigma|| + ||\hat{G}_1 - I||) + o_p(1)||\varsigma||^2 \xrightarrow{p} 0. \end{aligned}$$

The remainder of the proof of lemma A2 applies NPV holds as is, so that by (26) and (27) above, $\Omega^{-1/2}\hat{\Omega}\Omega^{-1/2} \xrightarrow{p} 1$ under Assumptions 1-6, giving the conclusion of Theorem 4.2.

PROOF OF THEOREM 4.2: Let $r(x_i)$ abbreviate $[\partial g(P_i)/\partial P]$. Suppose Assumption 7(ii) is satisfied. Then, $\Lambda = E(\tau(P^{jk})\ \xi^B(P^{jk}))$. Let $\tau_B(P^{jk})$ denote the mean square projection of $\tau(P^{jk})$ on ξ^B, i.e., $\tau_B(P^{jk}) = \Lambda\xi^B(P_i^{jk})$. Define

$$d_{SB}(x_i) = E[r(x_i)\tau_B(P^{jk})\ \chi^S(x_i)']\chi^S(x_i)$$

as the mean square projection of $r(x_i)\tau_B(P^{jk})$ on $\chi^S(x_i)$ (by $E(\chi^S(x_i)\chi^S(x_i)') = I$)and let $d_S(x_i) = E[r(x_i)\tau(P^{jk})\ \chi^S(x_i)']\chi^S(x_i)$. By the boundedness of $r(x_i)$ (Assumption 4) and Assumption 7(iv),

$$E[(\tau(P^{jk}) - \tau_B(P^{jk}))^2] \leq E[\{\tau(P^{jk}) - \ \xi^B(P^{jk})'\bar{\gamma}_B\}] \to 0 \ \text{ as } B \to \infty.$$

Notice that $d_{SB}(x_i)$ and $d_S(x_i)$ are respectively the mean square projections of $r(x_i)\tau_B(P^{jk})$ and $r(x_i)\tau(P^{jk})$ on $\chi^S(x_i)$. Then,

$$E[\{d_{SB}(x_i) - d_S(x_i)\}^2] \leq E[r(x_i)^2||\tau_B(P^{jk}) - \tau(P^{jk})||^2] \to 0 \text{ as } B \to \infty.$$

Further, by the definition of $\mathcal{S}$ as the mean-square limit of $\chi^S(x)'\beta$ over all possible χ^S as $S \to \infty$, $E[||d_S(x_i) - d(x_i)||^2] \to 0$ as $S \to \infty$. Since mean square convergence of d_S implies convergence of second moments, by the boundedness of $\text{Var}(y_i|x_i)$ (Assumption 1),

$$\Omega = E(d_S(x_i)\text{Var}(y_i|x_i)d_S(x_i)'] \to E(d(x_i)\text{Var}(y_i|x_i)d(x_i)'] = \bar{\Omega}$$

so that $\bar{\Omega}$ is bounded under Assumption 7.

It now follows identically as in the proof of Theorem 4.1that $n^{1/2}\Omega^{-1/2}[\lambda(\hat{g}) - \lambda(g_0)] = n^{-1/2}\Omega^{-1/2}\Lambda[\xi^B_J\zeta_{\gamma,\beta}\ \{\chi^{S\prime}_{J-1}\eta\}] + o_p(1)$, except that now $\Omega^{-1/2} \longrightarrow 1/\bar{\Omega}^{1/2}$ so that $n^{1/2}[\lambda(\hat{g}) - \lambda(g_0)] = n^{-1/2}\Lambda[\xi^B_J\zeta_{\gamma,\beta}\ \{\chi^{S\prime}_{J-1}\eta\}] + o_p(1) = n^{-1/2}\sum_i \Lambda[\xi^B_J(P)\zeta_{\gamma,\beta}\ \{\chi^{S\prime}_{J-1}(x_i)\eta_i\}$and $n^{1/2}[\lambda(\hat{g}) - \lambda(g_0)] = \Omega^{1/2}\ n^{1/2}[\lambda(\hat{g}) - \lambda(g_0)]\Omega^{-1/2} \xrightarrow{d} N(0, \bar{\Omega})$. Further, as in the conclusion of Theorem 4.2 $\Omega^{-1/2}\hat{\Omega}\Omega^{-1/2} \xrightarrow{p} 1$ so that with $\Omega^{-1/2} \longrightarrow 1/\bar{\Omega}^{1/2}$, it follows that $\hat{\Omega} \xrightarrow{p} \bar{\Omega}$.¥

References

[1] Andrews, D.W.K (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Models," *Econometrica*, 59, 307-345.

