RANK ESTIMATORS FOR MONOTONIC INDEX MODELS

By Christopher Cavanagh[a] and Robert P. Sherman[b1]

[a] *Department of Economics, Columbia University*
[b] *Division of Humanities and Social Sciences, Caltech*

Abstract

We present a new class of rank estimators of scaled coefficients in semiparametric monotonic linear index models. The estimators require no subjective bandwidth choices and have attractive computational properties. We establish $\sqrt{n}$-consistency and asymptotic normality, and provide the general form and consistent estimators of the asymptotic covariance matrix. We also provide a generalization covering single equation multiple indices models satisfying certain monotonicity constraints. An analogue of consistency when all explanatory variables are categorical is established, and an application is presented.

KEYWORDS: Rank estimators, semiparametric monotonic linear index models, computational efficiency, U-processes, models with multiple indices, categorical explanatory variables.
JEL Classification: C13; C14.

[1]Robert Sherman, Caltech, Division of Humanities and Social Sciences, Pasadena, CA 91125, phone: 818-395-4038, fax: 818-405-9841, email address: sherman@hss.caltech.edu

1. Introduction

In many econometric applications, it is natural to expect a monotonic relationship between a response variable and an associated index. We expect wages to be increasing in an index of human capital. We often assume that binary choice probabilities are increasing in an index of individual and choice characteristics. And so on. Moreover, we often take the index to be linear in the explanatory variables.

While it may be reasonable to assume a monotonic relationship between a response and a linear index, it is usually difficult to specify the exact nature of the monotonicity. Perhaps even more difficult is specifying the exact form of the error distribution for the model. It is well known that misspecifications of either type can cause standard estimators to produce inconsistent estimates of the index parameters and other estimands that depend on these parameters.

It is therefore desirable to develop estimators of index parameters for semiparametric monotonic linear index models: estimators that directly exploit monotonicity between a response and a linear index without requiring any knowledge, beyond regularity conditions, about the form of the monotonic relationship or the error distribution.

Let $Z = (Y, X)$ be an observation from a distribution P on a set $S \subseteq \mathbb{R} \otimes \mathbb{R}^d$, where Y is a response variable and X is a vector of regressor variables. Han (1987) introduced the semiparametric monotonic linear index model

$$Y = D \circ F(X'\beta_0, u) \tag{1}$$

where $X'\beta_0$ is the linear index with β_0 a d-dimensional vector of unknown parameters, u is a random disturbance independent of X, the function F is strictly increasing in each of its arguments, and the function D is nonconstant and increasing. The model is semiparametric in that no parametric assumptions are made about the distribution of u or the functional form of $D \circ F$. Many interesting econometric models fit into this framework, including linear regression, binary choice, ordered choice, censored regression, and various transformation and duration models.

Let Z_i, $i = 1, \ldots, n$, be a sample of independent observations from the distribution P. Han proposed estimating β_0 in (1) with

$$\underset{\mathcal{B}}{\operatorname{argmax}} \sum_{i<j} \left[\{Y_i > Y_j, X_i'\beta > X_j'\beta\} + \{Y_i < Y_j, X_i'\beta < X_j'\beta\} \right] \tag{2}$$

where $\mathcal{B}$ is a subset of $\mathbb{R}^d$ and $\{\cdot\}$ denotes the indicator function of a set. He called the estimator the maximum rank correlation (MRC) estimator since it maximizes Kendall's (1938) measure of rank correlation between Y and $X'\beta$. Han also proved strong consistency. Sherman (1993) established $\sqrt{n}$-consistency and asymptotic normality.

In addition to covering a variety of interesting econometric models, the MRC estimator has the appealing property that it requires no subjective bandwidth choice. This is in contrast to kernel-based competitors (see, for example, the work of Ichimura, 1993 and Klein and Spady, 1993) which produce a different estimate of β_0 for each of a wide range of possible bandwidth choices.

In this paper, we introduce a new class of rank estimators of β_0 for semiparametric monotonic linear index models. Like the MRC estimator, these rank estimators require no subjective bandwidth choices. In addition, the estimators exploit monotonicity in a more natural and verifiable fashion, may allow more flexibility in balancing robustness and efficiency objectives, cover a wider range of models, and in general, are far more computationally efficient than the MRC estimator.

Let M denote an increasing function on $\mathbb{R}$. For real numbers $a_1, \ldots, a_n$, let $R_n(a_i)$ denote the rank of a_i. We propose estimating β_0 in (1) with

$$\beta_n \equiv \underset{\mathcal{B}}{\operatorname{argmax}} \sum_i M(Y_i) R_n(X_i'\beta) \tag{3}$$

for an appropriate subset $\mathcal{B}$ of $\mathbb{R}^d$. For ease of notation we have suppressed the dependence of β_n on M.

When $M(\cdot) = R_n(\cdot)$, the maximand in (3) is a linear function of Spearman's (1904) measure of rank correlation between Y and $X'\beta$ (Lehmann, 1975, Chapter 7). One would expect this measure to be maximized at β_0 for any model satisfying (1).

If robustness is critical, $M(\cdot) = R_n(\cdot)$ is a natural choice. However, more efficient estimates may be obtained using the identity function $M(y) = y$. An intermediate choice would be the winsorized function

$$M(y) = a\{y < a\} + y\{a \le y \le b\} + b\{y > b\}$$

for real numbers $a < b$.

The key condition driving the consistency of β_n is

$$\mathbb{E}[M(Y) \mid X] \text{ is a nonconstant, increasing function of } X'\beta_0\,. \tag{4}$$

As suggested in the opening paragraph, condition (4) is a very natural one to assume in many econometric problems. Moreover, one can make a simple visual check of (4) by constructing a nonparametric regression estimator of the regression function in (4) based on the estimated index $X'\beta_n$. (We illustrate this in the final section.) Further, condition (4) holds for any model satisfying (1) since $\mathbb{E}M(D \circ F(t, \cdot))$ is increasing in t even if F is increasing only in its first argument. In fact, neither this weaker condition on F nor the independence of u and X are needed for condition (4) to hold. To illustrate, take $M(y) = y$ and consider the linear regression model

$$Y = X'\beta_0 + h(X'\beta_0)u$$

where h is a nonnegative function on the support of $X'\beta_0$. If $I\!E(u \mid X) = 0$, then no matter what h is,

$$I\!E[Y \mid X'\beta_0 = t] = t$$

and so condition (4) is satisfied. Moreover, if the support of both u and $X'\beta_0$ is $I\!R$, then it is easy to find nonnegative functions h for which $F(t,u) = t + h(t)u$ is not increasing in t for all u. Examples include $h(t) = |t|^\alpha$ for $\alpha > 0$. So, while (1) provides a convenient description of an interesting subclass of models for which (4) holds, we wish to put greater emphasis on condition (4), since it is not only the primary condition driving the consistency of β_n, but also includes a richer set of structural forms than does (1).

The key condition driving the consistency of the MRC estimator is

$$I\!P\{Y_i > Y_j \mid X_i, X_j\} \geq I\!P\{Y_i < Y_j \mid X_i, X_j\} \quad \text{when} \quad X_i'\beta_0 \geq X_j'\beta_0\,. \tag{5}$$

This condition follows from the single index nature of the model in (1), the monotonicity of $D \circ F$, and the assumed independence of u and X. It is easy to construct examples where (4) holds but (5) does not. Return to the previous example, but for simplicity, assume that u and X are independent and $h(0) = 0$. If $X_1'\beta_0 = t > 0$ and $X_2'\beta_0 = 0$, then (5) reduces to

$$I\!P\{u > -t/h(t)\} \geq I\!P\{u < -t/h(t)\}\,.$$

If the median of u is less than $-t/h(t)$ for any t, then (5) is violated, while (4) is not. It is also not clear how one would check assumption (5).

Next, note that in general, each evaluation of the maximand in (2) requires $O(n^2)$ operations, while only $O(n \log n)$ operations are required to compute the maximand in (3). This is because ranking the $X_i'\beta$'s is essentially a sorting procedure, and there exist sorting algorithms requiring only $O(n \log n)$ comparisons (see, for example, section 3.4 of Aho et alia (1976)). This can mean a substantial savings in computation time, especially if one wishes to bootstrap various quantities of interest. We illustrate some computational advantages in Section 6.

