Econometric Theory, **24**, 2008, 795–807. Printed in the United States of America.
doi: 10.1017/S0266466608080328

A PERMUTATION-BASED ESTIMATOR FOR MONOTONE INDEX MODELS

DEBOPAM BHATTACHARYA
Dartmouth College

This paper shows that the finite-dimensional parameters of a monotone-index model can be estimated by minimizing an objective function based on sorting the data. The key observation guiding this procedure is that the sum of distances between pairs of adjacent observations is minimized (over all possible permutations) when the observations are sorted by their values. The resulting estimator is a generalization of Cavanagh and Sherman's monotone rank estimator (MRE) (Cavanagh and Sherman, 1998, *Journal of Econometrics* 84, 351–381) and does not require a bandwidth choice. The estimator is $\sqrt{n}$-consistent and asymptotically normal with a consistently estimable covariance matrix. This least-squares estimator can also be used to estimate monotone-index panel data models. A Monte Carlo study is presented where the proposed estimator is seen to dominate the MRE in terms of mean-squared error and mean absolute deviation.

1. INTRODUCTION

This paper concerns the monotone-index model where the dependent variable, y, is an unknown function of the index, $x'\beta_0$, and an unobserved error term, ε, satisfying

$$\mathrm{E}(y|x) = g(x'\beta_0),$$

where $g(.)$ is monotone. It is shown here that β_0 can be estimated up to location and scale by minimizing an objective function that equals the sum of distances between an ordered set of observations. The resulting estimator, called here the monotone permutations estimator (MPE), does not require a subjective bandwidth choice and is $\sqrt{n}$-consistent and asymptotically normal with a consistently estimable covariance matrix. The MPE is shown to be a generalization of the monotone rank estimator (MRE) of Cavanagh and Sherman

I am grateful to Bo Honore, Yuichi Kitamura, Peter Phillips, and two anonymous referees for comments and suggestions. All errors are mine. Address correspondence to Debopam Bhattacharya, Department of Economics, 301 Rockefeller Hall, Dartmouth College, Hanover, NH 03755, USA; e-mail: debopam.bhattacharya@dartmouth.edu.

© 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

(1998) that appears to utilize more information than the MRE does. It should be clarified at the outset that the main purpose of this paper is to provide this generalization and not to derive a necessarily "better" estimator than the MRE. However, in a simple Monte Carlo study reported later, the proposed estimator is seen to dominate the MRE in terms of root-mean-squared error (RMSE) and mean absolute deviations (MAD).

Section 2 outlines the model, discusses the motivation behind the estimation procedure, describes the estimation technique in detail, and compares the estimator with the MRE. Section 3 contains a lemma and the theorem for consistency. Section 4 discusses asymptotic normality. Section 5 presents a small Monte Carlo study and concludes with directions for future research. The proof of consistency appears in the Appendix.

2. MODEL AND ESTIMATION PROCEDURE

The model is given by

$$y = h(x'\beta_0, \varepsilon) \tag{1}$$

with the parameter of interest β_0 and the function $h(.,.)$ unknown. It is assumed that $h(.,.)$ is such that $\mathrm{E}(y|x) = g(x'\beta_0)$, with $g(.)$ nonconstant and weakly increasing. Note that for $g(.)$ to be weakly increasing, it is sufficient but not necessary that $h(.)$ is strictly increasing in its first argument and so applies to binary response models. It also includes, as special cases, most common parametric models, including those for mean regression and models with censored and truncated dependent variables. The problem is to estimate β_0 (up to scale and location normalization), given a sample of n independent and identically distributed (i.i.d.) observations (y_i, x_i).

I now describe the MPE, provide an intuition for why it "works," and compare it with the MRE.

