Simple Estimators for Monotone Index Models

Hyungtaik Ahn
Dongguk University,

Hidehiko Ichimura
University College London,

James L. Powell
University of California, Berkeley
(powell@econ.berkeley.edu)

June 2004

Abstract

In this paper, estimation of the coefficients in a "single-index" regression model is considered under the assumption that the regression function is a smooth and strictly monotonic function of the index. The estimation method follows a "two-step" approach, where the first step uses a nonparametric regression estimator for the dependent variable, and the second step estimates the unknown index coefficients (up to scale) by an eigenvector of a matrix defined in terms of this first-step estimator. The paper gives conditions under which the proposed estimator is root-n-consistent and asymptotically normal.

JEL Classification: C24, C14, C13.

Acknowledgements

This research was supported by the National Science Foundation. Hyungtaik Ahn's research was supported by Dongguk Research Fund. We are grateful to Bo Honoré, Ekaterini Kyriazidou, Robin Lumsdaine, Thomas Rothenberg, Paul Ruud, and Mark Watson for their helpful comments.

1. Introduction

Estimation of the unknown coefficients β_0 in the single index regression model

$$E(y_i|\mathbf{x}_i) = G(\mathbf{x}_i'\beta_0), \tag{1.1}$$

where y_i and $\mathbf{x}_i$ are observable and $G(\cdot)$ is an unknown function, has been investigated in a number of papers in the econometric literature on semiparametric estimation. (A survey of these estimators is given in Powell (1994).) Some estimation methods, like the "average derivative" approach of Härdle and Stoker (1989) and Powell, Stock, and Stoker (1989) and the "density-weighted least squares" estimator of Ruud (1986) and Newey and Ruud (1991) exploit an assumption of smoothness (continuity and differentiability) of the unknown function G, but require all components of the regressor vector x to be jointly continuously distributed, which rarely applies in practice. Härdle and Horowitz (1996) has extended the average derivative estimator to allow for discrete regressors at the expense of introducing four additional nuisance parameters to be chosen by users of their estimator in addition to the standard smoothing parameter choice required in all nonparametric estimators. Other estimation methods which assume smoothness of G include the "single-index regression" estimators of Ichimura (1993a), Ichimura and Lee (1991), and, for the special case of a binary dependent variable, Klein and Spady (1993); these estimation methods permit general distributions of the regressors, but can be computationally burdensome, since they involve minimization problems with nonparametric estimators of G whose solutions cannot be written in a simple closed form. Still other estimators of were proposed for the "generalized regression model" proposed by Han (1987),

$$y_i = T(\mathbf{x}_i'\beta_0, \varepsilon_i), \tag{1.2}$$

where the unknown transformation $T(\cdot)$ is assumed to be monotonic in its first argument, and where the unobservable error term ε_i is assumed to be independent of $\mathbf{x}_i$. The assumed monotonicity of T, which implies monotonicity of G in (1.1), is fundamental for the consistency of the "maximum rank correlation" estimator of Han (1987) and the related monotonicity-based estimators of Cavanagh and Sherman (1991); like the "single index regression" estimators, computation of the "monotonicity" estimators is typically formidable, since it requires minimization of a criterion which may be discontinuous and involves a double sum over the data.

In this paper, which combines the results of Ahn (1995) and Ichimura and Powell (1996), both "smoothness" and monotonicity the nuisance function G are imposed – more specifically, it is assumed to be differentiable (up to a high order) and invertible in its argument. Simple "two-step" estimators are proposed under these restrictions; the first step obtains a nonparametric estimator of the conditional mean g_i of y_i given $\mathbf{x}_i$ using a standard (kernel) method, while the second step extracts an estimator of β_0 from a matrix defined using this first-step estimator. One estimator of the unknown coefficients is based upon the "eigenvector" approach that was used in a different context by Ichimura (1993b), and the corresponding second-step matrix estimator was considered (again in a different context) by Ahn and Powell (1993). An alternative, closed-form estimator of β_0 is also proposed; the relation of the "eigenvector" to the "closed form" estimation approach is analogous to the relation of limited information maximum likelihood (LIML) to two-stage least squares (2SLS) for simultaneous equations models. These estimators are computationally simple (since the second-step matrix estimator can be written in closed form), and do not require that all components of the regressor vector $\mathbf{x}_i$ are jointly continuously distributed. And, as shown below, they are root-n consistent (where n is the sample size) and asymptotically normal under regularity conditions that have been imposed elsewhere in the econometric literature on semiparametric estimation.