As a final point of comparison with the MRC estimator, we show that in the special case of the binary choice model, the MRC estimator and β_n are identical if either M is a deterministic function or $M(Y_i) = R_n(Y_i)$. For real numbers $a_1, \ldots, a_n$, define

$$R_n(a_i) = \sum_j \{a_i > a_j\}\,.$$

For $k = 0, 1$, define

$$R_n^{(k)}(X_i'\beta) = \sum_j \{Y_j = k\}\{X_i'\beta > X_j'\beta\}\,.$$

Also, let n_k denote the number of Y_i's equal to k. Note that $R_n(Y_i) = 0$ if $Y_i = 0$ and $R_n(Y_i) = n_0$ if $Y_i = 1$. The maximand in (2) equals

$$\begin{aligned} & \sum_i \sum_j \{Y_i > Y_j, X_i'\beta > X_j'\beta\} \\ = & \sum_i \sum_j \{Y_i = 1\}\{Y_j = 0\}\{X_i'\beta > X_j'\beta\} \\ = & \sum_i Y_i R_n^{(0)}(X_i'\beta)\,. \end{aligned}$$

When M is a deterministic function, the maximand in (3) equals

$$\begin{aligned} & \sum_i M(Y_i)\left[R_n^{(0)}(X_i'\beta) + R_n^{(1)}(X_i'\beta)\right] \\ = & \; c_n + c\sum_i Y_i R_n^{(0)}(X_i'\beta) + c\sum_i Y_i R_n^{(1)}(X_i'\beta) \end{aligned}$$

where $c_n = M(0)\sum_i R_n(X_i'\beta) = M(0)n(n+1)/2$ and $c = M(1) - M(0)$. The equivalence follows from

$$\sum_i Y_i R_n^{(1)}(X_i'\beta) = \sum_i \sum_j Y_i Y_j \{X_i'\beta > X_j'\beta\} = (n_1 - 1)n_1/2\,.$$

That is, this latter term does not depend on β. Similarly, when $M(Y_i) = R_n(Y_i)$, the maximand in (3) equals

$$n_0 \sum_i Y_i R_n(X_i'\beta) = n_0 \sum_i Y_i R_n^{(0)}(X_i'\beta) + n_0(n_1 - 1)n_1/2\,.$$

In the next section, we establish the consistency of β_n when M is either a deterministic function, or when $M(Y_i) = R_n(Y_i)$. In Section 3, we establish $\sqrt{n}$-consistency and asymptotic normality, and show how to obtain the general form and consistent estimators of the asymptotic covariance matrix. Section 4 provides a generalization to single equation multiple indices models satisfying certain monotonicity constraints. In Section 5, we establish a notion of consistency in the important special case when all the explanatory variables are categorical, and in Section 6, present an application. Section 7 summarizes and provides some concluding remarks.

2. Consistency

In this section we establish consistency of β_n when M is either deterministic or when $M(Y_i) = R_n(Y_i)$. The proof will be given for the former case, and then we will show how it easily extends to cover the latter.

Expand the rank function in (3) into a sum, and for each β in $\mathcal{B}$ write

$$G_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} M(Y_i)\{X_i'\beta > X_j'\beta\}\,.$$

Then β_n maximizes $G_n(\beta)$ over $\mathcal{B}$. Once again, for ease of notation we have suppressed the dependence of $G_n(\beta)$ on M. Note that $\{G_n(\beta) : \beta \in \mathcal{B}\}$ is a U-process of order 2.

We use the following assumptions in the consistency proof:

A0. $\mathbb{E}[M(Y) \mid X]$ depends on X only through $X'\beta_0$.
A1. $\mathbb{E}[M(Y) \mid X]$ is a nonconstant, increasing function of $X'\beta_0$.
A2. The support of X is not contained in a proper linear subspace of $\mathbb{R}^d$.
A3. The dth component of X has an everywhere positive Lebesgue density, conditional on the other components.
A4. $\mathcal{B}$ is a compact subset of $\{\beta \in \mathbb{R}^d : \beta_d = 1\}$.
A5. $\mathbb{E}[M(Y)]^2 < \infty$.

Assumption A0 is the single index assumption. Note that, in general, A0 is a weaker assumption than (1). Assumption A1 is the key monotonicity assumption, and along with A2 through A4, ensures that β_0 can be identified. Together, A2 and A4 imply that β_0 can only be identified up to location and scale. For the consistency proof that follows, assumption A5 can be weakened to a finite first moment for $M(Y)$, but the normality result presented in the next section requires a finite second moment. Note that when $M(\cdot)$ is either the rank function or the winsorized function discussed in the introduction, then $M(\cdot)$ is bounded and A5 is trivially satisfied.

THEOREM 1: *If A0 through A5 hold, then* $|\beta_n - \beta_0| = o_p(1)$.

PROOF. Write $G(\beta)$ for $\mathbb{E}M(Y_1)\{X_1'\beta > X_2'\beta\}$. Note that $G(\beta)$ is the expected value of $G_n(\beta)$. We will show

(i) $G(\beta)$ is uniquely maximized at β_0.

(ii) $\sup_{\mathcal{B}} |G_n(\beta) - G(\beta)| = o_p(1)$.

(iii) $G(\beta)$ is continuous.

Consistency will then follow from standard arguments using the compactness of $\mathcal{B}$. (See, for example, Amemiya (1985, pp. 106–107).)

Invoke A0 and write $H(X'\beta_0)$ for $\mathbb{E}[M(Y) \mid X]$. By symmetry,

$$G(\beta) = \frac{1}{2}\mathbb{E}[H(X_1'\beta_0)\{X_1'\beta > X_2'\beta\} + H(X_2'\beta_0)\{X_1'\beta < X_2'\beta\}]\,. \qquad (6)$$

If $\beta = \beta_0$, then A1 and A3 ensure that the indicators in (6) pick out the larger of $H(X_1'\beta_0)$ and $H(X_2'\beta_0)$ with probability one. Consequently,

$$G(\beta_0) = \frac{1}{2}\mathbb{E}\max\left(H(X_1'\beta_0), H(X_2'\beta_0)\right) .$$

Deduce that $G(\beta)$ is maximized at β_0.

We now show that β_0 is the unique maximizer.

Suppose that for some β in $\mathcal{B}$,

$$G(\beta) = \frac{1}{2}\mathbb{E}\max\left(H(X_1'\beta_0), H(X_2'\beta_0)\right) . \tag{7}$$

Deduce from (7) and (6) that

$$H(X_1'\beta_0) \geq H(X_2'\beta_0) \quad \text{when} \quad X_1'\beta > X_2'\beta . \tag{8}$$

Let $S_{\mathcal{X}}$ denote the support of $\mathcal{X} = (X_1, \ldots, X_{d-1})$ and write $CH_{\mathcal{X}}$ for the convex hull of $S_{\mathcal{X}}$. That is, $CH_{\mathcal{X}}$ is the smallest convex set containing $S_{\mathcal{X}}$. Assumption A2 implies that $CH_{\mathcal{X}}$ is a $(d-1)$-dimensional subset of $\mathbb{R}^{d-1}$ and so has a nonempty interior. Select a point μ from this interior and define

$$I_\mu = \{(\mu, t) : t \in \mathbb{R}\} .$$

Assumption A1 guarantees the existence of a point t_0 in the support of $X'\beta_0$ for which

$$H(t) < H(t_0) \quad \text{for} \quad t < t_0 .$$

Choose ι_0 in I_μ for which $\iota_0'\beta_0 = t_0$. Such a point can always be found since A3 and A4 together imply that $\{\iota'\beta_0 : \iota \in I_\mu\} \equiv \mathbb{R}$. Define the d-dimensional open wedges

$$W_1(\beta) = \{x'\beta_0 < \iota_0'\beta_0\}\{x'\beta > \iota_0'\beta\}$$

and

$$W_2(\beta) = \{x'\beta_0 > \iota_0'\beta_0\}\{x'\beta < \iota_0'\beta\} .$$

(Note that we can replace β and β_0 with their respective unit vectors without changing $W_1(\beta)$ and $W_2(\beta)$. Thus, for each x in $\mathbb{R}^d$ and each β in $\mathcal{B}$, we may view $x'\beta$ as the orthogonal projection of x onto the space spanned by β.) If $X_1 \in W_1(\beta)$ and $X_2 \in W_2(\beta)$ then

$$H(X_1'\beta_0) < H(X_2'\beta_0) \quad \text{while} \quad X_1'\beta > X_2'\beta .$$

Then in order for (8) to hold we must have

$$\mathbb{P}\{X_1 \in W_1(\beta)\}\mathbb{P}\{X_2 \in W_2(\beta)\} = 0 . \tag{9}$$

We now show that (9) only holds for $\beta = \beta_0$.