Let l, u denote the minimum and maximum possible values of y. If y is not of bounded support, we can transform y to $\Phi(y)$ where $\Phi(.)$ denotes the normal or any other known cumulative distribution function (c.d.f.) and the transformed model

$$z \equiv \Phi(y) = \Phi(h(x'\beta, \varepsilon))$$

still has the same structure as (1). Then $l = 0$ and $u = 1$.[1]

First choose a β from the parameter space. Then pick m observations without replacement from the original n observations where $m \geq 2$ (m will stay fixed as n tends to infinity for the asymptotics of the estimator). There will be $\binom{n}{m}$ possible m-tuples. For each such m-tuple, order the m observations by the values of $x'\beta$ and compute the sum of squared differences between the adja-

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

cent y's with l, u subtracted, respectively, from the first and last (ordered) observation. Average across the $\binom{n}{m}$ m-tuplets and choose the β that minimizes the final sum. Formally, the MPE is defined as

$$\hat{\beta} = \arg\min_{\beta \in B} Q_n(\beta)$$

$$= \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < i_2 < \cdots < i_m \le n} q_m(y_{i_1}, y_{i_2}, \ldots, y_{i_m}, x_{i_1}, x_{i_2}, \ldots, x_{i_m}; \beta),$$

where

$$q_m(y_{i_1}, y_{i_2}, \ldots, y_{i_m}, x_{i_1}, x_{i_2}, \ldots, x_{i_m}; \beta)$$

$$= \sum_{\substack{\{j_1, \ldots, j_m\} \\ \in \wp\{i_1, \ldots, i_m\}}} \left[\left(\sum_{k=1}^{m-1} (y_{j_k} - y_{j_{k+1}})^2 + (y_{j_1} - l)^2 + (y_{j_m} - u)^2 \right) \times 1(x'_{j_1}\beta < x'_{j_2}\beta < \cdots < x'_{j_m}\beta) \right]$$

and $\wp\{i_1, \ldots, i_m\}$ denotes the set of permutations of the integers $\{i_1, \ldots, i_m\}$.

The key observation that motivates the MPE is that the sum of distances between pairs of adjacent observations is minimized (over all possible permutations) when the observations are sorted by their values. Formally, if $a < y_1 < y_2 < y_3 < \cdots < y_m < b$ are m real numbers and $y_{i_1}, y_{i_2}, y_{i_3}, \ldots, y_{i_m}$ is any permutation of the y's, then

$$(y_1 - a)^2 + \sum_{j=2}^{m} (y_j - y_{j-1})^2 + (y_m - b)^2$$

$$\le (y_{i_1} - a)^2 + \sum_{j=2}^{m} (y_{i_k} - y_{i_{k-1}})^2 + (y_{i_m} - b)^2.$$

The intuition behind our estimation procedure is as follows. When one has the true β_0, sorting (arranging in ascending order) the data by $x'\beta_0$ is "like" (i.e., but for the random error) sorting the data by the true y's. But when one sorts by $x'\beta$ for $\beta \neq \beta_0$, one produces a permutation of the true ordering of the y's. Therefore, the sum of distances (measured by the sum of squared differences) between the adjacent y's is likely to be the smallest when one has sorted by the true $x'\beta_0$. Therefore, minimizing this (ordered) sum of squares with respect to β will yield the true β.

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

2.1. Relation with MRE

The MPE can be viewed as a generalization of the MRE. Setting $u = 1$ and $l = 0$ without loss of generality and expanding the squares in $q_m(.)$ from earlier in this section to

$$q_m(y_{i_1}, y_{i_2}, \ldots, y_{i_m}, x_{i_1}, x_{i_2}, \ldots, x_{i_m}; \beta)$$
$$= \sum_{\substack{\{j_1,\ldots,j_m\} \\ \in \wp\{i_1,\ldots,i_m\}}} \left[\left(\sum_{k=1}^{m-1} y_{j_k} y_{j_{k+1}} + y_{j_m} \right) 1(x'_{j_1}\beta < x'_{j_2}\beta < \cdots < x'_{j_m}\beta) \right]$$
$$+ \text{terms not involving } \beta.$$

Thus, for $m = 2$, the objective function is equivalent to

$$-\frac{2}{n(n-1)} \sum_{i \neq j} \{(y_i y_j + y_j) 1(x'_i\beta < x'_j\beta) + (y_j y_i + y_i) 1(x'_j\beta < x'_i\beta)\}$$
$$\simeq -\frac{2}{n(n-1)} \sum_{i \neq j} y_i 1(x'_j\beta < x'_i\beta) + \text{terms not dependent on } \beta,$$

which is the negative of the Cavanagh and Sherman (1998) objective function with $M(y) = y$.[2] Thus the MPE can be viewed as a generalization of the MRE, where one uses larger subsets of observations than pairs and uses not just the largest (i.e., y corresponding to the largest $x'\beta$) of them as in the MRE but products of the successive pairs of largest observations for each choice of a subset (i.e., each choice of m).

