

A Permutation-Based Approach to Estimating Monotone Index

Models*

Debopam Bhattacharya[†]

Department of Economics,

Dartmouth College.

January 8, 2007.

Abstract

This note 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 and does not require a bandwidth choice. The estimator is \sqrt{n} -consistent, 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.

*I am grateful to Bo Honore, Yuichi Kitamura, Peter Phillips and two anonymous referees for comments and suggestions. All errors are mine.

[†]Address for correspondence Department of Economics, 301 Rockefeller Hall, Dartmouth College, Hanover, NH 03755; e-mail: debopam.bhattacharya@dartmouth.edu

1 Introduction

This note 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

$$E(y|x) = g(x'\beta_0)$$

where $g(\cdot)$ is monotone. It is shown here that β_0 can be estimated up to location and scale by minimizing an objective function which 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, 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 (1998, henceforth CS) which appears to utilize more information than the MRE does. It should be clarified at the outset that the main purpose of this note is to provide this generalization and not to derive a necessarily "better" estimator than the MRE. However, in a simple Monte Carlo study reported below, the proposed estimator is seen to dominate the MRE in terms of root-mean-squared error and mean absolute deviations.

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 main 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 & 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 $E(y|x) = g(x'\beta_0)$, with $g(.)$ non-constant 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 i.i.d. observations (y_i, x_i) .

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

Let l, u denote the minimum and maximum possible value 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 cdf 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

¹If Y is bounded wp 1, then our results depend only on the property that $E(Y|X)$ 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 $E(\Phi(Y)|X)$ depends on X through $X'\beta$ and is monotonic in $X'\beta$.

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 adjacent 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 then defined as:

$$\hat{\beta} = \arg \min_{\beta \in B} Q_n(\beta) = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} q_m(y_{i_1}, y_{i_2}, \dots, y_{i_m}, x_{i_1}, x_{i_2}, \dots, x_{i_m}; \beta)$$

where

$$q_m(y_{i_1}, y_{i_2}, \dots, y_{i_m}, x_{i_1}, x_{i_2}, \dots, x_{i_m}; \beta) = \sum_{\substack{\{j_1, \dots, j_m\} \\ \in \wp\{i_1 \dots i_m\}}} \left[\left(\sum_{k=1}^{m-1} (y_{j_k} - y_{j_{k+1}})^2 + (y_{j_1} - l)^2 + (y_{j_m} - u)^2 \right) 1(x'_{j_1}\beta < x'_{j_2}\beta < \dots < x'_{j_m}\beta) \right]$$

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

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 $a < y_1 < y_2 < y_3 < \dots < y_m < b$ are m real numbers and $y_{i_1}, y_{i_2}, y_{i_3}, \dots, y_{i_m}$ is a permutation of the y 's, then

$$(y_1 - a)^2 + \sum_{j=2}^m (y_j - y_{j-1})^2 + (y_m - b)^2 \leq (y_{i_1} - a)^2 + \sum_{j=2}^m (y_{i_k} - y_{i_{k-1}})^2 + (y_{i_m} - b)^2.$$

The intuition behind our estimation procedure is the following. 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 .

2.1 Relation with MRE

The MPE can be viewed as a generalization of the MRE. Setting $u = 1$ and $l = 0$ w.l.o.g and expanding the squares in $q_m(\cdot)$ of the previous section to

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

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

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

which is the negative of CS's 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

²For $m = 3$, the objective function is equivalent to

$$\frac{1}{n(n-1)(n-2)} \sum_{k \neq i \neq j} Q_3(x_i, y_i, x_j, y_j, x_k, y_k), \text{ where} \\ -Q_3(x_i, y_i, x_j, y_j, x_k, y_k) \\ = (y_j y_i + y_k y_j + y_k) \mathbf{1}(x'_i \beta < x'_j \beta < x'_k \beta) \\ + (y_k y_i + y_k y_j + y_j) \mathbf{1}(x'_i \beta < x'_k \beta < x'_j \beta) \\ + (y_j y_i + y_k y_i + y_k) \mathbf{1}(x'_j \beta < x'_i \beta < x'_k \beta) \\ + (y_j y_k + y_k y_i + y_i) \mathbf{1}(x'_j \beta < x'_k \beta < x'_i \beta) \\ + (y_j y_k + y_i y_j + y_i) \mathbf{1}(x'_k \beta < x'_j \beta < x'_i \beta) \\ + (y_k y_i + y_i y_j + y_j) \mathbf{1}(x'_k \beta < x'_i \beta < x'_j \beta)$$