2. The Model and Estimator

Rewriting the single-index regression model (1.1) as

$$y_i \equiv g_i + u_i \equiv G(\mathbf{x}_i'\beta_0) + u_i, \tag{2.1}$$

where

$$g_i \equiv G(\mathbf{x}_i'\beta_0) \equiv E[y_i|\mathbf{x}_i] \tag{2.2}$$

is the conditional mean of the (scalar) dependent variable y_i given the p-dimensional vector of regressors $\mathbf{x}_i$ (so the unobservable u_i has $E[u_i|\mathbf{x}_i] = 0$), the maintained assumption that G is monotonic implies

$$\lambda(g_i) \equiv G^{-1}(g_i) = \mathbf{x}_i'\beta_0. \tag{2.3}$$

That is, given the conditional mean g_i of y_i,

$$0 = \mathbf{x}_i'\beta_0 - \lambda(g_i) \tag{2.4}$$

for some unknown transformation $\lambda(g_i)$ of g_i . Clearly β_0 could only be identified up to a scale normalization from this relation; given such a normalization, though, (2.4) can be used to identify the remaining components of β_0, provided the regressors $\mathbf{x}_i$ are sufficiently variable when the conditional mean g_i is held fixed. Specifically, for a pair of observations with the same conditional mean g_i, the parameter vector β_0 must be orthogonal to the difference in regressors, i.e.,

$$\begin{aligned} 0 &= (\mathbf{x}_i - \mathbf{x}_j)'\beta_0 + \lambda(g_j) - \lambda(g_i) \\ &= (\mathbf{x}_i - \mathbf{x}_j)'\beta_0 \qquad\qquad \text{if} \qquad g_i = g_j. \end{aligned} \tag{2.5}$$

Therefore, letting $w(\mathbf{x}_i, \mathbf{x}_j) \equiv w_{ij}$ be any nonnegative weighting function of the pair of regressors $\mathbf{x}_i$ and $\mathbf{x}_j$, the coefficient vector β_0 satisfies

$$\beta_0'\mathbf{\Sigma}_w\beta_0 = 0, \tag{2.6}$$

where $\mathbf{\Sigma}_w \equiv E[\mathbf{\Sigma}_w(g_j)]$ and

$$\Sigma_w(s) = E[w(\mathbf{x}_i, \mathbf{x}_j) \cdot (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)' | g_i = s] \tag{2.7}$$

assuming these moments exist. Provided the matrix $\mathbf{\Sigma}_w$ has rank $(p-1)$ — so that any other nontrivial linear combination $(\mathbf{x}_i - \mathbf{x}_j)'\alpha$ of the difference of regressors has nonzero variance conditional on $(\mathbf{x}_i - \mathbf{x}_j)'\beta_0 = 0$ and α is not proportional to β_0 — the parameter vector is identified (up to scale) as the eigenvector corresponding to the unique zero eigenvalue of the matrix $\mathbf{\Sigma}_w$, which depends only on the joint distribution of the observable $(y_i, \mathbf{x}_i')$.

A natural approach to transform this identification result into an estimation method for β_0 would be to first estimate the unobservable conditional expectation terms $g_i \equiv E[y_i|\mathbf{x}_i]$ by some nonparametric method, then estimate a sample analogue to the matrix $\mathbf{\Sigma}_w$ using pairs of observations with estimated values $\hat{g}_i$ of g_i that were approximately equal. Such an estimation strategy was proposed, in a somewhat different context, by Ahn and Powell (1993); in that paper, the first-step nonparametric estimator was the familiar kernel estimator, which takes the form of a weighted average of the dependent variable,

$$\hat{g}_i \equiv \frac{\sum_{i=1}^n K_{ij} \cdot y_j}{\sum_{i=1}^n K_{ij}}, \tag{2.8}$$

with weights K_{ij} given by

$$K_{ij} \equiv K\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{h_1}\right), \tag{2.9}$$

for $K(\cdot)$ a "kernel" function which tends to zero as the magnitude of its argument increases, and $h_1 \equiv h_{1n}$ a first-step "bandwidth" which is chosen to tend to zero as the sample size n increases. Given this estimator $\hat{g}_i$ of the conditional mean variable g_i, a second-step estimator of a matrix analogous to $\mathbf{\Sigma}_w$ was defined by Ahn and Powell (1993) as

$$\hat{\mathbf{S}} \equiv \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \hat{\omega}_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'; \tag{2.10}$$

the weights $\hat{\omega}_{ij}$ took the form

$$\hat{\omega}_{ij} \equiv \frac{1}{h_2}\, k\left(\frac{\hat{g}_i - \hat{g}_j}{h_2}\right)\, t_i t_j, \tag{2.11}$$

where $k(\cdot)$ is a univariate kernel analogous to K above, $h_2 \equiv h_{2n}$ is a bandwidth sequence for the second-step estimator $\hat{\mathbf{S}}$, and $t_i = t(\mathbf{x}_i)$ is a "trimming" term which is chosen to equal zero for observations where $\hat{g}_i$ is known to be imprecise (i.e., where $\mathbf{x}_i$ is outside some prespecified compact subset of its support). The weighting function $\hat{\omega}_{ij}$ in (2.11) declines to zero as $\hat{g}_i - \hat{g}_j$ increases relative to the bandwidth h_2; thus, the conditioning event "$g_i = g_j$" in the definition of $\mathbf{\Sigma}_w$ is ultimately imposed as this bandwidth shrinks with the sample size (and the nonparametric estimator of g_i converges to its true value in probability).