2.2. Panel Data

Consider a panel data model where the data are available for at least two periods for each individual. The model is given by

$$y_{it} = h(\alpha_i, x'_{it}\beta_0, \varepsilon_{it}), \tag{2}$$

where $h(.,.,.)$ is increasing and nonconstant in the second argument; ε_{it} is independent of x_{it} for all i, t; given i, t ranges from 1 to $T_i \geq 2$; i ranges from 1 to n; and α_i is a fixed effect. If $\mathrm{E}(y_{it}|x_{it}, \alpha_i) = \mathrm{E}(y_{it}|x'_{it}\beta_0, \alpha_i)$ is monotone in $x'_{it}\beta$, then the idea of the MPE estimator can be used here too. For any β in the parameter space (suitably normalized for identification), for every individual i, sort the T_i observations according to $x'_{it}\beta$. Compute the sum of squared differences between the y_{it}'s for $t = 1, \ldots, T_i$ for each i. Average this sum over $i = 1, \ldots, n$ and choose β to minimize this average. The continuity and uniform convergence results for this estimate are completely analogous to the single-

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

cross-section case described below. But, unlike the cross-section case, the objective function in the panel case is not a U-process and resembles the objective function of Manski's maximum score estimator. Therefore it is no longer $\sqrt{n}$-consistent, and one may replace the indicator functions in the objective function by smooth alternatives, along the lines of Horowitz's (1992) smoothed maximum score estimator, to get arbitrarily close to $\sqrt{n}$-consistency.

3. CONSISTENCY

From now on, I shall work only for the estimator computed with $m = 3$, which is the smallest value of m for which the MPE is different from the MRE. The proof for a general m is completely analogous but notationally messier.

Assumptions. The following assumptions will be used in the proof of consistency:

(A0) (y_i, x_i) are i.i.d., $\mathrm{E}(y|x)$ depends on x only through the index $x'\beta_0$: $\mathrm{E}(y|x) = g(x'\beta_0)$ with $g(.)$ weakly increasing.
(A1) The support of x is not contained in a proper linear subspace of $\mathbb{R}^d$.
(A2) The dth component of x has an everywhere positive Lebesgue density, conditional on the other components.
(A3) The parameter space B is a compact subset of $\{\beta \in \mathbb{R}^d : \beta_d = 1\}$.
(A4) $\mathrm{E}y^2 < \infty$.

To prove consistency of the estimator, I shall use the following lemma.

LEMMA 1. *Consider three observations* (x_1, y_1), (x_2, y_2), (x_3, y_3) *from the preceding model satisfying A0–A4. Let* $f(x_1, x_2, x_3)$ *be any nonnegative (with probability one) function defined on the support of* $x = (x_1, x_2, x_3)$. *Then for all* $i \neq j \neq k \in \{1, 2, 3\}$,

$$\begin{aligned}
&\mathrm{E}\{f(x_1, x_2, x_3)1(x_1'\beta_0 < x_2'\beta_0 < x_3'\beta_0) \\
&\quad \times ((y_1 - l)^2 + (y_2 - y_1)^2 + (y_3 - y_2)^2 + (y_3 - u)^2)\} \\
&\quad \leq \mathrm{E}\{f(x_1, x_2, x_3)1(x_1'\beta_0 < x_2'\beta_0 < x_3'\beta_0) \\
&\qquad \times ((y_i - l)^2 + (y_j - y_i)^2 + (y_k - y_j)^2 + (y_k - u)^2)\}.
\end{aligned}$$

Lemma 1 essentially translates the motivating observation of Section 2.1 for real numbers into the analogous inequality for (conditional expectations of) random variables. This lemma will be used repeatedly in proving the main consistency theorem.

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

THEOREM 1. *Under Assumptions A0–A4,*

$$\hat{\beta} - \beta_0 = o_p(1).$$

Proof of consistency is done through the following steps:

1. (a) $\mathrm{E}(Q_n(\beta)) = \mathrm{E}(q_3(\beta))$ is minimized at $\beta = \beta_0$ and (b) this minimum is unique;
2. $\sup_{\beta \in B} \| Q_n(\beta) - \mathrm{E}(Q_n(\beta)) \| \to 0$ in probability;
3. $\mathrm{E}(q_3(\beta))$ is continuous in β.