For each β in $\mathcal{B}$, define

$$H_\beta = \{x'\beta = \iota_0'\beta\}$$

and

$$L_\beta = H_\beta \cap H_{\beta_0}\,.$$

Note that $W_1(\beta)$ and $W_2(\beta)$ are delimited by the $(d-1)$-dimensional hyperplanes H_β and H_{β_0}, and for $\beta \neq \beta_0$, L_β is a $(d-2)$-dimensional hyperplane in $\mathbb{R}^d$. Consider the projections

$$P_0(\beta) = \{x \in CH_\mathcal{X} : (x,t) \in L_\beta \text{ for some } t \in \mathbb{R}\}$$

and, for $j = 1, 2$,

$$P_j(\beta) = \{x \in CH_\mathcal{X} : (x,t) \in W_j(\beta) \text{ for some } t \in \mathbb{R}\}\,.$$

That is, $P_0(\beta)$ projects L_β into $CH_\mathcal{X}$ and $P_j(\beta)$ projects $W_j(\beta)$ into $CH_\mathcal{X}$. Also note that $\{P_j(\beta),\, j = 0, 1, 2\}$ partitions $CH_\mathcal{X}$.

Since both H_β and H_{β_0} contain ι_0, L_β must contain ι_0. Since ι_0 is an element of I_μ, $P_0(\beta)$ must contain μ. Since μ is an interior point of $CH_\mathcal{X}$, $P_0(\beta)$ cannot contain an entire $(d-2)$-dimensional face of $CH_\mathcal{X}$. But then each $P_j(\beta)$ must contain at least one point of $S_\mathcal{X}$, implying

$$\int_{P_j(\beta)\cap S_\mathcal{X}} G_\mathcal{X}(dx) > 0 \tag{10}$$

where $G_\mathcal{X}(\cdot)$ denotes the distribution of $\mathcal{X}$.

For each x in $S_\mathcal{X}$, write l_x for the line through x parallel to the dth coordinate axis. If $\beta \neq \beta_0$, then there must be a nonzero angle between H_β and H_{β_0} and so at least one of H_β and H_{β_0} must intersect l_x. Write $t_\beta(x)$ for the dth component of $H_\beta \cap l_x$. If $H_\beta \cap l_x$ is null, define $t_\beta(x) = \infty$. Then

$$\mathbb{P}\{X \in W_j(\beta)\} = \int_{P_j(\beta)\cap S_\mathcal{X}} \left[\int_{\min(t_{\beta_0}(x),\, t_\beta(x))}^{\max(t_{\beta_0}(x),\, t_\beta(x))} f(t \mid x)\, dt\right] G_\mathcal{X}(dx)$$

where $f(\cdot \mid x)$ denotes the conditional density of X_d given $\mathcal{X} = x$. Since $\beta \neq \beta_0$, $t_{\beta_0}(x) \neq t_\beta(x)$ for each x in $S_\mathcal{X}$. This, A3, and (10) imply that $\mathbb{P}\{X \in W_j(\beta)\} > 0$, contradicting (9). This establishes (i).

Next, recall that $Z = (Y, X)$ denotes an observation from the distribution P on the set S. For each β in $\mathcal{B}$ and each (z_1, z_2) in $S \otimes S$ define

$$f(z_1, z_2, \beta) = M(y_1)\{x_1'\beta > x_2'\beta\} - G(\beta)\,. \tag{11}$$

Then

$$G_n(\beta) - G(\beta) = U_n f(\cdot, \cdot, \beta)$$

where U_n denotes the random measure putting mass $1/[n(n-1)]$ on each pair (Z_i, Z_j), $i \neq j$. That is, $\{U_n f(\cdot,\cdot,\beta) : \beta \in \mathcal{B}\}$ is a zero-mean U-process of order 2. A trivial modification of the argument given in Sherman (1993, Section 6) shows that $\{f(\cdot,\cdot,\beta) : \beta \in \mathcal{B}\}$ is Euclidean for the envelope $|M(y_1)| + P|M(\cdot)|$. Deduce from A5 and Corollary 7 of Sherman (1994, Section 6) that

$$\sup_{\mathcal{B}} |U_n f(\cdot,\cdot,\beta)| = O_p(1/\sqrt{n})\,.$$

This is more than enough to establish (ii).

Finally, fix $\beta \in \mathcal{B}$ and let $\{\beta(m)\}$ denote a sequence of elements of $\mathcal{B}$ converging to β as m tends to infinity. Let Q denote the product measure $P \otimes P$. Assumption A3 implies that

$$Q\{x_1'\beta = x_2'\beta\} = 0\,.$$

This in turn implies that

$$M(y_1)\{x_1'\beta(m) > x_2'\beta(m)\} - M(y_1)\{x_1'\beta > x_2'\beta\} \to 0 \quad \text{as} \quad m \to \infty \qquad (12)$$

for Q almost all (z_1, z_2). Take expectations, then apply the dominated convergence theorem and A5 with $2|M(y_1)|$ as the dominating function to establish (iii). This proves the theorem. □

REMARK. The first time through the proof of Theorem 1 it may be somewhat difficult to follow the argument for unique maximization of $G(\beta)$ at β_0. This difficulty can be overcome by working through the argument with the following simple example in mind: Take $d = 2$, $\mathcal{B} = \{(t,1) : t \in [-r,r], r > 0\}$, and $\beta_0 = (0,1)$. Note that $\mathcal{X} = X_1$. Take $S_{\mathcal{X}} = \{-1,1\}$ so that $CH_{\mathcal{X}} = [-1,1]$. Take $\mu = 0$ so that $I_\mu = \{(0,t) : t \in I\!R\}$. Suppose $H(X'\beta_0)$ has a point of increase at $X'\beta_0 = 0$ so that we may take $t_0 = 0$. This implies that $\iota_0 = (0,0)$ and so $W_1(\beta) = \{x'\beta_0 < 0\}\{x'\beta > 0\}$, $W_2(\beta) = \{x'\beta_0 > 0\}\{x'\beta < 0\}$, $H_\beta = \{x'\beta = 0\}$, and for $\beta \neq \beta_0$, $L_\beta = \iota_0 = (0,0)$. This, in turn, implies that $P_0(\beta) = \{0\}$, $P_1(\beta) = [-1,0)$, and $P_2(\beta) = (0,1]$. The faces of $CH_{\mathcal{X}}$ are the points -1 and 1. The lines $l_{\pm 1} = \{(\pm 1, t) : t \in I\!R\}$, $t_{\beta_0}(\pm 1) = \infty$, and $t_\beta(\pm 1)$ for $\beta \neq \beta_0$ are real numbers.

Once this simple example is mastered, it is easy to see how the proof works in general for $d = 2$, and then for $d > 2$.

Finally, we show how to extend the proof of Theorem 1 to cover the special case $M(y) = R_n(y)$. If $M(Y_i) = R_n(Y_i) = \sum_j \{Y_i > Y_j\}$, and we rescale by $(n)_3^{-1}$ where $(n)_3 = n(n-1)(n-2)$, then the maximand in (3) becomes

$$(n)_3^{-1} \sum_{i,j,k} \{Y_i > Y_j\}\{X_i'\beta > X_k'\beta\}\,.$$

It is easy to show that the $j = k$ terms make a negligible asymptotic contribution and so we may take

$$G_n(\beta) = (n)_3^{-1} \sum_{\boldsymbol{i_3}} \{Y_i > Y_j\}\{X_i'\beta > X_k'\beta\}$$

where $\boldsymbol{i_3} = (i, j, k)$ ranges over the $(n)_3$ triples of distinct integers from the set $\{1, \ldots, n\}$. The function $H(X'\beta_0)$ becomes

$$\mathbb{E}[M(Y_1) \mid X_1] = (n-1)\mathbb{P}[\{Y_1 > Y_3\} \mid X_1]$$

and

$$G(\beta) = \mathbb{P}\{Y_1 > Y_3\}\{X_1'\beta > X_2'\beta\}\,.$$

Replace $M(y_1)$ with $\{y_1 > y_3\}$ in (11) and (12) and the proof goes through exactly as before, except that in establishing (ii), we invoke Corollary 7 of Sherman (1994, Section 6) with $k = 3$ instead of $k = 2$, since here $G_n(\beta) - G(\beta)$ is a zero-mean U-process of order 3.