Adopting this estimator $\hat{\mathbf{S}}$ of $\mathbf{\Sigma}_w$ (which implies a particular definition of the weighting function $w(\mathbf{x}_i, \mathbf{x}_j)$ in (2.6), described below), a corresponding estimator of β_0 would exploit a sample analogue of relation (2.6) based on the eigenvectors of $\hat{\mathbf{S}}$. Though the matrix $\mathbf{\Sigma}_w$ will be positive semi-definite under the regularity conditions to be imposed, the estimator $\hat{\mathbf{S}}$ need not be for any finite sample. Hence, the estimator $\hat{\beta}$ of β_0 is defined here as the eigenvector for the eigenvalue of $\hat{\mathbf{S}}$ that is closest to zero in magnitude. That is, defining $(\hat{\nu}_1, ..., \hat{\nu}_p)$ to be the p solutions to the determinantal equation

$$|\hat{\mathbf{S}} - \nu \mathbf{I}| = 0, \tag{2.12}$$

the estimator $\hat{\beta}$ is defined as an appropriately-normalized solution to

$$(\hat{\mathbf{S}} - \hat{\nu}\mathbf{I})\hat{\beta} = \mathbf{0}, \tag{2.13}$$

where

$$\hat{\nu} \equiv \arg\min_j |\hat{\nu}_j|. \tag{2.14}$$

A convenient normalization for $\hat{\beta}$ (and β_0) imposes the additional restriction that a particular component of β_0 (say, the first) is known to be nonzero, and is normalized to unity, so that the remaining coefficients are identified relative to that value. Specifically, writing $\hat{\beta}$ and β_0 as

$$\hat{\beta} = \begin{pmatrix} 1 \\ -\hat{\theta} \end{pmatrix}, \qquad \beta_0 = \begin{pmatrix} 1 \\ -\theta_0 \end{pmatrix}, \tag{2.15}$$

and partitioning $\hat{\mathbf{S}}$ conformably as

$$\hat{\mathbf{S}} \equiv \begin{bmatrix} \hat{\mathbf{S}}_{11} & \hat{\mathbf{S}}_{12} \\ \hat{\mathbf{S}}_{21} & \hat{\mathbf{S}}_{22} \end{bmatrix}, \tag{2.16}$$

the solution to (2.13) takes the form

$$\hat{\theta} \equiv [\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1} \cdot \hat{\mathbf{S}}_{21} \tag{2.17}$$

for this normalization.

The "smallest eigenvalue" approach used here was used in an earlier paper by Ichimura (1993b) to construct a two-step estimator for the binary response model under a conditional median restriction, suggested as a computationally-simpler alternative to the maximum score estimator for this model proposed by Manski (1975, 1985). For the binary response model with a continously-distributed, conditional-mean-zero error term, the conditional mean $g_i \equiv E[y_i|\mathbf{x}_i]$ of the binary dependent variable equals one-half if and only if the underlying regression function $\mathbf{x}_i'\beta_0$ also equals zero; by the same reasoning as given for the estimator $\hat{\theta}$ in the present paper, Ichimura proposed estimation of β_0 using the eigenvector for the smallest eigenvalue of a matrix of weighted averages of the cross-products of the regressors, with kernel weights (like those above) depending upon the deviation of the estimated conditional means $\{\hat{g}_i\}$ from one-half. Though this estimator was shown to be consistent and asymptotically normal, its rate of convergence was slower than the square root of the sample size, unlike the asymptotic theory for the present estimator $\hat{\theta}$ derived in the next section.

An alternative "closed form" estimator $\tilde{\theta}$ of θ_0 can be defined as

$$\tilde{\theta} \equiv [\hat{\mathbf{S}}_{22}]^{-1} \cdot \hat{\mathbf{S}}_{21}; \tag{2.18}$$

this estimator is motivated by rewriting (2.4) as

$$x_{i1} = \mathbf{x}_{i2}'\theta_0 + \lambda(g_i) + v_i, \tag{2.19}$$

which is in the same form as the "selectivity bias" model treated by Ahn and Powell (1993), with $\mathbf{x}_i = (x_{i1}, \mathbf{x}_{i2}')'$ and with error term v_i which is identically zero for this application. In light of the motivation for $\hat{\theta}$ given above, the alternative estimator $\tilde{\theta}$ can be viewed as exploiting the fact (to be verified below) that $\hat{\nu}$ tends to zero in probability, since the smallest eigenvalue of the probability limit $\mathbf{\Sigma}_w$ of $\hat{\mathbf{S}}$ is zero. The relation of $\hat{\theta}$ to $\tilde{\theta}$ here is analogous to the relationship between the two classical single-equation estimators for simultaneous equations systems, namely, limited-information maximum likelihood, which has an alternative derivation as a least-variance ratio (LVR) estimator, and two-stage least squares (2SLS), which can be viewed as a modification of LVR which replaces an estimated eigenvalue by its known (zero) probability limit. The analogy to these classical estimators extends to the asymptotic distribution theory for $\hat{\theta}$ and $\tilde{\theta}$, which, under the conditions imposed below, will be asymptotically equivalent, like LVR and 2SLS. Their relative advantages and disadvantages are also analogous – e.g., $\tilde{\theta}$ is simpler to compute, while $\hat{\theta}$ will be equivariant with respect to choice of which (nonzero) component of β_0 to normalize to unity. As noted in the introduction, both estimators will be much easier to calculate than many of the existing estimators for the single-index regression model, which typically require solution of a p-dimensional minimization problem with a criterion involving a double-summation over the observations.