The proofs of step 2 and step 1(b) are analogous to Cavanagh and Sherman (1998). The proof of the minimization step 1(a) is substantially different and is worked out in detail in the Appendix, which also includes a proof of step 3. The minimization proof is different from Cavanagh and Sherman simply because the objective function is different unless $m = 2$.

4. ASYMPTOTIC NORMALITY

The proof of asymptotic normality is analogous to the proof of asymptotic normality of the maximum rank correlation estimator, proposed by Sherman (1993). Instead of repeating the entire proof therefore, I shall only point out the modifications to that proof that the MPE warrants. The details can be obtained from relevant results in Sherman (1993, 1994).

First note that the dth component of the parameter and its estimate are 1. Let $\beta = \beta(\gamma) = (\gamma, 1)$. Define $Z = (Z_1, Z_2, Z_3) \equiv \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ and

$$\tau(z, \gamma) = \mathrm{E}_{12} q_3(.,.,z, \beta(\gamma)) + \mathrm{E}_{13} q_3(., z, ., \beta(\gamma)) + \mathrm{E}_{23} q_3(z, ., ., \beta(\gamma)),$$

where $\mathrm{E}_{13} q_3(., z, ., \beta)$ denotes the expectation of $q_3(Z_1, Z_2, Z_3, \beta)$ taken with respect to the first and third arguments, keeping the second argument at z, etc. The proof of asymptotic normality works by showing that the objective function is asymptotically equivalent to an empirical process, indexed by γ with kernel $\tau(., \gamma)$. Minimization of this empirical process with respect to γ yields the result. The asymptotic normality result is that under appropriate regularity conditions (bounded moments and differentiability),

$$\sqrt{n}(\hat{\beta} - \beta_0) \Rightarrow_d (W, 0),$$

where W follows $N(0, \Omega)$ with

$$\Omega = V^{-1} \Delta V^{-1},$$

$$V = \frac{1}{3} \mathrm{E} \nabla_2 \tau(., \beta(\gamma_0)), \qquad \Delta = \mathrm{E}(\nabla_1 \tau(., \beta(\gamma_0)) \nabla_1 \tau(., \beta(\gamma_0))').$$

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

For a general m, the asymptotic variance will be

$$\Omega = V^{-1}\Delta V^{-1},$$

$$V = \frac{1}{m}\mathrm{E}\nabla_2\tau(.,\beta(\gamma_0)), \qquad \Delta = \mathrm{E}(\nabla_1\tau(.,\beta(\gamma_0))\nabla_1\tau(.,\beta(\gamma_0))'),$$

where

$$\tau(z,\gamma) = \sum_{r=1}^{m}\mathrm{E}_{-r}q_m(\ldots,z,\ldots;\beta(\gamma))$$

and $\mathrm{E}_{-r}q_m(\ldots,z,\ldots;\beta(\gamma))$ denotes the expectation of $q_m(\ldots,.,\ldots;\beta(\gamma))$ with respect to all arguments except the rth one and the rth argument is z.

Exactly as in Cavanagh and Sherman (1998), one can consistently estimate V and Δ by numeric derivatives.

5. MONTE CARLO STUDY

The relative performances of the MRE and MPE are now compared in a small Monte Carlo study. Because the objective function for this problem is both nonsmooth and highly nonconvex, estimation with large-dimensional covariates will be time consuming. Because the purpose of the Monte Carlo study is only expository, I choose a design with only two covariates so that there is just one parameter to estimate.

The model is

$$y = e^{x_1+2x_2} + \xi(x_1,x_2)\varepsilon,$$

and ε, x_1 and x_2 are generated as standard normal variates independent of each other. The $\xi(.,.)$ function was set equal to 1 for the homoskedastic design and $|0.5x_1 + x_2|$ for the heteroskedastic design. Five hundred replications were run for each design, and sample sizes considered are 50, 100, and 200. The single parameter estimated is the coefficient on x_2 whose true value is 2. Means, medians, RMSE, and MAD across the replications are reported in Table 1. Optimization was done using the Nelder–Mead algorithm using a randomly chosen initial value (in (0,1)) and using an initial size 10 of the simplex side. The IMSL routine UMPOL written in FORTRAN 77 was used in these exercises. There was no problem of convergence in any of the exercises, and varying initial values did not alter the point to which the optimization converged.