Remark. Theorem 1 can be strengthened to almost-sure convergence if we replace (ii) with a uniform strong law of large numbers. The latter result can be obtained provided $\mathbb{E}[M(Y)]^{4+\epsilon} < \infty$ for some $\epsilon > 0$. This follows from applying the Hoeffding decomposition (see pp. 177–178 of Serfling (1980)) to $G_n(\beta) - G(\beta)$ and then invoking Corollary 9 in Section 6 of Sherman (1994) to handle each of the degenerate pieces.

3. The Limiting Distribution of β_n

In this section, we establish $\sqrt{n}$-consistency and asymptotic normality of β_n when M is either deterministic or when $M(Y_i) = R_n(Y_i)$. We also show how to obtain the general form and consistent estimators of the asymptotic covariance matrix.

Recall once again that $Z = (Y, X)$ denotes an observation from the distribution P on the set $S \subseteq \mathbb{R} \otimes \mathbb{R}^d$ and that the parameter space $\mathcal{B}$ is a compact subset of $\{\beta \in \mathbb{R}^d : \beta_d = 1\}$.

First consider the case where M is deterministic. For each (z_1, z_2) in $S \otimes S$ and each β in $\mathcal{B}$ define

$$f(z_1, z_2, \beta) = M(y_1)\{x_1'\beta > x_2'\beta\}\,. \tag{13}$$

Then, for each z in S and each β in $\mathcal{B}$ define

$$\tau(z, \beta) = f(z, P, \beta) + f(P, z, \beta) \tag{14}$$

where $f(z, P, \beta)$, for example, is short for the conditional expectation of $f(\cdot, \cdot, \beta)$ given its first argument. The function $\tau(\cdot, \beta)$ turns out to be the kernel function

of the empirical process that drives the asymptotic behavior of β_n, and is derived from the Hoeffding decomposition of $f(\cdot,\cdot,\beta)$ into its orthogonal components (again, see Serfling (1980, pp. 177-178)).

Write ∇_m for the mth partial derivative operator applied to the first $d-1$ components of β, and

$$|\nabla_m|\sigma(\beta) \equiv \sum_{i_1,\ldots,i_m} \left| \frac{\partial^m}{\partial\beta_{i_1}\cdots\partial\beta_{i_m}}\sigma(\beta) \right| .$$

The symbol $\|\cdot\|$ denotes the matrix norm: $\|(a_{ij})\| = (\sum_{i,j} a_{ij}^2)^{1/2}$.

We utilize two more assumptions to establish asymptotic normality.

A6. The first $d-1$ components of β_0 belong to the interior of a compact subset of $\mathbb{R}^{d-1}$.

A7. Let $\mathcal{N}$ denote a neighborhood of β_0.

(i) For each z in S, all mixed second partial derivatives of $\tau(z,\cdot)$ exist on $\mathcal{N}$.

(ii) There is an integrable function $\Gamma(z)$ such that for all z in S and β in $\mathcal{N}$

$$\|\nabla_2\tau(z,\beta) - \nabla_2\tau(z,\beta_0)\| \le \Gamma(z)|\beta-\beta_0| .$$

(iii) $\mathbb{E}|\nabla_1\tau(\cdot,\beta_0)|^2 < \infty$.

(iv) $\mathbb{E}|\nabla_2|\tau(\cdot,\beta_0) < \infty$.

(v) The $(d-1)\otimes(d-1)$ matrix $\mathbb{E}\nabla_2\tau(\cdot,\beta_0)$ is negative definite.

The conditions of A7 are standard regularity conditions sufficient to support an argument based on a Taylor expansion of $\tau(z,\cdot)$ about β_0. See Section 8 of Sherman (1993) for a discussion of simple sufficient conditions for satisfying A7.

THEOREM 2: *If M is deterministic and A0 through A7 hold, then*

$$\sqrt{n}(\beta_n - \beta_0) \Longrightarrow (W, 0)$$

where W has a $N(0, V^{-1}\Delta V^{-1})$ distribution with $\Delta = \mathbb{E}\nabla_1\tau(\cdot,\beta_0)[\nabla_1\tau(\cdot,\beta_0)]'$ and $2V = \mathbb{E}\nabla_2\tau(\cdot,\beta_0)$.

Theorem 2 follows directly from the arguments given in Section 5 of Sherman (1993) to prove $\sqrt{n}$-consistency and asymptotic normality of the MRC estimator.

We now present the general form of Δ and V in terms of the model primitives. Notice that

$$\tau(Z,\beta) = \int_{x'\beta<X'\beta} S(Y,x'\beta_0)\,G(dx) + \int \rho(x'\beta_0)\,G(dx)$$

where $G(\cdot)$ denotes the probability distribution of X,

$$S(y,t) = M(y) - I\!E\left[M(Y) \mid X'\beta_0 = t\right],$$

and

$$\rho(t) = I\!E[M(Y) \mid X'\beta_0 = t].$$

Let $\mathcal{X}$ denote the first $d-1$ components of the vector of regressors, and $g_0(\cdot)$ the marginal density of $X'\beta_0$. Write $\tilde{\mathcal{X}}_0$ for $I\!E(\mathcal{X} \mid X'\beta_0)$.

THEOREM 3: *If derivatives $\rho'(t)$ and $g_0'(t)$ exist, and if $I\!E|X|^2 < \infty$, then*

$$\Delta = I\!E(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)' \, S(Y, X'\beta_0)^2 \, g_0(X'\beta_0)^2 \tag{15}$$

and

$$V = I\!E(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)' \, \rho'(X'\beta_0) \, g_0(X'\beta_0). \tag{16}$$

The proof of Theorem 3 is essentially the proof of Theorem 4 in Sherman (1993), using the fact that $I\!E[S(Y,t) \mid X'\beta_0 = t] = 0$.[2]

A consistent estimator of $V^{-1}\Delta V^{-1}$ from Theorem 2 can be obtained by constructing consistent estimators of the components Δ and V using numerical derivatives as in Section 7 of Sherman (1993). Alternatively, one could apply kernel methods (nonparametric regression and density estimation) to obtain consistent estimates of the components $\tilde{\mathcal{X}}_0$, $S(Y, X'\beta_0)$, $g_0(X'\beta_0)$, and $\rho'(X'\beta_0)$, then average out to obtain consistent estimates of Δ and V.

Now consider the case $M(\cdot) = R_n(\cdot)$. For each (z_1, z_2, z_3) in $S \otimes S \otimes S$ and each β in $\mathcal{B}$ define

$$f(z_1, z_2, z_3, \beta) = \{y_1 > y_3\}\{x_1'\beta > x_2'\beta\}. \tag{17}$$

Then, for each z in S and each β in $\mathcal{B}$ define

$$\tau(z, \beta) = f(z, P, P, \beta) + f(P, z, P, \beta) + f(P, P, z, \beta) \tag{18}$$

where $f(z, P, P, \beta)$ denotes the conditional expectation of $f(\cdot, \cdot, \cdot, \beta)$ given its first argument, and so on.

If we require that assumptions A0 through A6 hold, and that assumption A7 hold for the function τ in (18), then it is easy to obtain results comparable to Theorem 2 when $M(Y_i) = R_n(Y_i)$. The only difference in the statement of the theorem is that $2V$ is replaced by $3V$. This difference is due to the fact that when $M(Y_i) = R_n(Y_i)$, the maximand in (3) expands to a U-process of order three rather than one of order two. The proof is given in Section 7.3 of Sherman (1994).

[2] We would like to thank Myoung-Jae Lee for pointing out that the factor of 2 that appears in Theorem 4 of Sherman (1993) should not be there.