3. Large Sample Properties of the Estimator

Since the definition of the estimator $\hat{\beta} = (1, -\hat{\theta}')'$ is based on the same form of a "pairwise difference" matrix estimator $\hat{\mathbf{S}}$ analyzed in Ahn and Powell (1993), it is most convenient to impose the same regularity conditions from that paper, and to derive the asymptotic theory for the present estimator using the large-sample characterizations previously obtained. The appendix below lists analogues of the eleven assumptions imposed by Ahn and Powell (1993), modified to fit the present problem and notation. The necessity and generality of those assumptions were discussed at length in that previous paper; here, then, those conditions are only briefly reviewed, noting any differences between the current assumptions and their earlier counterparts.

In the assumptions in the appendix, the regression vector $\mathbf{x}_i$ is assumed to have only discretely- and continuously-distributed components, and high-order moments (namely, six) of y_i and $\mathbf{x}_i$ are assumed to exist. The conditional mean $g_i = E[y_i|\mathbf{x}_i] = G(\mathbf{x}_i'\beta_0)$ is assumed to be continuously

distributed, with density function denoted by $f(\cdot)$; it is also assumed that the density f and various conditional expectations of functions of $\mathbf{x}_i$ given $g_i = g$ are very smooth in the argument g, i.e., they have high-order derivatives which have well-behaved distributions when evaluated at $g_i = g$. One of these functions is the conditional expectation of the trimming variable $\tau(\mathbf{x}_i)$, which is assumed to be bounded above and to decline smoothly to zero outside some compact set $\mathbf{X}$ for which $f(g) = f(G(\mathbf{x}'\beta_0))$ is bounded away from zero on $\mathbf{X}$. The functions $K(\cdot)$ and $k(\cdot)$ for the first- and second-step estimators are assumed to be "higher-order" kernels, with the number of vanishing moments of K depending upon the number of continuous components of $\mathbf{x}_i$, and with the first three moments of k equalling zero (i.e., k is a "fourth-order kernel"). Likewise, the bandwidth terms h_1 and h_2 are assumed to converge to zero at particular rates as the sample size n increases; these conditions, combined with the "smoothness" and "higher-order kernel" assumptions, ensure that the bias of various implicit nonparametric estimators is of smaller order than the square root of the sample size, and is therefore negligible for the first-order distribution theory.

Under these conditions, the results of Ahn and Powell (1993) imply that the estimator $\hat{\mathbf{S}}$ of (2.10) above converges in probability to a matrix $\mathbf{\Sigma}_0$, which is a special case of the general matrix $\mathbf{\Sigma}_w$ of (2.7), with the particular weighting function

$$w_0(\mathbf{x}_i, \mathbf{x}_j) \equiv t_i t_j (f_i f_j)^{1/2} = t_i t_j f_i, \tag{3.1}$$

where $f_i \equiv f(g_i)$ is the density of g_i and the last equality imposes the conditioning event $g_i = g_j$. Using iterated expectations, the matrix $\mathbf{\Sigma}_0$ can be rewritten as

$$\begin{aligned} \mathbf{\Sigma}_0 &\equiv E\left[2 f_i [E(t_i|g_i) \cdot E(t_i \mathbf{x}_i \mathbf{x}_i'|g_i) - E(t_i \mathbf{x}_i|g_i) \cdot E(t_i \mathbf{x}_i'|g_i)\right] \\ &\equiv E\left[2 f(g_i) [\tau_i(g_i) \cdot \mu_{\mathbf{xx}}(g_i) - \mu_{\mathbf{x}}(g_i) \cdot \mu_{\mathbf{x}}(g_i)'\right], \end{aligned} \tag{3.2}$$

$$\begin{aligned} \tau(g) &\equiv E[t_i|g_i = g], \\ \mu_{\mathbf{x}}(g) &\equiv E[t_i \mathbf{x}_i|g_i = g], \qquad \text{and} \\ \mu_{\mathbf{xx}}(g) &\equiv E[t_i \mathbf{x}_i \mathbf{x}_i'|g_i = g]. \end{aligned} \tag{3.3}$$

It is easy to verify that $\mathbf{\Sigma}_0 \beta_0 = 0$; the final assumption in the appendix is the identification condition, which asserts that the null space of $\mathbf{\Sigma}_0$ only consists of scalar multiples of β_0. And,

imposing the normalization $\beta_0 = (1, -\theta_0)$, this requires that $rank(\mathbf{\Sigma}_0) = p-1 = rank(\mathbf{\Sigma}_{22})$, where $\mathbf{\Sigma}_{22}$ is the lower-right $(p-1)$-dimensional submatrix of $\mathbf{\Sigma}_0$.