The results are presented in Table 1, where MRE denotes the Cavanagh and Sherman (1998) estimator and corresponds to the MPE estimator with $m = 2$ and MPE3 and MPE4 are the estimators developed in the present paper and correspond to $m = 3$ and $m = 4$, respectively. It can be seen that increasing m generally improves the performance and MPE4 dominates MRE uniformly and by a significant amount. This dominance is more pronounced when the sample size is small. For $n = 50$, the RMSE falls by about 44% as m moves from 2 to

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

Table 1. Simulation Results

	Mean	Median	RMSE	MAD
a. Homoskedastic case				
$n = 50$				
MPE4	2.076	1.998	0.155	0.138
MPE3	2.048	2.030	0.261	0.142
MRE	2.119	2.010	0.283	0.198
$n = 100$				
MPE4	1.999	1.987	0.077	0.071
MPE3	2.029	1.988	0.106	0.085
MRE	2.005	1.993	0.109	0.103
$n = 200$				
MPE4	1.996	1.986	0.054	0.0466
MPE3	2.099	2.010	0.074	0.066
MRE	1.999	2.007	0.083	0.068
b. Heteroskedastic case				
$n = 50$				
MPE4	2.019	2.018	0.128	0.108
MPE3	2.057	2.006	0.231	0.147
MRE	2.08	2.04	0.232	0.146
$n = 100$				
MPE4	2.007	1.991	0.067	0.062
MPE3	2.003	1.995	0.103	0.068
MRE	2.000	1.992	0.089	0.071
$n = 200$				
MPE4	2.025	2.017	0.054	0.046
MPE3	2.098	2.017	0.056	0.048
MRE	1.987	1.984	0.057	0.051

4. For the homoskedastic case, even for a sample of size 200 (which is large relative to the number of parameters which is 1), MPE4 does significantly better than the MRE. Moreover, the RMSE and mean absolute deviation for the MPE3 lie between those of MRE and MPE4 almost everywhere. This seems to confirm the intuition that for higher m the MPE utilizes more information (e.g., uses triplets rather than pairs) and so is potentially more efficient. Only for the heteroskedastic case with a large n, MPE4 and MRE seem to have somewhat similar performance.

However, the computational burden increases nontrivially as m rises. This is both because the number of distinct m-tuples $\binom{n}{m}$ will increase with m when m is small and also, more importantly, because calculation of the MPE involves sorting each of these distinct m-tuples for each choice of the parameter vector.

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

For all sample sizes considered in the Monte Carlo study, calculating MRE took about 1 second on a 585 MHz processor with 632 megabytes of RAM. The MRE4 took 1 second for sample size 50, 3 seconds for sample size equal to 100, and about 1 minute for sample size equal to 200.

The two obvious remaining questions are (i) whether the improvement in performance continues as one keeps increasing m and (ii) how m should be chosen in practice. A theoretical comparison of asymptotic (and finite-sample) RMSE and MAD across the different values of m is hard, tedious, and outside the scope of this short paper. Further analysis of these issues is therefore left to future research. The Monte Carlo results reported in the present paper suggest that the generalized estimator can potentially offer significant efficiency gains, particularly in small samples.

NOTES

1. If Y is bounded with probability one, then our results depend only on the property that $\mathrm{E}(Y|X)$ is an increasing function of $X'\beta_0$. It is not necessary that the entire distribution of $(Y|X)$ should depend on X through $X'\beta_0$. If, however, Y cannot be assumed to be bounded, then we need this transformation and require that $\mathrm{E}(\Phi(Y)|X)$ depends on X through $X'\beta$ and is monotonic in $X'\beta$.

2. For $m = 3$, the objective function is equivalent to $(1/n(n-1)(n-2))\sum_{k\neq i\neq j} Q_3(x_i, y_i, x_j, y_j, x_k, y_k)$, where $-Q_3(x_i, y_i, x_j, y_j, x_k, y_k) = (y_j y_i + y_k y_j + y_k)1(x_i'\beta < x_j'\beta < x_k'\beta) + (y_k y_i + y_k y_j + y_j)1(x_i'\beta < x_k'\beta < x_j'\beta) + (y_j y_i + y_k y_i + y_k)1(x_j'\beta < x_i'\beta < x_k'\beta) + (y_j y_k + y_k y_i + y_i)1(x_j'\beta < x_k'\beta < x_i'\beta) + (y_j y_k + y_i y_j + y_i)1(x_k'\beta < x_j'\beta < x_i'\beta) + (y_k y_i + y_i y_j + y_j)1(x_k'\beta < x_i'\beta < x_j'\beta)$.