Next, notice that

$$\begin{aligned}\tau(Z,\beta) &= \int_{x'\beta<X'\beta} S(Y,x'\beta_0)\,G(dx) + \int \rho(x'\beta_0)\,G(dx) \\ &+ \int\int_{x_1'\beta>x_2'\beta} \lambda(Y,x_1'\beta_0)\,G(dx_1)\,G(dx_2)\end{aligned}$$

where, as before, $G(\cdot)$ denotes the probability distribution of X,

$$S(y,t) = H(y) - I\!E\left[H(Y) \mid X'\beta_0 = t\right],$$

where $H(y) = I\!P\{y > Y\}$,

$$\rho(t) = I\!E[H(Y) \mid X'\beta_0 = t],$$

and

$$\lambda(y,t) = I\!E[\{Y > y\} \mid X'\beta_0 = t].$$

Since $I\!E[S(Y,t) \mid X'\beta_0 = t] = 0$, we may apply Theorem 3 to obtain the general form for Δ and V.

Finally, as before, either numerical derivatives or kernel methods can be used to construct consistent estimators of Δ and V.

4. Single Equation Multiple Indices Models

Let $Z = (Y,X,W)$ denote an observation from a distribution P on a set $S \subseteq I\!R \otimes I\!R^d \otimes I\!R^k$, where Y is a response variable and X and W are vectors of regressors. Consider the single equation double index model

$$Y = \Lambda(X'\beta_0, W'\gamma_0, u, \epsilon) \tag{19}$$

where β_0 is a d-dimensional vector of parameters, γ_0 is a k-dimensional vector of parameters, u and ϵ are random disturbances, and Λ is increasing in its first two arguments. One example of such a model is the bivariate choice model with partial observability discussed in Poirier (1980). In this model, one observes

$$Y = \{X'\beta_0 \geq u\}\{W'\gamma_0 \geq \epsilon\}. \tag{20}$$

That is, one only observes whether or not *both* of the conditions in (20) are satisfied. Another example is the sample selection model (see, for example, Heckman (1974)) in which observations are available only on individuals who satisfy a certain condition. Here, one observes

$$Y = (X'\beta_0 + u)\{W'\gamma_0 \geq \epsilon\}.$$

In this section, we propose a rank estimator of (β_0, γ_0) in (19) that is $\sqrt{n}$-consistent and asymptotically normal provided certain monotonicity conditions hold. The results readily generalize to cover models with any number of indices.

Let $Z_i = (Y_i, X_i, W_i)$ denote a sample of independent observations from P, and let M denote an increasing function on $\mathbb{R}$. We propose estimating (β_0, γ_0) in (19) with

$$(\beta_n, \gamma_n) = \underset{\mathcal{B}\otimes\mathcal{G}}{\operatorname{argmax}} \sum_i M(Y_i)\left[R_n(X_i'\beta) + R_n(W_i'\gamma)\right] \tag{21}$$

where $\mathcal{B}$ and $\mathcal{G}$ are appropriate subsets of $\mathbb{R}^d$ and $\mathbb{R}^k$, respectively. Note that this means that β_n and γ_n can be obtained by separate maximizations.

Suppose that assumptions A0 through A5 hold, and in addition, that A0 through A5 hold with X, X_d, $\mathcal{B}$, β_d, and β_0 replaced by W, W_k, $\mathcal{G}$, γ_k, and γ_0, respectively. If M is either deterministic or $M(Y_i) = R_n(Y_i)$, then the consistency proof presented in Section 2 generalizes in an obvious way and we get that (β_n, γ_n) consistently estimates (β_0, γ_0).

The asymptotic normality result also easily generalizes. Consider the case where M is deterministic. Expand the rank functions in (21) into sums to see that the analogue of the τ function in Section 3 is

$$\tau(z, \beta, \gamma) = f(z, P, \beta, \gamma) + f(P, z, \beta, \gamma)$$

where for each (z_1, z_2) in $S \otimes S$ and each (β, γ) in $\mathcal{B} \otimes \mathcal{G}$,

$$f(z_1, z_2, \beta, \gamma) = M(y_1)\left[\{x_1'\beta > x_2'\beta\} + \{w_1'\gamma > w_2'\gamma\}\right].$$

If we assume that the analogues of A6 and A7 hold for the τ function just defined, then the asymptotic normality proof goes through exactly as before and we get a result analogous to Theorem 2 for (β_n, γ_n). Consistent estimators of asymptotic covariance matrices and explicit forms of these matrices can be obtained as in Section 3. Similar observations hold for the case $M(Y_i) = R_n(Y_i)$.

The key conditions required to apply these results are assumption A2 and its analogue for $W'\gamma_0$. That is, we require

(i) $\mathbb{E}[M(Y) \mid X]$ is increasing in $X'\beta_0$.

(ii) $\mathbb{E}[M(Y) \mid W]$ is increasing in $W'\gamma_0$.

To see that both these conditions can be met, take $M(y) = y$ and consider the bivariate choice model with partial observability given in (20). For simplicity, assume that u and X are independent and that ϵ and W are independent. Then (i) reduces to

$$\mathbb{E}[G(X'\beta_0, W'\gamma_0) \mid X] \quad \text{is increasing in } X'\beta_0 \tag{22}$$

where

$$G(t, s) = \mathbb{P}\{u \le t, \epsilon \le s\}.$$

Note that (22) is trivially satisfied if X and W are independent, since G is increasing in each of its arguments. However, such an assumption will rarely be met in practice. Still, there do exist more realistic conditions under which conditions (i) and (ii) hold.

For example, write $X'\beta_0$ as $\mathcal{T}_0 + X_d$ and $W'\gamma_0$ as $\Sigma_0 + W_k$. Suppose that X and W are jointly normal so that given $\mathcal{T}_0 = \tau$ and $\Sigma_0 = \sigma$, the pair (X_d, W_k) has a joint normal distribution. Without loss of generality, take the conditional means and variances to be zeros and ones. Let ρ denote the conditional correlation. It follows that the conditional distribution of $(X'\beta_0, W'\gamma_0)$ given $\mathcal{T}_0 = \tau$ and $\Sigma_0 = \sigma$ is $N(\tau, \sigma, 1, 1, \rho)$. Deduce that the distribution of $W'\gamma_0$ given $X'\beta_0 = t$, $\mathcal{T}_0 = \tau$, and $\Sigma_0 = \sigma$ is $N(\mu, 1 - \rho^2)$ where $\mu = \rho(t - \tau) + \sigma$. Let $f(\cdot \mid t, \tau, \sigma)$ denote the density associated with this conditional distribution. The expectation in (22) equals

$$I\!E^{\mathcal{T}_0, \Sigma_0} \left[\int G(t, s) f(s \mid t, \tau, \sigma)\, ds \right] .$$

Let $\lambda(t, \tau, \sigma)$ denote the integral in brackets. If λ is increasing in t for each fixed pair (τ, σ), then (22) will be satisfied. For simplicity, assume that $\frac{\partial}{\partial t} G(t, s)$ exists for each s in the support of $W'\gamma_0$, and that we may differentiate under the integral sign. We have that

$$\frac{\partial}{\partial t} \lambda(t, \tau, \sigma) = \int \frac{\partial G(t, s)}{\partial t} f(s \mid t, \tau, \sigma)\, ds \; + \; c(\rho) \int G(t, s)(s - \mu) f(s \mid t, \tau, \sigma)\, ds$$

where $c(\rho) = \rho/(1 - \rho^2)$. Since $\frac{\partial}{\partial t} G(t, s) \geq 0$, the first integral is nonnegative. Turn to the second integral. If ρ is nonnegative then so is $c(\rho)$. When $s = \mu + \delta$ for $\delta \geq 0$ the contribution to the integrand is

$$G(t, \mu + \delta) \delta f(\mu + \delta \mid t, \tau, \sigma) .$$

When $s = \mu - \delta$ the contribution is

$$-G(t, \mu - \delta) \delta f(\mu - \delta \mid t, \tau, \sigma) .$$

Since $f(\cdot \mid t, \tau, \sigma)$ is symmetric about μ, the sum of these two contributions is

$$\delta f(\mu + \delta \mid t, \tau, \sigma) \left[G(t, \mu + \delta) - G(t, \mu - \delta) \right] .$$

Since G is increasing in each of its arguments, the last quantity is nonnegative. Deduce that (22) and therefore (i) holds. By symmetry, (ii) also holds.