Under these conditions, spelled out precisely in the appendix, obvious modifications of the arguments for Lemma 3.1 and Theorem 3.1 of Ahn and Powell (1993) yield the following large-sample properties of the estimator $\hat{\mathbf{S}}$:

Lemma 3.1: Under Assumptions A.1 through A.11 in the appendix below,

$$\begin{aligned} (i) \qquad \hat{\mathbf{S}} - \mathbf{\Sigma}_0 &= o_p(1), \qquad \text{and} \\ (ii) \qquad \sqrt{n}\hat{\mathbf{S}}\,\boldsymbol{\beta}_0 &= \frac{2}{\sqrt{n}} \sum_{i=1}^{n} t_i f_i \lambda'(g_i)[\tau(g_i)\mathbf{x}_i - \mu_{\mathbf{x}}(g_i)] \cdot u_i + o_p(1), \end{aligned}$$

where $u_i \equiv y_i - g_i$, $\lambda'(g) = d\lambda(g)/dg$, for λ is defined in (2.3), and the remaining terms are defined in (3.1) - (3.3) above.■

Consistency of the estimator $\hat{\theta}$ in (2.17) above could be verified directly using result (i) and the identification condition A.4, but it is simpler to derive the asymptotic distribution of $\hat{\theta}$ using result (ii), from which consistency of $\hat{\theta}$ immediately follows. The asymptotic linearity expression in (ii) above is the analogue to the result (ii) of Theorem 3.1 of Ahn and Powell (1993), exploiting the relation $\lambda(g) = \mathbf{x}_i'\beta_0$ (so $\hat{\mathbf{S}}\,\boldsymbol{\beta}_0$ is the same as "$\hat{\mathbf{S}}_{z\lambda}$" of that paper, with $\mathbf{z}_i \equiv \mathbf{x}_i$). The terms in the normalized average in expression (ii) have zero mean and finite variance, so the Lindeberg-Levy central limit theorem implies that $\hat{\mathbf{S}}\,\boldsymbol{\beta}_0$ is asymptotically normal; however, this asymptotic distribution will be singular, since

$$\beta_0'[\tau(g_i)\mathbf{x}_i - \mu_{\mathbf{x}}(g_i)] = [\tau(g_i)\lambda(g_i) - E(t_i\lambda(g)|g_i)] = 0, \tag{3.4}$$

again using $\lambda(g) = \mathbf{x}_i'\beta_0$. It follows that

$$\sqrt{n}\beta_0'\hat{\mathbf{S}}\,\boldsymbol{\beta}_0 = o_p(1), \tag{3.5}$$

which further implies that the smallest (in magnitude) eigenvalue $\hat{\nu}$ converges in probability to zero faster than the square root of the sample size, because

$$\sqrt{n}\,|\hat{\nu}| = \sqrt{n} \min_{\alpha \neq 0} |\alpha'\hat{\mathbf{S}}\,\alpha|/\|\alpha\|^2 \leq \sqrt{n}\,|\beta_0'\hat{\mathbf{S}}\,\beta_0|/\|\beta_0\|^2 = o_p(1). \tag{3.6}$$

To derive the asymptotic distribution of the proposed estimator $\hat{\theta}$ of (2.17), the normalized difference of $\hat{\theta}$ and θ_0 can be decomposed as

$$\begin{aligned}\sqrt{n}\,(\hat{\theta}-\theta_0) &= [\hat{\mathbf{S}}_{22}-\hat{\nu}\mathbf{I}]^{-1}\sqrt{n}\,[\hat{\mathbf{S}}_{12}-(\hat{\mathbf{S}}_{22}-\hat{\nu}\mathbf{I})\,\theta_0] \\ &= [\hat{\mathbf{S}}_{22}-\hat{\nu}\mathbf{I}]^{-1}\sqrt{n}\,\hat{\mathbf{s}}-\sqrt{n}\,\hat{\nu}\,[\hat{\mathbf{S}}_{22}-\hat{\nu}\mathbf{I}]^{-1}\theta_0, \end{aligned} \tag{3.7}$$

where

$$\hat{\mathbf{s}} \equiv \hat{\mathbf{S}}_{12}-\hat{\mathbf{S}}_{22}\theta_0 \equiv [\hat{\mathbf{S}}\,\beta_0]_2, \tag{3.8}$$

i.e., $\hat{\mathbf{s}}$ is the subvector of $\hat{\mathbf{S}}$ corresponding to the free coefficients θ_0 . From result (i) of Lemma 3.1 and (3.6), it follows that

$$\hat{\mathbf{S}}_{22}-\hat{\nu}\,\mathbf{I} \rightarrow^{p} \mathbf{\Sigma}_{22} \tag{3.9}$$

and

$$\sqrt{n}(\hat{\theta}-\theta_0) = [\mathbf{\Sigma}_{22}]^{-1}\sqrt{n}\,\hat{\mathbf{s}}+o(1), \tag{3.10}$$

from which the consistency and asymptotic normality of $\hat{\theta}$ follow from the asymptotic normality of $\hat{\mathbf{S}}\,\boldsymbol{\beta}_0$. A similar argument yields the asymptotic equivalence of Ahn's estimator $\tilde{\theta}$ of (2.18) and the estimator $\hat{\theta}$, since