REFERENCES

Cavanagh, C. & R.P. Sherman (1998) Rank estimators for monotonic index models. *Journal of Econometrics* 84, 351–381.

Horowitz, J.L. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–531.

Sherman, R.P. (1993) The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61, 123–137.

Sherman, R.P. (1994) Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics* 22, 439–459.

APPENDIX

Proof of Lemma 1. Let $Z = 1(x_1'\beta_0 < x_2'\beta_0 < x_3'\beta_0)$ and consider without loss of generality the permutation $\{2,1,3\}$. Then we have

$$\begin{aligned}
&\mathrm{E}(fZ((y_2 - l)^2 + (y_2 - y_1)^2 + (y_3 - y_1)^2 + (y_3 - u)^2)) \\
&\quad - \mathrm{E}(fZ((y_1 - l)^2 + (y_2 - y_1)^2 + (y_3 - y_2)^2 + (y_3 - u)^2)) \\
&= \mathrm{E}[fZ\{(y_2 - l)^2 + (y_3 - y_1)^2 - (y_1 - l)^2 - (y_3 - y_2)^2\}].
\end{aligned}$$

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

The term inside {.} equals

$$L = -2y_2 l - 2y_1 y_3 + 2y_1 l + 2y_2 y_3.$$

Taking expectations, conditional on x_1, x_2, x_3

$$\begin{aligned}
&\mathrm{E}(L|x_1, x_2, x_3) \\
&\quad = -2g(x_2'\beta_0)l - 2g(x_1'\beta_0)g(x_3'\beta_0) + 2g(x_1'\beta_0)l + 2g(x_2'\beta_0)g(x_3'\beta_0) \\
&\quad = -2g(x_2'\beta_0)(l - g(x_3'\beta_0)) + 2g(x_1'\beta_0)(l - g(x_3'\beta_0)) \\
&\quad = 2\{g(x_3'\beta_0) - l\}\{g(x_2'\beta_0) - g(x_1'\beta_0)\}.
\end{aligned}$$

Note that when $Z = 1$, we have $x_1'\beta_0 < x_2'\beta_0 < x_3'\beta_0$. Given that $g(.)$ is monotone increasing and that $l < g(x_i'\beta_0)$ for $i = 1, 2, 3$ with probability one in x, we have

$$\{g(x_3'\beta_0) - l\}\{g(x_2'\beta_0) - g(x_1'\beta_0)\} \geq 0 \quad \text{w.p. 1.}$$

Therefore,

$$\begin{aligned}
&\mathrm{E}(fZ((y_2 - l)^2 + (y_2 - y_1)^2 + (y_3 - y_1)^2 + (y_3 - u)^2)) \\
&\quad \geq \mathrm{E}(fZ((y_1 - l)^2 + (y_2 - y_1)^2 + (y_3 - y_2)^2 + (y_3 - u)^2)).
\end{aligned}$$

■

Proof of Theorem 1.
Step 1. Minimization.

$$\begin{aligned}
\mathrm{E}q_3(\beta) = {} & \mathrm{E}([(y_i - l)^2 + (y_j - y_i)^2 + (y_k - y_j)^2 + (y_k - u)^2]1(x_i'\beta < x_j'\beta < x_k'\beta)) \\
& + \mathrm{E}([(y_i - l)^2 + (y_k - y_i)^2 + (y_j - y_k)^2 + (y_j - u)^2]1(x_i'\beta < x_k'\beta < x_j'\beta)) \\
& + \mathrm{E}([(y_j - l)^2 + (y_i - y_j)^2 + (y_k - y_i)^2 + (y_k - u)^2]1(x_j'\beta < x_i'\beta < x_k'\beta)) \\
& + \mathrm{E}([(y_j - l)^2 + (y_k - y_j)^2 + (y_i - y_k)^2 + (y_i - u)^2]1(x_j'\beta < x_k'\beta < x_i'\beta)) \\
& + \mathrm{E}([(y_k - l)^2 + (y_i - y_k)^2 + (y_j - y_i)^2 + (y_j - u)^2]1(x_k'\beta < x_i'\beta < x_j'\beta)) \\
& + \mathrm{E}([(y_k - l)^2 + (y_j - y_k)^2 + (y_i - y_k)^2 + (y_i - u)^2]1(x_k'\beta < x_j'\beta < x_i'\beta)).
\end{aligned}$$