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

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

where $h(\cdot, \cdot, \cdot)$ is increasing and non-constant 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 $E(y_{it}|x_{it}, \alpha_i) = E(y_{it}|x'_{it}\beta_0, \alpha_i)$ 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 $t = 1, \dots, T_i$ for each i . Average this sum over $i = 1, \dots, n$ and choose β to minimize this average. The continuity and uniform convergence results for this estimate are completely analogous to the single-cross-section case described above. But now, the objective function is no longer 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 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., $E(y|x)$ depends on x only through the index $x'\beta_0$: $E(y|x) = g(x'\beta_0)$ with $g(\cdot)$ weakly increasing.
- (A1) The support of x is not contained in a proper linear subspace of \mathbb{R}^d .
- (A2) The d -th 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) $Ey^2 < \infty$.

In order 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 above model satisfying A0-A4. Let $f(x_1, x_2, x_3)$ be any non-negative (wp1) function defined on the support of $x = (x_1, x_2, x_3)$. Then for all $i \neq j \neq k \in \{1, 2, 3\}$,*

$$\begin{aligned} & E \left\{ f(x_1, x_2, x_3) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \left((y_1 - l)^2 + (y_2 - y_1)^2 + (y_3 - y_2)^2 + (y_3 - u)^2 \right) \right\} \\ & \leq E \left\{ f(x_1, x_2, x_3) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \left((y_i - l)^2 + (y_j - y_i)^2 + (y_k - y_j)^2 + (y_k - u)^2 \right) \right\} \end{aligned}$$

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.

Theorem 1 *Under assumptions A0-A3 and A4,*

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

Proof of consistency is done through the following steps:

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

The proofs of step 2 and step 1 (b) are analogous to CS. 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 CS 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 d th 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

$$\tau(z, \gamma) = E_{12}q_3(\cdot, \cdot, z, \beta(\gamma)) + E_{13}q_3(\cdot, z, \cdot, \beta(\gamma)) + E_{23}q_3(z, \cdot, \cdot, \beta(\gamma))$$

where $E_{13}q_3(\cdot, z, \cdot, \beta)$ 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 $\tau(\cdot, \gamma)$. 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),

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

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

$$\begin{aligned}\Omega &= V^{-1}\Delta V^{-1}, \\ V &= \frac{1}{3}E\nabla_2\tau(\cdot, \beta(\gamma_0)), \Delta = E(\nabla_1\tau(\cdot, \beta(\gamma_0))\nabla_1\tau(\cdot, \beta(\gamma_0))').\end{aligned}$$

For a general m , the asymptotic variance will be

$$\begin{aligned}\Omega &= V^{-1}\Delta V^{-1}, \\ V &= \frac{1}{m}E\nabla_2\tau(\cdot, \beta(\gamma_0)), \Delta = E(\nabla_1\tau(\cdot, \beta(\gamma_0))\nabla_1\tau(\cdot, \beta(\gamma_0))')\end{aligned}$$

where

$$\tau(z, \gamma) = \sum_{r=1}^m E_{-r}q_m(\dots, z, \dots; \beta(\gamma))$$

and $E_{-r}q_m(\dots, z, \dots; \beta(\gamma))$ denotes the expectation of $q_m(\dots, \cdot, \dots; \beta(\gamma))$ w.r.t. all arguments except the r th one and the r th argument is z .

Exactly as in CS, one can consistently estimate V and Δ by numeric derivatives.