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\tilde{\theta}) &= \sqrt{n}([\hat{\mathbf{S}}_{22}-\hat{\nu}\,\mathbf{I}]^{-1}-[\hat{\mathbf{S}}_{22}]^{-1})\hat{\mathbf{S}}_{12} \\ &= -\sqrt{n}\,\tilde{\nu}\,[\hat{\mathbf{S}}_{22}-\hat{\nu}\,\mathbf{I}]^{-1}\hat{\theta} \\ &= o_p(1) \end{aligned} \tag{3.11}$$

for $\tilde{\nu}$ an intermediate value between $\hat{\nu}$ and zero.

The results of these calculations are summarized in the following proposition:

Theorem 3.1: Under Assumptions A.1 through A.11, the estimator $\hat{\theta}$ defined in (2.17) has the asymptotic linear representation

$$\sqrt{n}(\hat{\theta}-\theta_0) = \mathbf{\Sigma}_{22}^{-1}\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i + o_p(1),$$

where

$$\psi_i \equiv 2\,t_i\,f(g_i)\,\lambda'(g_i)(E[t_i|g_i]\mathbf{x}_{i2}-E[t_i\mathbf{x}_{i2}|g_i])\,(y_i-g_i),$$

and is asymptotically normal,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d \mathcal{N}(0, \mathbf{\Sigma}_{22}^{-1}\mathbf{\Omega}\mathbf{\Sigma}_{22}^{-1}),$$

where $\mathbf{\Omega} \equiv E[\psi_i \psi_i']$. Also, $\hat{\theta}$ and the estimator $\tilde{\theta}$ of (2.18), proposed by Ahn (1995), are asymptotically equivalent,

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}) = o_p(1).\blacksquare$$

A final requirement for conducting the usual large-sample normal inference procedures is a consistent estimator of the asymptotic covariance matrix $\mathbf{\Sigma}_{22}^{-1}\mathbf{\Omega}\mathbf{\Sigma}_{22}^{-1}$ of $\hat{\theta}$. Estimation of $\mathbf{\Sigma}_{22}^{-1}$ is straightforward; by result (i) of Lemma 3.1 and (3.6) above, either $[\hat{\mathbf{S}}_{22} - \hat{\nu}\mathbf{I}]^{-1}$ or $[\hat{\mathbf{S}}_{22}^{-1}]^{-1}$ will be consistent, with the former being more natural for $\hat{\theta}$ and the latter for $\tilde{\theta}$. Consistent estimation of the matrix $\mathbf{\Omega}$ is less straightforward, but, as Ahn (1995) points out, one consistent estimator would be

$$\hat{\mathbf{\Omega}} \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_i \hat{\psi}_i, \tag{3.12}$$

for

$$\hat{\psi}_i \equiv \frac{1}{n-1}\sum_{i=1}^{n} \hat{\delta}_{ij}(\mathbf{x}_{i2} - \mathbf{x}_{j2})(\mathbf{x}_{i2} - \mathbf{x}_{j2})', \tag{3.13}$$

where

$$\hat{\delta}_{ij} \equiv \left(\frac{1}{h_2}\right)^2 k'\left(\frac{\hat{g}_i - \hat{g}_j}{h_2}\right) t_i\, t_j \tag{3.14}$$

and $k'(\cdot)$ denotes the derivative of the second-step kernel $k(\cdot)$. The argument for Ahn's (1995) Theorem 3.2 applies here as well, and yields consistency of $\hat{\mathbf{\Omega}}$.

4. Topics for Further Research

Though the large-sample theory of the previous section used a specific (kernel) form for the first-step nonparametric estimator $\hat{g}_i$ of $g_i = E[y_i|\mathbf{x}_i]$, it seems likely the results of Lemma 3.1 and Theorem 3.1 could be established using other initial nonparametric estimators of this conditional mean – like series, nearest neighbor, or locally linear regression methods – under analogous regularity conditions for those estimators. (Indeed, Ahn's (1995) analysis used a slightly different specification of the kernel estimator than in the present paper.) This would be useful because the first-step kernel estimator, while theoretically convenient, may well be problematic for practical implementation of