[2] Ai, C, X. Chen and R. Blundell (2001): "Semiparametric Engel Curves with Endogenous Expenditure", working paper, UCL.

[3] Card, D. (2001): Estimating the return to schooling: Progress on some persistent econometric problems, *Econometrica*, 69, 5, 1127-1160.

[4] Cavanagh, C. and R.P. Sherman (1998): "Rank estimators for monotonic index models", *Journal of Econometrics*, 84, 351-381.

[5] Chamberlain, G.C (1994): "Panel Data" in Handbook of Econometrics, Vol 2, Z. Griliches and M. Intriligator (eds.), North-Holland, Amsterdam.

[6] Chay, Ken (1995): "Evaluating the Impact of the 1964 Civil Rights Act on the Economics Status of Black Men using Censored Longitudinal Earnings Data", working paper, Dept of Economics, Berkeley.

[7] Chay, Ken and J. Powell (2001): "Semiparametric Censored Regression Models", working paper.

[8] Chiang, C (1956): "On Regular Best Asymptotically Normal Estimates", *Annals of Mathematical Statistics*, 2, 336-351.

[9] Das, M (2001), "Estimation of a Panel Data Model with Insufficient Exclusion Restrictions", Working Paper, Dept. of Economics, Columbia University.

[10] Das, M, W. Newey and F. Vella (2001): "Nonparametric Sample Selection Models", working paper, Department of Economics, Columbia University.

[11] Gallant, A. R and G. Souza, (1991), "On the Asymptotic Normality of Fourier flexible functional form estimates", *Journal of Econometrics*, 50, 329-353.

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

[13] Han, A.K. (1987): "Nonparametic analysis of a generalized regression model: the maximum rank correlation estimator", *Journal of Econometrics*, 35, 303-316.

[14] Hausman, J.A and W.K. Newey (1995): "Nonparametric Estimation of exact Consumer Surplus and Deadweight Loss," *Econometrica*, 63, 1445-1476.

[15] Ichimura, H (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single index models", *Journal of Econometrics*, 58, 71-120.

[16] Klein, R and R.P Sherman (2001): "Shift Restrictions and Semiparametric Estimation in Ordered Response Models", Rutgers University and Cal Tech working paper, Forthcoming in *Econometrica*.

[17] Linton, O. and J. Nielsen (1995), "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration", *Biometrika*, 82, 93-100.

[18] Lorentz, G.G., *Approximation of Functions,* Chelsea, New York.

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

[20] ——— (1997): "Convergence Rates and Asymptotic Normality of Series Estimators," *Journal of Econometrics*, 79, 147-168.

[21] Newey, W.K., J.L Powell and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565-603.

[22] Porter, J (1996): "Nonparametric Regression Estimation for a Flexible Panel Data Model", MIT Thesis.

[23] Powell, M.J.D (1981), Approximation Theory and Methods. Cambridge University Press.

[24] Robinson, P.M (1976):"Instrumental Variables Estimation of Differential Equations," *Econometrica*, 4, 756-776.

[25] Schultz, T.P, 1997, Human Capital, Schooling and Health, IUSSP, XXIII General Population Conference, Yale University.

[26] Stewart, M. B. (1983): "On Least Squares Estimation when the Dependent Variable is Grouped", *Review of Economic Studies*, 737-753.

[27] Stoker, T.M (1986): "Consistent Estimation of Scaled Coefficients", *Econometrica*, 1461-1481.

[28] Stone, C.J. (1985): "Additive Regression and other Nonparametric Models", *Annals of Statisticsc*, 13, 689-705.

[29] White, H., (1980), A Heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity, *Econometrica*, 48, 817-838.

TABLE 1a
CROSS VALIDATION DISTRIBUTION OF SUMMANDS

S	POWER SERIES[a]	JOIN POINTS	CV Statistic 1964	CV Statistic 1971
2	x_1, x_1x_2	0,0	424.25	372.82
3	(S=2) + x_1^2	1,0	420.77	349.98
4	(S=3) + $x_1^2x_2$	1,1	419.99	347.97
5	(S=4) + $x_1^2x_2^2$	2,1	417.52	346.21
6	(S=5) + x_1^3	2,2	414.35	348.52
7	(S=6) + $x_1^3x_2^2$	3,2	416.34	345.15
8	(S=7) + $x_1^3x_2^3$	3,3	418.23	350.38

B	POWER SERIES[b]	JOIN POINTS	1964	1971
1	$(P^j)-(P^k)$	0,0	1952.22	1571.10
2	(B=1) + $\{(P^j)^2-(P^k)^2\}$	1,0	1938.50	1544.91
3	(B=2) + $\{(P^j)^3-(P^k)^3\}$	1,1	1905.89	1453.44
4	(B=3) + $\{(P^j)^4-(P^k)^4\}$	2,1	1904.11	1432. 67
5	(B=4) + $\{(P^j)^5-(P^k)^5\}$	2,2	1911.76	1441.80