5 Monte Carlo

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

The model is

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

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. 500 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, root mean squared error (RMSE) and mean absolute deviations (MAD) across the replications are reported in the table. 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 the table where MRE denotes the CS estimator and corresponds to the MPE estimator with $m = 2$, 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 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 more importantly

calculation of the MPE involves sorting each of these distinct m -tuples for each choice of the parameter vector. For all sample sizes considered in the Monte Carlo, 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 should m be chosen in practice. A theoretical comparison of asymptotic (and finite-sample) RMSE and mean absolute deviations across the different values of m is hard, tedious and outside the scope of this short note. Further analysis of these issues is therefore left to future research. The Monte Carlos reported in the present paper suggest that the generalized estimator can potentially offer significant efficiency gains, particularly in small samples.

References

- [1] Cavanagh, C. & Sherman, R.P. (1998) Rank estimators for monotonic index models. *Journal of Econometrics*, **84**, 351-381.
- [2] Han, A.K. (1987) Non-parametric analysis of a generalized regression model. *Journal of Econometrics* **35**, 303-316.
- [3] Horowitz, J.L. (1992) A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica* **60**, issue 3, 505-31.
- [4] Sherman, R.P. (1994) Maximal Inequalities for Degenerate U-processes with applications to optimization estimators. *Annals of Statistics* **22**, 439-459.
- [5] Sherman, R.P. (1993) The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123-137.

6 APPENDIX

Proof of Lemma 1:

Consider three observations $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ from the above model satisfying A0-A4a.

Let $f(x_1, x_2, x_3)$ be any non-negative w.p.1 function defined on the support of $x = (x_1, x_2, x_3)$. Then for all $i \neq j \neq k \in \{1, 2, 3\}$,

$$\begin{aligned} & E \left\{ f(x_1, x_2, x_3) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \left((y_1 - l)^2 + (y_2 - y_1)^2 + (y_3 - y_2)^2 + (y_3 - u)^2 \right) \right\} \\ & \leq E \left\{ f(x_1, x_2, x_3) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \left((y_i - l)^2 + (y_j - y_i)^2 + (y_k - y_j)^2 + (y_k - u)^2 \right) \right\} \end{aligned}$$

Proof. Let $Z = 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0)$ and consider w.l.o.g the permutation $\{2, 1, 3\}$. Then we have

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

The term inside $\{.\}$ equals

$$L = -2y_2l - 2y_1y_3 + 2y_1l + 2y_2y_3.$$

Taking expectations, conditional on x_1, x_2, x_3

$$\begin{aligned} & E(L|x_1, x_2, x_3) \\ & = -2g(x'_2\beta_0)l - 2g(x'_1\beta_0)g(x'_3\beta_0) + 2g(x'_1\beta_0)l + 2g(x'_2\beta_0)g(x'_3\beta_0) \\ & = -2g(x'_2\beta_0)(l - g(x'_3\beta_0)) + 2g(x'_1\beta_0)(l - g(x'_3\beta_0)) \\ & = 2\{g(x'_3\beta_0) - l\}\{g(x'_2\beta_0) - g(x'_1\beta_0)\} \end{aligned}$$

Note that when $Z = 1$, we have $x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0$. Given that $g(\cdot)$ is monotone increasing and

that $l < g(x'_i\beta_0)$ for $i = 1, 2, 3$ w.p. 1 in x , we have

$$\{g(x'_3\beta_0) - l\} \{g(x'_2\beta_0) - g(x'_1\beta_0)\} \geq 0 \text{ w.p. } 1$$

Therefore,

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

■

Proof of theorem 1:

Proof. Minimization

$$\begin{aligned} & Eq_3(\beta) \\ = & E\left(\left[(y_i - l)^2 + (y_j - y_i)^2 + (y_k - y_j)^2 + (y_k - u)^2\right] 1(x'_i\beta < x'_j\beta < x'_k\beta)\right) \\ & + E\left(\left[(y_i - l)^2 + (y_k - y_i)^2 + (y_j - y_k)^2 + (y_j - u)^2\right] 1(x'_i\beta < x'_k\beta < x'_j\beta)\right) \\ & + E\left(\left[(y_j - l)^2 + (y_i - y_j)^2 + (y_k - y_i)^2 + (y_k - u)^2\right] 1(x'_j\beta < x'_i\beta < x'_k\beta)\right) \\ & + E\left(\left[(y_j - l)^2 + (y_k - y_j)^2 + (y_i - y_k)^2 + (y_i - u)^2\right] 1(x'_j\beta < x'_k\beta < x'_i\beta)\right) \\ & + E\left(\left[(y_k - l)^2 + (y_i - y_k)^2 + (y_j - y_i)^2 + (y_j - u)^2\right] 1(x'_k\beta < x'_i\beta < x'_j\beta)\right) \\ & + E\left(\left[(y_k - l)^2 + (y_j - y_k)^2 + (y_i - y_k)^2 + (y_i - u)^2\right] 1(x'_k\beta < x'_j\beta < x'_i\beta)\right). \end{aligned}$$

Now write the first term as the sum of 6 terms, each term corresponding one ordering of $(x'_1\beta_0, x'_2\beta_0, x'_3\beta_0)$

$$\begin{aligned}
T_1 &= E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \right) \\
&+ E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_1\beta_0 < x'_3\beta_0 < x'_2\beta_0) \right) \\
&+ E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_2\beta_0 < x'_1\beta_0 < x'_3\beta_0) \right) \\
&+ E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_2\beta_0 < x'_3\beta_0 < x'_1\beta_0) \right) \\
&+ E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_3\beta_0 < x'_1\beta_0 < x'_2\beta_0) \right) \\
&+ E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_3\beta_0 < x'_2\beta_0 < x'_1\beta_0) \right) \\
&= T_{11} + T_{12} + T_{13} + T_{14} + T_{15} + T_{16}, \text{ say.}
\end{aligned}$$

Let

$$\begin{aligned}
S_1 &= E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \right) \\
&+ E \left(\left((y_1 - l)^2 + (y_3 - y_1)^2 + (y_2 - y_3)^2 + (y_2 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_1\beta_0 < x'_3\beta_0 < x'_2\beta_0) \right) \\
&+ E \left(\left((y_2 - l)^2 + (y_1 - y_2)^2 + (y_1 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_2\beta_0 < x'_1\beta_0 < x'_3\beta_0) \right) \\
&+ E \left(\left((y_2 - l)^2 + (y_3 - y_2)^2 + (y_1 - y_3)^2 + (y_1 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_2\beta_0 < x'_3\beta_0 < x'_1\beta_0) \right) \\
&+ E \left(\left((y_3 - l)^2 + (y_1 - y_3)^2 + (y_2 - y_1)^2 + (y_2 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_3\beta_0 < x'_1\beta_0 < x'_2\beta_0) \right) \\
&+ E \left(\left((y_3 - l)^2 + (y_3 - y_2)^2 + (y_2 - y_1)^2 + (y_1 - u)^2 \right) 1(x'_1\beta < x'_2\beta < x'_3\beta) 1(x'_3\beta_0 < x'_2\beta_0 < x'_1\beta_0) \right) \\
&= S_{11} + S_{12} + S_{13} + S_{14} + S_{15} + S_{16}, \text{ say.}
\end{aligned}$$