In this last example, notice that if, conditional on $\mathcal{T}_0 = \tau$ and $\Sigma_0 = \sigma$, X_d and W_k are independent, then the conditional distribution of $W'\gamma_0$ given $X'\beta_0 = t$, $\mathcal{T}_0 = \tau$, and $\Sigma_0 = \sigma$ will not depend on t. Consequently, the second integral comprising $\frac{\partial}{\partial t} \lambda(t, \tau, \sigma)$ will be equal to zero, implying that (i) and (ii) will be satisfied. This is true more generally. That is, regardless of marginal behavior,

if, conditional on all other components, any component of X is independent of any component of W, then a slight modification of the argument above shows that both (i) and (ii) will hold.

Requiring conditional joint normality of (X_d, W_k) or the conditional independence of one component of X and one component of W may still be too restrictive to be useful in practice. Nonetheless, the hope is that assumptions (i) and (ii) are true more generally, not only for the bivariate choice model, but for other models as well. Certainly, simple visual checks of these assumptions can be made by constructing nonparametric regression estimators of the regression functions in (i) and (ii) based on the estimated indices $X'\beta_n$ and $W'\gamma_n$.

5. Estimation with Categorical Explanatory Variables

A key assumption in the consistency proof in Section 2 is condition A2 requiring that one of the explanatory variables have an everywhere positive Lebesgue density. In many important economic applications, all the explanatory variables are categorical. In this section, we establish an analogue of consistency in this case, showing that β_n must converge to a region containing β_0 at a very fast rate.

Suppose the support of X is a finite set $\{x_1, \ldots, x_k\}$ and the distribution of X assigns probability $p_i > 0$ to x_i. Subdivide the parameter space $\mathcal{B}$ into open regions $B_1, \ldots, B_m$ such that for each β in B_i the rank ordering of $\{x_1'\beta, \ldots, x_k'\beta\}$ is the same, and for β_i in B_i and β_j in B_j the rank orderings of $\{x_1'\beta_i, \ldots, x_k'\beta_i\}$ and $\{x_1'\beta_j, \ldots, x_k'\beta_j\}$ are different. Notice that for each β in B_i, the set $\{x_1'\beta, \ldots, x_k'\beta\}$ must contain k distinct points. These regions are bounded by hyperplanes of the form $\{x_i'\beta = x_j'\beta, i \neq j\}$, so the totality of regions can be determined by plotting all such hyperplanes. For example, suppose X consists of two binary variables so that the support of X is

$$\{(0,0), (0,1), (1,0), (1,1)\}.$$

If we normalize on the coefficient of the second variable being unity, then the regions are

$$\{\beta < -1\}, \quad \{-1 < \beta < 0\}, \quad \{0 < \beta < 1\}, \quad \{\beta > 1\}.$$

As the number of points in the support of X increases, the number of regions in parameter space increases rapidly. For example, if there are three binary explanatory variables, then there are 8 points in the support of X, and in principle, 28 bounding hyperplanes. In fact, many of these hyperplanes are either redundant or null, but the 12 distinct hyperplanes divide the parameter space into 52 regions.

Select a representative β_i from each B_i and define the finite parameter space $\mathcal{B}^* = \{\beta_1, \ldots, \beta_m\}$. For simplicity, assume that the true parameter value, β_0, is contained in one of the B_i.[3] Since we only use information on the rank ordering of the $x_i'\beta$ to estimate β_0, we can only identify the region containing β_0. Thus, we can rewrite the estimator equation (3) as

$$\beta_n = \underset{\mathcal{B}^*}{\operatorname{argmax}} \sum_i M(Y_i) R_n(X_i'\beta)\,. \tag{23}$$

Finally, for notational simplicity, we take β_0 as the representative of its region.

The following modified regularity condition is used in the consistency proof:

A2′. The function $H(X'\beta_0) = I\!E[M(Y) \mid X]$ is strictly increasing on its points of definition.

THEOREM 3: *If the support of X is finite, the parameter space $\mathcal{B}^*$ is defined as above, and assumptions A2′ and A5 hold, then the β_n defined in (23) satisfies $I\!P\{\beta_n \neq \beta_0\} = O(1/n)$. If, in addition, $M(\cdot)$ is bounded, then $I\!P\{\beta_n \neq \beta_0\} = O(\rho^n)$ with $0 < \rho < 1$.*

PROOF. First, for comparison with Theorem 1, we establish the consistency of β_n. As in the proof of Theorem 1, define $G(\beta) = I\!E G_n(\beta)$. As before, we will show

(i) $G(\beta)$ is uniquely maximized at β_0.

(ii) $\sup_{\mathcal{B}^*} |G_n(\beta) - G(\beta)| = o_p(1)$.

(iii) $G(\beta)$ is continuous.

Endow $\mathcal{B}^*$ with the discrete topology. Since all functions are continuous in this topology, (iii) holds trivially. Since $\mathcal{B}^*$ is finite, uniform convergence in probability (condition (ii)) follows from pointwise convergence, which in turn follows from standard results on U-statistics (Serfling, 1980, Chapter 5). To establish (i), apply equation (6) to get

$$G(\beta) = \frac{1}{2} \sum_{i \neq j} \left[H(x_i'\beta_0)\{x_i'\beta > x_j'\beta\} + H(x_j'\beta_0)\{x_i'\beta < x_j'\beta\} \right] p_i p_j\,.$$

If $\beta \neq \beta_0$, then there must exist at least one pair (x_i, x_j) in the support of $X \otimes X$ such that

$$x_i'\beta > x_j'\beta \quad \text{and} \quad x_i'\beta_0 < x_j'\beta_0\,.$$

[3] This assumption holds generically since the boundaries form a closed, nowhere dense subset of the parameter space $\mathcal{B}$. One might justify this assumption by adopting a Bayesian stance, viewing β_0 as a realization of a continuous random variable with support $\mathcal{B}$. It would then follow that β_0 would lie on a boundary with probability zero.

This pair contributes $H(x_i'\beta_0)p_ip_j$ to $G(\beta)$ and $H(x_j'\beta_0)p_ip_j$ to $G(\beta_0)$. By A2′,

$$H(x_i'\beta_0) < H(x_j'\beta_0)\,.$$

This establishes (i), proving that β_n is consistent for the region containing β_0.

To establish the rates of convergence in probability, first note that $\beta_n \neq \beta_0$ if and only if $G_n(\beta_0) < G_n(\beta)$ for some β in $\mathcal{B}^*$. Consequently,

$$I\!P\{\beta_n \neq \beta_0\} \leq \sum_{\mathcal{B}^*} I\!P\{G_n(\beta_0) < G_n(\beta)\}\,.$$

Let

$$2\epsilon = \min_{i\neq j} \frac{1}{2}|H(x_i'\beta_0) - H(x_j'\beta_0)|p_ip_j\,.$$

By A2′, $2\epsilon > 0$. Deduce from $I\!P\{X < Y\} \leq I\!P\{X < c\} + I\!P\{Y > c\}$ for any random variables X and Y and any real number c that

$$I\!P\{G_n(\beta_0) < G_n(\beta)\} \leq I\!P\{G_n(\beta_0) < G(\beta_0) - \epsilon\} + I\!P\{G_n(\beta) > G(\beta_0) - \epsilon\}\,.$$

By A5 and standard results on U-statistics, $G_n(\beta)$ is a U-statistic with mean $G(\beta)$ and variance of order $O(1/n)$. Let Kn^{-1} be a uniform bound on these variances. Each term on the right-hand side of the last inequality is bounded by $K/\epsilon^2 n$ by Chebyshev's inequality and the fact that $G(\beta) < G(\beta_0) - 2\epsilon$ for $\beta \neq \beta_0$. The result follows.

To establish the exponential bound, note that when $M(\cdot)$ is bounded, $G_n(\beta)$ is a U-statistic with a bounded kernel and hence has a finite moment generating function. By a large deviations theorem for U-statistics (see, for example, Theorem B on page 201 of Serfling (1980)), each term on the right-hand side of the last inequality has an exponential bound $C\rho^n$ with $0 < \rho < 1$. The result follows by combining these finitely many bounds. □

Remark. In practice, explanatory variables used in economic applications are often discrete but take on many values. Years of schooling, age, etc. are examples. Furthermore, even when the explanatory variables are continuously distributed, we may want to condition on these variables in computing distributional approximations since they are an exact ancillary statistic. The results of this section establish a notion of consistency in such cases. However, in practice, the asymptotic normal approximations derived in Section 3 may be useful even in these cases. The normal approximations depend on the quality of a quadratic approximation to the objective function and this can be assessed by plotting the observed objective function. This idea is pursued in the application in the next section.