Now write the first term as the sum of six terms, each term corresponding to one ordering of $(x_1'\beta_0, x_2'\beta_0, x_3'\beta_0)$:

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

$$
\begin{aligned}
T_{1.} &= \mathrm{E}\big(((y_1-l)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_1'\beta_0<x_2'\beta_0<x_3'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_1-l)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_1'\beta_0<x_3'\beta_0<x_2'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_1-l)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_2'\beta_0<x_1'\beta_0<x_3'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_1-l)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_2'\beta_0<x_3'\beta_0<x_1'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_1-l)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_3'\beta_0<x_1'\beta_0<x_2'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_1-l)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_3'\beta_0<x_2'\beta_0<x_1'\beta_0)\big)\\
&= T_{11}+T_{12}+T_{13}+T_{14}+T_{15}+T_{16},\quad \text{say.}
\end{aligned}
$$

Let

$$
\begin{aligned}
S_{1.} &= \mathrm{E}\big(((y_1-l)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_1'\beta_0<x_2'\beta_0<x_3'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_1-l)^2+(y_3-y_1)^2+(y_2-y_3)^2+(y_2-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_1'\beta_0<x_3'\beta_0<x_2'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_2-l)^2+(y_1-y_2)^2+(y_1-y_3)^2+(y_3-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_2'\beta_0<x_1'\beta_0<x_3'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_2-l)^2+(y_3-y_2)^2+(y_1-y_3)^2+(y_1-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_2'\beta_0<x_3'\beta_0<x_1'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_3-l)^2+(y_1-y_3)^2+(y_2-y_1)^2+(y_2-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_3'\beta_0<x_1'\beta_0<x_2'\beta_0)\big)\\
&\quad+\mathrm{E}\big(((y_3-l)^2+(y_3-y_2)^2+(y_2-y_1)^2+(y_1-u)^2)1(x_1'\beta<x_2'\beta<x_3'\beta)\\
&\qquad\times 1(x_3'\beta_0<x_2'\beta_0<x_1'\beta_0)\big)\\
&= S_{11}+S_{12}+S_{13}+S_{14}+S_{15}+S_{16},\quad \text{say.}
\end{aligned}
$$

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

Lemma 1 with $f(x_1, x_2, x_3) = 1(x_1'\beta < x_2'\beta < x_3'\beta)$ implies

$$T_{1j} \geq S_{1j} \quad \text{for } j = 1, \ldots, 6 \tag{A.1}$$

so that $T_{1.} \geq S_{1.}$.

Similarly,

$$\begin{aligned}
T_{2.} &= \mathrm{E}([(y_1 - l)^2 + (y_3 - y_1)^2 + (y_2 - y_3)^2 + (y_2 - u)^2]1(x_1'\beta < x_3'\beta < x_2'\beta)) \\
&\geq S_{2.} \\
&= \mathrm{E}\big(((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2)1(x_1'\beta < x_3'\beta < x_2'\beta) \\
&\qquad \times 1(x_1'\beta_0 < x_2'\beta_0 < x_3'\beta_0)\big) \\
&\quad + \mathrm{E}\big(((y_1 - l)^2 + (y_3 - y_1)^2 + (y_2 - y_3)^2 + (y_2 - u)^2)1(x_1'\beta < x_3'\beta < x_2'\beta) \\
&\qquad \times 1(x_1'\beta_0 < x_3'\beta_0 < x_2'\beta_0)\big) \\
&\quad + \mathrm{E}\big(((y_2 - l)^2 + (y_1 - y_2)^2 + (y_1 - y_3)^2 + (y_3 - u)^2)1(x_1'\beta < x_3'\beta < x_2'\beta) \\
&\qquad \times 1(x_2'\beta_0 < x_1'\beta_0 < x_3'\beta_0)\big) \\
&\quad + \mathrm{E}\big(((y_2 - l)^2 + (y_3 - y_2)^2 + (y_1 - y_3)^2 + (y_1 - u)^2)1(x_1'\beta < x_3'\beta < x_2'\beta) \\
&\qquad \times 1(x_2'\beta_0 < x_3'\beta_0 < x_1'\beta_0)\big) \\
&\quad + \mathrm{E}\big(((y_3 - l)^2 + (y_1 - y_3)^2 + (y_2 - y_1)^2 + (y_2 - u)^2)1(x_1'\beta < x_3'\beta < x_2'\beta) \\
&\qquad \times 1(x_3'\beta_0 < x_1'\beta_0 < x_2'\beta_0)\big) \\
&\quad + \mathrm{E}\big(((y_3 - l)^2 + (y_3 - y_2)^2 + (y_2 - y_1)^2 + (y_1 - u)^2)1(x_1'\beta < x_3'\beta < x_2'\beta) \\
&\qquad \times 1(x_3'\beta_0 < x_2'\beta_0 < x_1'\beta_0)\big) \\
&= S_{21} + S_{22} + S_{23} + S_{24} + S_{25} + S_{26}, \quad \text{say,}
\end{aligned}$$