the procedure. In particular, the proposed estimation method exploits the monotonicity of g_i in terms of the index $\mathbf{x}_i'\beta_0$, but the kernel estimator may require relatively large samples to accurately reflect this monotonicity, and the second-step estimator will be sensitive to "oversmoothing" in the first step. Consider, for example, a sample with $t_i = 1$ for all observations (so that all $\mathbf{x}_i$ lie in the prespecified compact set $\mathbf{X}$). In this case, as the first-step bandwidth h_1 tends to infinity, $\hat{g}_i$ tends to the sample average $\bar{y}$ of the dependent variable for all observations, and the matrix $\hat{\mathbf{S}}$ tends to a constant multiple of the sample covariance of the regressors, whose smallest eigenvalue needs not tend to zero in large samples, and whose corresponding eigenvector bears no necessary relation to β_0. This suggests that a first-step estimation method whose "oversmoothed" limit was non-constant might have better finite-sample performance than the present kernel estimator. (For example, g_i might be estimated by the sum of a linear least-squares fit of y_i on $\mathbf{x}_i$ and a kernel regression estimate of the conditional mean of the residuals from that fit.) Whether such alternative first-step estimators are theoretically valid and practically useful is a good topic for further work.

On a related topic, the "faster-than-root-n" convergence of the smallest eigenvalue $\hat{\nu}$ to zero would be expected to fail if the single-index specification (1.1) for the conditional mean of y_i is not satisfied, which suggests that a normalized version of $\hat{\nu}$ might be used to test whether the single-index specification is indeed correct. However, the derivation of the asymptotic distribution of $\hat{\nu}$ is not straightforward, and is related to the "asymptotic singularity" issue that arises in the nonparametric specification testing literature (e.g., see ãit Sahalia, Bickel, and Stoker (1994)), so the large-sample properties of $\hat{\nu}$ will require additional work.

5. APPENDIX: Regularity Conditions

With modifications for the present problem and notation, conditions 3.1 through 3.11 of Ahn and Powell (1993) are translated here as follows:

Assumption A.1 (Random Sampling and Bounded Moments): The vectors $(y_i, \mathbf{x}_i')'$ are independently and identically distributed across i, with all components having finite sixth-order moments.

Assumption A.2 (Correctly-Specified Model): The data satisfy the monotone single-index

regression model described in (2.1), (2.2), and (2.3) above.

Assumption A.3 (Continuous Distribution of Index): The conditional distribution of $g_i \equiv E[y_i|\mathbf{x}_i] = G(\mathbf{x}_i'\beta_0)$ is absolutely continuous with respect to Lebesgue measure, with (conditional) density function $f(\cdot)$ that is continuous and bounded from above.

Assumption A.4 (Identification): The matrix $\mathbf{\Sigma}_0$, defined in (3.2), and its lower-right $(p-1)\times(p-1)$ submatrix $\mathbf{\Sigma}_{22}$ have rank $p-1$.

Assumption A.5 (Kernel Regularity, Second Step): The kernel function $k(\cdot)$ used to define the weights $\hat{\omega}_{ij}$ in (2.11) above satisfies

(i) $k(u)$ is twice differentiable, with $k''(u) < k_0$ for some k_0;

(ii) $k(u) = k(-u)$;

(iii) $k(u) = 0$ if $|u| > l_0$ for some $l_0 > 0$; and

(iv) $\int uk(u)du = 0$ for $u = 1$, 2, and 3.

Assumption A.6 (Bandwidth Rates, Second Step): The bandwidth sequence h_2 used to define the weights $\hat{\omega}_{ij}$ of (2.11) is of the form

$$h_2 = c_n \cdot n^{-\delta},$$

where the positive sequence c_n has $c_0 < c_n < c_0^{-1}$ for some $c_0 > 0$, and $\delta \in (1/8, 1/6)$.

Assumption A.7 (Smooth Density and Conditional Expectations): The conditional density function $f(u)$ of g_i , the functions $\tau(g)$, $\mu_{\mathbf{x}}(g)$, and $\mu_{\mathbf{xx}}(g)$ (defined in (3.3) above), and the function $\lambda(g) \equiv G^{-1}[g]$ and its derivative $\lambda'(g) \equiv d\lambda(g)'dg$ are all fourth-order continuously differentiable, with derivatives that are bounded for all g in the support of g_i.

Assumption A.8 (Distribution of Conditioning Variables): After an appropriate reordering, the vector $\mathbf{x}_i$ of regressors can be partitioned as $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})'$, where $\mathbf{x}_i^{(1)}$ is continuously distributed and $\mathbf{x}_i^{(2)}$ is discrete. Furthermore, if $\phi(\mathbf{x}^{(1)}|\mathbf{x}^{(2)})$ is the conditional density function of $\mathbf{x}_i^{(1)}$ given $\mathbf{x}_i^{(2)} = \mathbf{x}^{(2)}$, then for each $\mathbf{x}$ in some known, compact subset $\mathbf{X}$ of the support of $\mathbf{x}_i$, the following conditions hold:

(i) $\phi(\mathbf{x}^{(1)}|\mathbf{x}^{(2)}) > \phi_0$ for some $\phi_0 = \phi_0(\mathbf{x}^{(2)}) > 0$.