L	POWER SERIES[a]	JOIN POINTS	CV Statistic 1964	CV Statistic 1971
2	x_1, x_1x_2	0,0	420.71	390.16
3	(K=2) + x_1^2	1,0	418.74	372.82
4	(K=3) + $x_1^2x_2$	1,1	410.56	353.02
5	(K=4) + $x_1^2x_2^2$	2,1	404.95	349.97
6	(K=5) + x_1^3	2,2	403.13	349.38
7	(K=6) + $x_1^3x_2^2$	3,2	406.32	345.15
8	(K=7) + $x_1^3x_2^3$	3,3	411.83	345.52

[a] x_1x_2 is the interaction of x_1 and x_2. The first entry of Join Points corresponds to x_1 and the second to x_2; each series includes a constant and the two binary regressors. L denotes the number of summands in the comparison estimator that replaces grouped dependent variable with an ad hoc number from its interval (see text).

[b] This series does not include a constant; the first entry of Join Points corresponds to P^j and the second to P^k; each series is the preceding series plus the additional term given above.

TABLE 2

EMPIRICAL APPLICATION: ESTIMATES[a]

	PARAMETER ESTIMATES							
	(1) Linear OLS		(2) Max. Likelihood		(3) Nonparametric		(4) Nonparametric MD	
	1964	1971	1964	1971	1964	1971	1964	1971
							First stage estimates	
x_1	.0275	.0256	.085	.079	-.438	-.524	-.0454	-.0685
	(.0042)	(.0048)	(.0047)	(.0045)	(.0743)	(.1451)	(.0223)	(.0291)
x_3	-.1090	-0.108	-0.588	-0.47	-.0936	.0796	.0213	.0287
	(.0478)	(.0542)	(.062)	(.059)	(.0480)	(.054)	(.015)	(.0157)
x_1^2					.0384	.0723	.0065	.0052
					(.0074)	(.0203)	(.0026)	(.0023)
x_1x_2					.0088	.0057	.0018	.0014
					(.0012)	(.0014)	(.0012)	(.0010)
$x_1^2x_2$					3.62e-5	.0001	-.0001	6.398780
					(8.38e-5)	(9.6e-5)	(.00005)	4e-006
$x_1^2x_2^2$					-1.35e-5	-1.2e-5	1.7e-6	1.8e-6
					(1.13e-6)	(1.3e-6)	(7.42e-7)	(1.21e-6)
x_1^3					-.0025	-.0032	-.0002	0010
					(.0007)	(.0008)	(.0009)	(.0008)
$1(x_1>12)$					.0297	.0497	.0138	.0326
$(x_1-12)^2$					(.0168)	(.0191)	(.0100)	(.0391)
							Second stage	
$\hat{P}^j - \hat{P}^k$							-5.627	-7.620
							(1.264)	(2.021)
$(\hat{P}^j)^2 - (\hat{P}^k)^2$							59.51	75.73
							(12.64)	(15.61)
$(\hat{P}^j)^3 - (\hat{P}^k)^3$							-106.05	-131.30
							(52.59)	(69.91)
$(\hat{P}^j)^4 - (\hat{P}^k)^4$							-36.78	-42.88
							(68.95)	(80.41)

	(B) AVERAGE DERIVATIVE ESTIMATES							
	(1)		(2)		(3)		(4)	
	1964	1971	1964	1971	1964	1971	1964	1971
$x_1 \le 12$					.0146	.0165	.0517	.0526
					(.0041)	(.0046)	(.0142)	(.0165)
$x_1 \in [13,16]$					.0239	.0186	.0723	.0654
	.0275	.0256	.085	.0796	(.0061)	(.0048)	(.0203)	(.0197)
$x_1 \ge 17$	(.0042)	(.0048)	(.0047)	(.0045)	.0337	.0280	.1003	.0985
					(.0231)	(.0159)	(.0673)	(.0635)
All x_1					. 0256	.0229	.0677	.0635
					(.0111)	(.0089)	(.0376)	(.0332)

[a] Standard errors in parenthesis. In column (4), first stage estimates are given for the y^2 binary regression. x_1 is education, x_2 is age and x_3 is race (Black=1). All regressions include a constant and a marriage indicator, columns (1)-(2) also includes x_2^2;columns (3)-(4) includes two additional join points in each year, a marriage indicator, $x_1^3x_2^3$ and $x_1^3x_2^2$ for column (3) and column (4), 1971.