and so on up to $T_{6.} \geq S_6 \ldots$. Therefore,

$$\mathrm{E}(q_3(\beta)) \geq \sum_{\tau=1}^{6} S_{\tau.}.$$

Note that

$$\sum_{\tau=1}^{6} S_{\tau 1} = \mathrm{E}\big(((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2)1(x_1'\beta_0 < x_2'\beta_0 < x_3'\beta_0)\big)$$

$$\ldots$$

$$\sum_{\tau=1}^{6} S_{\tau 6} = \mathrm{E}\big(((y_3 - l)^2 + (y_3 - y_2)^2 + (y_2 - y_1)^2 + (y_1 - u)^2)1(x_3'\beta_0 < x_2'\beta_0 < x_1'\beta_0)\big)$$

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press

so that

$$\sum_{\tau=1}^{6} S_{\tau\cdot} = \sum_{\tau=1}^{6} S_{\tau 1} + \sum_{\tau=1}^{6} S_{\tau 2} + \cdots + \sum_{\tau=1}^{6} S_{\tau 6} = \mathrm{E}(q_2(\beta_0)).$$

This proves that β_0 minimizes $\mathrm{E}(q_2(\beta))$.

Step 3. Continuity. We want to show for any sequence $\beta_k \to \beta$, $\mathrm{E}(q_3(\beta_k)) \to \mathrm{E}(q_3(\beta))$.

Consider a sequence $\beta_k \to \beta$ as $k \to \infty$. Consider

$$\begin{aligned} t(\beta_k) - t(\beta) &= [(y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2]1(x_1'\beta_k < x_2'\beta_k < x_3'\beta_k) \\ &\quad - [(y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - u)^2 + (y_3 - u)^2]1(x_1'\beta < x_2'\beta < x_3'\beta) \\ &= [(y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2] \\ &\quad \times [1(x_1'\beta_k < x_2'\beta_k < x_3'\beta_k) - 1(x_1'\beta < x_2'\beta < x_3'\beta)]. \end{aligned}$$

For β_k lying in a small enough neighborhood of β, $1(x_1'\beta_k < x_2'\beta_k < x_3'\beta_k)$ and $1(x_1'\beta < x_2'\beta < x_3'\beta)$ must be either both 0 or both 1 with probability one. Suppose not. Then, e.g., we shall have with positive probability

$$x_1'\beta_k < x_2'\beta_k \quad \text{and} \quad x_1'\beta \geq x_2'\beta,$$

i.e.,

$$(x_1 - x_2)'\beta \geq 0 > (x_1 - x_2)'\beta_k.$$

Taking limits as $k \to \infty$,

$$(x_1 - x_2)'\beta \geq 0 \geq (x_1 - x_2)'\beta,$$

which can hold iff $(x_1 - x_2)'\beta = 0$, a zero probability event under Assumption A2. It follows that with probability one, $t(\beta_k) - t(\beta) \to 0$. Now Assumption A4, together with the dominated convergence theorem, implies

$$\mathrm{E}(t(\beta_k) - t(\beta)) \to 0.$$

A similar argument holds for the other terms in $q_3(\beta)$. ■

https://doi.org/10.1017/S0266466608080328 Published online by Cambridge University Press