6. An Application

In this section, we apply our estimation method to examine an index model formulation of a fairly standard wage equation. We consider the annual wages of a sample of prime-age white males as a function of a single linear index in education (completed years of schooling), potential job market experience (age – education – 6), experience squared, and marital status. The application illustrates (at least) three features of our rank estimators: (1) the computational efficiency of these methods is important since the sample size is fairly large (18,967) but not atypically so for large survey samples; (2) the explanatory variables we use are discrete, but fairly rich, so that the arguments of Section 6 are relevant; (3) the computational issues that arise in both optimization and estimation of the asymptotic covariance matrix of the index parameters can be handled by standard packages.

The data come from the March 1989 Current Population Survey (CPS). We restrict attention to white men, between the ages of 25 and 64, who worked full-time year-round, earned at least \$1/hour and were not self-employed. Table 1 provides some summary statistics for this data set. The index model we use is derived from a standard human-capital earnings function with the coefficient on education normalized to unity:

$$S + \beta_1 E + \beta_2 E^2 + \beta_3 MS .$$

In this expression, S denotes completed years of schooling, E potential job market experience, and MS marital status (1 if married, 0 otherwise). In the last five columns of Table 2, we give point estimates and standard errors for the β_i's using the following methods:

(i) least squares of log(earnings) on the index;

(ii) the rank estimator defined by equation (3) with $M(y) = y$ and $y =$ earnings;

(iii) the rank estimator with $M(y) = y$ and $y =$ log(earnings);

(iv) the rank estimator with $M(y) = R_n(y)$;

(v) the SLS (semiparametric least squares) estimator of Ichimura (1993).

(The first column of estimates in Table 2 are least squares estimates of unscaled slope parameters and a constant term.) Note that since ranks are invariant to any strictly increasing transformation, we can use earnings or log(earnings) or any strictly increasing function of earnings as the argument of the rank function in (iv) without changing parameter estimates.

All estimates of the scaled coefficients are close. The largest deviations are between the monotone regression estimates and the SLS estimates - approximately 20% variation in the coefficient on experience squared – but even these differences yield very small differences in the estimated index. For example, the correlation between the index based on the SLS estimates and that based on the monotone wage regression is .996.

Two different standard errors are given for each of the estimates.[4] The standard errors in parentheses are computed from the limiting distributions and related results of Theorems 2 and 3 for the rank estimators and similar results for the SLS estimator. More details on the computation of these standard errors are given below. The standard errors in square brackets are computed by randomly resampling (bootstrapping) 250 samples of size 500 from the original sample, computing parameter estimates for these subsamples and rescaling the bootstrap standard errors by $\sqrt{500/18967}$.[5] These bootstrap standard errors are in good agreement with the limiting distribution standard errors and both indicate that the monotone rank estimators perform well in terms of efficiency. The rank estimates are quite close to the scaled least squares estimates in terms of efficiency – within 5-10% depending on the coefficient and the standard error used – and outperform the Ichimura estimator by as much as 10-20%.

Computational Issues

1. Optimization

To maximize the objective function, we used an iterative application of the Nelder-Mead (N-M) simplex method. The N-M algorithm was translated from the FORTRAN version in Numerical Recipes (1992) to GAUSS. The iterative scheme was: (1) at each iteration we decreased the convergence tolerance by a fixed factor λ_1:

$$\text{ftolerance}\,(\text{iteration i}) = \lambda_1\,\text{ftolerance}\,(\text{iteration i}-1);$$

(2) the initial simplex at each iteration had as one vertex the optimizing vertex of the preceding iteration's final simplex, and the shape of the initial simplex was the same as that of the preceding iteration scaled down by λ_2:

$$\begin{aligned}\text{initial simplex}\,(\text{iteration i}) &= \text{optimal vertex}\,(\text{final simplex}\,(\text{iteration i}-1))\\ &+ \lambda_2\,\text{initial simplex}\,(\text{iteration i}-1)\,.\end{aligned}$$

Experimentation with this method indicated that it worked well. For this estimation problem, seven iterations with $\lambda_1 = .6$ and $\lambda_2 = .4$ yielded estimates that were consistently close (to three significant digits) to the final results given in Table 2 for a variety of starting simplices. Least squares estimates based on the log-wage model provided excellent starting values.[6] The iterative scheme

[4] These standard error estimates are heuristic since strictly speaking, the asymptotic normality result does not cover the case of all categorical regressors. However, below we present evidence supporting the usefulness of the normal approximation even in this case.

[5] Bootstrap standard errors computed using the percentile method based on 90% confidence intervals yield very similar estimates – all within 2% of the bootstrap sample standard deviations. Percentile methods using larger confidence intervals were less reliable, because of the relatively small number of bootstrap replications.

[6] In practice, there is usually a reasonable parametric model that will yield good parametric starting values so that fewer iterations are typically required to yield stable and reliable optima.

reduced the possibility of locating a local optimum that was not a global optimum or becoming trapped on a lower dimensional hyperplane as can sometimes occur with the N-M method.

In addition to the N-M method, we tried quasi-Newton methods with numerical derivatives. The theoretical justification for such methods is that the sample objective function converges to a smooth population objective function, so that asymptotically, the gradient and hessian of the population objective function can be computed by numerical derivatives. However, in practice, these methods were very sensitive to the choice of stepsize.

To put the computational complexity of this method in context, a single evaluation of any of the three semiparametric objective functions defined above takes about 2 seconds on a 486-33 PC with 16 Megabytes of RAM. An application of the iterative N-M method requires several thousand function evaluations, so the total time to compute point estimates is typically around 1 hour. By comparison, a single function evaluation of the objective function for the maximum rank correlation estimator (Han, 1987) takes about 500 times longer – approximately 20 minutes. We did not attempt to compute this estimator since it probably would have required about 500 hours of computer time.

The SLS estimator was also estimated using the N-M algorithm. The nonparametric regression step was carried out using Fast Fourier transform methods (see, for example, Härdle, 1992). With a fixed bandwidth choice in the nonparametric regression step, the SLS estimator took about 4 times longer to solve than the rank-based estimators. Of course, if one were to use cross-validation or local smoothing in the nonparametric step, this would increase the computational cost of the SLS estimator considerably. We also note that the SLS estimates reported here are based on nonparametric estimates using a Gaussian kernel and a fixed bandwidth proportional to $n^{-1/8}$. We also computed SLS estimates for bandwidths proportional to $n^{-1/10}$, $n^{-1/6}$ and $n^{-3/20}$. The maximum variations in the parameter estimates over these alternative bandwidth choices are roughly one standard deviation for each of the parameters – .015 for the experience coefficient, .00052 for the experience squared coefficient, and .092 for the marital status coefficient.

2. Standard Error Estimation

The theoretical calculations of Section 3 suggest two natural methods for estimating the asymptotic covariance matrices of the rank estimators. One method is based on numerical derivatives of the objective function, as discussed in Sherman (1993). The other approach is based on nonparametric estimation of the model primitives and application of Theorem 3. Numerical derivatives were numerically unstable in the sense that small changes in step size lead to large changes in the estimated covariance matrix. The standard errors in Table 2 are based on nonparametric estimates of the following quantities:

(i) $g_0(\cdot)$, the index density function;

(ii) $\tilde{\mathcal{X}}_0$, the regression of the explanatory variables on the index;

(iii) $\rho(\cdot)$, the regression of $M(Y)$ on the index;

(iv) $\rho'(\cdot)$, the derivative of the ρ function.

These nonparametric estimates were calculated using kernel methods for a variety of kernel and bandwidth choices. The kernel functions $K(x)$ used included:

(i) $K(x) = \{|x| \leq 1\}$, a square kernel;

(ii) $K(x) = \phi(x)\{|x| \leq 1\}$, a truncated normal kernel;

(iii) $K(x) \propto (5 - x^2)\{|x| \leq \sqrt{5}\}$, the Epanechnikov kernel.