Lemma 1 with $f(x_1, x_2, x_3) = 1(x'_1\beta < x'_2\beta < x'_3\beta)$ implies

$$T_{1j} \geq S_{1j} \text{ for } j = 1, \dots, 6 \quad (3)$$

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

Similarly,

$$\begin{aligned}
T_2 &= E \left(\left[(y_1 - l)^2 + (y_3 - y_1)^2 + (y_2 - y_3)^2 + (y_2 - u)^2 \right] 1(x'_1\beta < x'_3\beta < x'_2\beta) \right) \\
&\geq S_2 \\
&= E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_3\beta < x'_2\beta) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \right) \\
&\quad + E \left(\left((y_1 - l)^2 + (y_3 - y_1)^2 + (y_2 - y_3)^2 + (y_2 - u)^2 \right) 1(x'_1\beta < x'_3\beta < x'_2\beta) 1(x'_1\beta_0 < x'_3\beta_0 < x'_2\beta_0) \right) \\
&\quad + E \left(\left((y_2 - l)^2 + (y_1 - y_2)^2 + (y_1 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta < x'_3\beta < x'_2\beta) 1(x'_2\beta_0 < x'_1\beta_0 < x'_3\beta_0) \right) \\
&\quad + E \left(\left((y_2 - l)^2 + (y_3 - y_2)^2 + (y_1 - y_3)^2 + (y_1 - u)^2 \right) 1(x'_1\beta < x'_3\beta < x'_2\beta) 1(x'_2\beta_0 < x'_3\beta_0 < x'_1\beta_0) \right) \\
&\quad + E \left(\left((y_3 - l)^2 + (y_1 - y_3)^2 + (y_2 - y_1)^2 + (y_2 - u)^2 \right) 1(x'_1\beta < x'_3\beta < x'_2\beta) 1(x'_3\beta_0 < x'_1\beta_0 < x'_2\beta_0) \right) \\
&\quad + E \left(\left((y_3 - l)^2 + (y_3 - y_2)^2 + (y_2 - y_1)^2 + (y_1 - u)^2 \right) 1(x'_1\beta < x'_3\beta < x'_2\beta) 1(x'_3\beta_0 < x'_2\beta_0 < x'_1\beta_0) \right) \\
&= S_{21} + S_{22} + S_{23} + S_{24} + S_{25} + S_{26}, \text{ say.}
\end{aligned}$$

and so on up to $T_6 \geq S_6$. Therefore,

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

Note that

$$\begin{aligned}
\sum_{\tau=1}^6 S_{\tau 1} &= E \left(\left((y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right) 1(x'_1\beta_0 < x'_2\beta_0 < x'_3\beta_0) \right) \\
&\quad \dots \\
\sum_{\tau=1}^6 S_{\tau 6} &= E \left(\left((y_3 - l)^2 + (y_3 - y_2)^2 + (y_2 - y_1)^2 + (y_1 - u)^2 \right) 1(x'_3\beta_0 < x'_2\beta_0 < x'_1\beta_0) \right)
\end{aligned}$$

so that

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

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

Step 3: Continuity

We want to show for any sequence $\beta_k \rightarrow \beta$, $E(q_3(\beta_k)) \rightarrow E(q_3(\beta))$.

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

$$\begin{aligned}
& t(\beta_k) - t(\beta) \\
&= \left[(y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right] 1(x'_1\beta_k < x'_2\beta_k < x'_3\beta_k) \\
&\quad - \left[(y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - u)^2 + (y_3 - u)^2 \right] 1(x'_1\beta < x'_2\beta < x'_3\beta) \\
&= \left[(y_1 - l)^2 + (y_1 - y_2)^2 + (y_2 - y_3)^2 + (y_3 - u)^2 \right] \\
&\quad \times \left[1(x'_1\beta_k < x'_2\beta_k < x'_3\beta_k) - 1(x'_1\beta < x'_2\beta < x'_3\beta) \right]
\end{aligned}$$

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 wp1. Suppose not. Then, for instance, we shall have with positive probability

$$x'_1\beta_k < x'_2\beta_k \text{ and } x'_1\beta \geq x'_2\beta$$

i.e.

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

Taking limits as $k \rightarrow \infty$,

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

which can hold iff $(x_1 - x_2)' \beta = 0$, a zero probability event under assumption A2. It follows that wp1, $t(\beta_k) - t(\beta) \rightarrow 0$. Now assumption A4, together with the dominated convergence theorem imply

$$E(t(\beta_k) - t(\beta)) \rightarrow 0.$$

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

SIMULATION RESULTS: Homoskedastic Case

n=50

	Mean	Median	RMSE	MAD
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

	Mean	Median	RMSE	MAD
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

	Mean	Median	RMSE	MAD
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

SIMULATION RESULTS: Heteroskedastic Case

n=50

	Mean	Median	RMSE	MAD
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

	Mean	Median	RMSE	MAD
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

	Mean	Median	RMSE	MAD
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