(ii) Defining $\xi_i \equiv \xi(\mathbf{x}_i) \equiv (1, g_i, f_i, \tau_i, \lambda'(g_i), \mathbf{x}_i', \mu_{\mathbf{x}}(g)')'$, the function $\xi(\mathbf{x}) \cdot \phi(\mathbf{x}^{(1)}|\mathbf{x}^{(2)})$ is bounded and M-times continuously differentiable with bounded derivatives in $\mathbf{x}^{(1)}$, for some even integer $M > m/(1/3 - 2\delta)$, where $m = dim(\mathbf{x}^{(1)})$ and δ is given in Assumption 3.6 above.

(iii) The functions $E[y_i^2|\mathbf{x}_i = \mathbf{x}] \cdot \phi(\mathbf{x}^{(1)}|\mathbf{x}^{(2)})$ and $g(\mathbf{x}) \cdot \phi(\mathbf{x}^{(1)}|\mathbf{x}^{(2)})$ are continuous on $\mathbf{X}$.

(iv) The number of points of support of $\mathbf{x}^{(2)}$ in $\mathbf{X}$ is finite.

Assumption A.9 (Exogenous Trimming): The indicator variable t_i is constructed so that $t_i > 0$ only if $\mathbf{x}_i \in \mathbf{X}$, where the compact set $\mathbf{X}$ satisfies the restrictions in Assumption A.8 above.

Assumption A.10 (Kernel Regularity, First Step): The kernel function $K(\cdot)$ used to define the estimator $\hat{g}_i$ in (2.8) and (2.9) above is of the form

$$K(u) = \sum_{i=1}^{M/2} a_j \rho(u; b_j C),$$

where

(i) the even integer M satisfies the conditions of Assumption A.8 (ii);

(ii) $\rho(u; C)$ is the density function of a $\mathcal{N}(0, C)$ random vector;

(iii) C is an arbitrary positive definite matrix;

(iv) $b_1, ..., b_{M/2}$ are arbitrary, distinct, positive constants; and

(v) the constants $a_1, ..., a_{M/2}$ satisfy the linear equations

$$\sum_j a_j = 1, \qquad \sum_j a_j b_j^q = 0 \qquad \text{for } q = 1, ..., M/2 - 1.$$

Assumption A.11 (Bandwidth Rates, First Step): The bandwidth sequence h_1 used to define the estimator g_i in (2.8) and (2.9) is of the form

$$h_1 = d_n \cdot n^{-\gamma},$$

where the positive sequence d_n has $d_0 < d_n < d_0^{-1}$ for some $d_0 > 0$, and $\gamma \in (1/2M, (1/6 - \delta)/m)$, for M, δ, and m given in Assumptions 3.6 and 3.8 (ii).

6. References

Ahn, H., 1995, "Estimation of monotonic single index models," manuscript, Department of Economics, Virginia Polytechnic Institute.

Ahn, H. and J.L. Powell, 1993, "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3-29.

Aït Sahalia, Y., P. Bickel, and T. Stoker, 1995, "Goodness of fit tests for regression using kernel methods," manuscript, Sloan School of Management, M.I.T.

Cavanagh, C. and R. Sherman, 1991, "Rank estimators for monotonic regression models," manuscript, Bellcore.

Han, A.K., 1987, "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics* 35, 303-316.

Härdle, W. and T.M. Stoker, 1989, "Investigating smooth multiple regression by the method of average derivatives," *Journal of the American Statistical Association*, 84, 986-995.

Härdle, W. and Horowitz, 1996, "Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632-1640.

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

Ichimura, H., 1993b, "Local quantile regression estimation of binary response models with conditional heteroskedasticity," manuscript, Department of Economics, University of Minnesota.

Ichimura, H. and L. Lee, 1991, "Semiparametric least squares estimation of multiple index models: single equation estimation," in: W.A. Barnett, J.L. Powell, and G. Tauchen, eds., *Nonparametric and semiparametric methods in econometrics and statistics*, Cambridge: Cambridge University Press.

Ichimura, H. and J.L. Powell, 1996, "A simple estimator for monotone single index models, " manuscript, Department of Economics, Princeton University.

Klein, R.W. and R.S. Spady, 1993, "An efficient semiparametric estimator of the binary response model," *Econometrica*, 61, 387-422.

Manski, C.F., 1975, "Maximum score estimation of the stochastic utility model of choice," *Journal of Econometrics*, 3, 205-228.

Manski, C.F., 1985, "Semiparametric Analysis of discrete response, asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27, 205-228.

Newey, W.K. and P. Ruud, 1991, "Density weighted least squares estimation," manuscript, Department of Economics, U.C. Berkeley.

Powell, J.L., 1994, "Estimation of semiparametric models," in R.F. Engle and D.F. McFadden, eds., *Handbook of Econometrics, Volume 4*, Amsterdam: North Holland.

Powell, J.L., J.H. Stock and T.M. Stoker, 1989, "Semiparametric estimation of weighted average derivatives," *Econometrica* 57, 1403-1430.

Ruud, P., 1986, "Consistent estimation of limited dependent variable models despite misspecification of distribution," *Journal of Econometrics*, 32, 157-187.