The bandwidths used were $\hat{\sigma}n^{-1/3}$, $\hat{\sigma}n^{-1/4}$, $\hat{\sigma}n^{-1/5}$, and $\hat{\sigma}n^{-1/6}$ with $\hat{\sigma}$ equal to the sample standard deviation of the estimated index. Thus, a total of 12 different asymptotic covariance estimators were calculated for each model. The results given in the table correspond to the truncated gaussian kernel with bandwidth $\hat{\sigma}n^{-1/5}$. The maximum deviation for all three estimated models across the other 11 estimates was less than 10%.

Useful by-products of these calculations are the nonparametric estimates of the model primitives. In particular, the estimates of the density of the index and of the function of the data that is assumed to be monotone in the index can be particularly enlightening. Figure 1a shows an estimate of the index density for the monotone log-wage regression model based on the truncated normal kernel and bandwidth $\hat{\sigma}n^{-1/5}$. Figure 1b shows a locally smoothed version of the same density, using local bandwidths $\hat{\sigma}n^{-1/5}/\hat{f}(x)^{1/2}$, where $\hat{f}(x)$ is the density shown in Figure 1a. (See Silverman, 1986, Chapter 5 for more on adaptive methods.) These plots reveal that the index density is reasonably well behaved at this scale even though the components of the index are categorical. We elaborate on this point below. Figures 2a and 2b show estimates of the regression of log-wage on the index with and without local smoothing. Figure 2a shows substantial non-monotonicity in the lower tail of the estimated index distribution, but this non-monotonicity effectively disappears when local smoothing is applied. These non-monotonicities arise from the extremely small number of points in this tail, as revealed by Figures 1a and 1b. Figure 2b provides qualitative support for the underlying assumption of monotonicity used to estimate this model. However, a small non-monotonicity persisits in the upper tail of the index distribution.

As mentioned before, the explanatory variables in this model are all categorical variables. There are 1133 distinct cells, (or combinations) of the education, experience and marital status variables in this sample. Hence, the assumptions underlying the asymptotic normality result of Section 3 are problematic. However, the key qualitative feature underlying the proof of asymptotic normality – that the objective function is approximately quadratic at the $1/\sqrt{n}$ scale – may hold approximately with this rich set of categorical explanatory variables.

To illustrate this point, we have plotted one dimensional slices of the objective function through the maximizing point for the monotone log-wage regression model. Each slice varies one of the parameters over a range of ± 5 estimated standard deviations around the maximizing value and holds the other parameters fixed at their maximizing values. These three slices are shown in Figure 3, with Figures 3a, 3b, and 3c corresponding to the parameter estimates associated with the variables E, E^2, and MS, respectively. The picture that emerges is broadly consistent with the argument that the objective function is approximately quadratic at this scale. The least smooth slice corresponds to the marital status variable – the least smooth of the explanatory variables, but even this slice exhibits an approximate quadratic shape.

To further assess the asymptotic normal approximation in this model, we computed 250 bootstrap samples of size 18,967. In Figures 4a, 4b, and 4c, we show QQ plots for the parameters based on these bootstrap samples compared to standard normal variables. All these plots are close to linear and as such are consistent with the approximate normality of the estimates. Some small bumpiness in the QQ plots might be explained by the discrete nature of the explanatory variables or from the small Monte Carlo sample size, but the usefulness of the normal approximation does not seem to be in doubt.

Finally, as mentioned above, nonparametric regression of the log-wage on the estimated index exhibits a small non-monotonicity at the upper tail of the index. These index values correspond to high values of the education and experience measures. This suggests that the corresponding interaction term might belong in the model. In Table 3, we give parameter estimates and standard errors for this model for all the semiparametric estimators considered previously. These estimates support the inclusion of the interaction term. In addition, Figure 5 shows cubic smoothing spline estimates (see, for example, Green and Silverman, 1994) of the nonparametric regression of log-wages on indices with and without interaction. Figure 5a plots the spline regression of log-wage against the index without interaction using cross-validation to choose the smoothing parameter. Figure 5b plots the spline regression against the index with interaction using the same smoothing parameter as in Figure 5a. These figures show that inclusion of the interaction term eliminates the small non-monotonicity appearing in the upper tail of the index of the model without interaction. This qualitative feature persists over a wide range of smoothing parameter choices. Taken together, these facts lend strong support to the adequacy of the model with interaction.

7. Summary and Remarks

This paper presents a new class of rank estimators of scaled coefficients in semiparametric monotonic linear index models. These estimators exploit monotonicity between a response variable and an associated index in a natural way and may allow flexibility in balancing robustness and efficiency objectives.

Further, they are $\sqrt{n}$-consistent and asymptotically normal, computationally attractive, and require no subjective bandwidth choices as are needed for kernel-based competitors.

REMARK 1. It should be noted that the kernel-based SLS estimator of Ichimura (1993), for example, does not require the monotonicity assumption A1 to be consistent. As such it covers some models not covered by the rank estimators. However, the SLS estimator (like other kernel-based estimators) requires very smooth index and error distributions to control asymptotic bias in order to achieve $\sqrt{n}$-consistency. It also requires trimming and can be more variable and more computationally intensive than the rank estimators even when the Fast Fourier Transform is used without local smoothing (see Section 6).

REMARK 2. We have no specific suggestions on how best to choose the function $M(\cdot)$ in any particular application. This is an interesting and difficult question requiring consideration on a case by case basis. However, we can offer the following general guidelines: If a practitioner does not know the nature of the monotonic relationship between the response variable and the associated index but is comfortable assuming that the error distribution has finite variance, then greater efficiency may be achieved by using a specification like $M(y) = y$ rather than $M(y) = R_n(y)$ since more information on the response is used with the former specification. If the practitioner is not comfortable making the finite variance assumption, then $M(y) = R_n(y)$ is a safe, robust choice. Winsorizing is also robust and may be more efficient than the rank specification, though choice of the winsorizing constants may be crucial to the performance of the estimator in smaller samples.

REMARK 3. Consider the application presented in Section 6. Prior to introducing the interaction term between education and experience one could question whether the relationship between wages and the original index was indeed monotonic, at least in the upper tail of the index distribution. Even after adding the interaction term, one could, in theory, still entertain doubts. As strong as the evidence is in favor of a monotonic relationship between wages and the index with interaction, it is still true that a monotonic picture (Figure 5b) need not imply a monotonic relationship. One can formally test the monotonic specification. For example, suppose the single linear index model holds and that all the conditions required by the SLS procedure hold. Then, as mentioned above, the SLS estimator will be consistent even if monotonicity fails. Assume further that all the conditions of rank estimation hold, except possibly the monotonicity condition A1. Under these conditions, a generalized Hausman-type test for monotonicity can be constructed using the SLS estimates and the rank estimates (see Amemiya 1985, p.145).[7] Similar tests of monotonicity can be developed based

[7]We thank an anonymous referee for suggesting this test.

on other estimators that remain consistent when monotonicity fails. Examples include the estimator of Klein and Spady (1993) for the binary choice model, and various weighted average derivative estimators.

REFERENCES

Aho, A. V., Hopcroft, J. E., and Ullman, J. D. (1976): *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Mass.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard Univ. Press, Cambridge, Mass.

Green, P. J. and Silverman, B. W. (1994): *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London, UK.

Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316.

Heckman, J. J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–693.

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted Least Squares Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

Kendall, M. G. (1938): "A New Measure of Rank Correlation," *Biometrica*, 30, 81–93.

Klein, R. W. and Spady, R. H. (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61, 387–421.

Lehmann, E. L. (1975): *Nonparametrics: Statistical Methods Based on Ranks*. California: Holden Day.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators." *Econometrica*, 57, 1027–1057.

Poirier, D. J. (1980): "Partial Observability in Bivariate Probit Models," *Journal of Econometrics*, 12, 209–217.

Press, W. P., Flannery, B. P., Teulkolsky, S. A., and Vetterling, W. T. (1992): *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Serfling, R. J. (1980): *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

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

———- (1994): "Maximal Inequalities for Degenerate U-processes with Applications to Optimization Estimators," *Annals of Statistics*, 22, 439–459.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Spearman, C. (1904): "The Proof and Measurement of Association between Two Things," *Am. J. Psychol.*, 15, 72–101.