MONOTONICITY CONDITIONS AND INEQUALITY IMPUTATION FOR SAMPLE-SELECTION AND NON-RESPONSE PROBLEMS

(January, 2004)

Myoung-jae Lee
School of Economics and Social Sciences
Singapore Management University
469 Bukit Timah Road
Singapore 259756

mjlee@smu.edu.sg

fax: 65-6822-0833

Under a sample selection or non-response problem where a response variable y is observed only when a condition $\delta = 1$ is met, the identified mean $E(y|\delta = 1)$ is not equal to the desired mean $E(y)$. But the monotonicity condition $E(y|\delta = 1) \leq E(y|\delta = 0)$ yields an informative bound $E(y|\delta = 1) \leq E(y)$, which is enough for certain inferences. For example, in a majority voting with δ being vote-turnout, it is enough to know if $E(y) > 0.5$ or not, for which $E(y|\delta = 1) > 0.5$ is sufficient under the monotonicity. The main question is then whether the monotonicity condition is testable, and if not, when it is plausible. Answering to these queries, when there is a "proxy" variable z related to y but fully observed, we provide a test for the monotonicity; when z is not available, we provide primitive conditions and plausible models for the monotonicity. Going further, when both y and z are binary, bivariate monotonicities of the type $P(y, z|\delta = 1) \leq P(y, z|\delta = 0)$ are considered, which can lead to sharper bounds for $P(y)$. As an empirical example, a data set on the 1996 US presidential election is analyzed to see if the Republican candidate could have won had everybody voted, i.e., to see if $P(y) > 0.5$ where $y = 1$ is voting for the Republican candidate.

Key Words: sample selection, non-response, monotonicity, imputation, orthant dependence.

1. Introduction

Many surveys suffer from sample-selection (or non-response) problems, where the selection (or decision to respond) δ is done in a way it matters for a response variable y of interest. In income surveys, people at either tail of the income range may not respond; the observed response is then δy with $\delta = 1[y$ not in either tail$]$, where $1[A] = 1$ if A holds and 0 otherwise. Our interest is in $E(y|x)$ for a covariate vector x, but what is readily available from the sample is $E(y|x, \delta = 1)$. The question is then what can be learned about $E(y|x)$ under the selection problem.

Since

$$E(y|x) = E(y|x, \delta = 0) \cdot P(\delta = 0|x) + E(y|x, \delta = 1) \cdot P(\delta = 1|x), \qquad (1.1)$$

under $0 < P(\delta = 0|x)$, we get

$$E(y|x, \delta = 0) = E(y|x, \delta = 1) \iff E(y|x) = E(y|x, \delta = 1). \qquad (1.2)$$

Two extreme approaches exist for the selection problem in the literature. First, the *non-verifiable* (1.2) is assumed, amounting to ignoring the selection problem. Call this "ignoring approach"; "ignoring" is taken from Rosenbaum and Rubin (1983). Second, if there is

a "proxy" variable z closely related to y but fully observed, (1.3)

then z is used to impute for y; call this "imputation approach". Here, the sense of "y and z closely related" should come from a *verifiable* condition: at least,

$$E(y|x, \delta = 1) = E(z|x, \delta = 1). \qquad (1.4)$$

The two approaches are diametrically different ways to come up with a complete data: in the former, the missing-mechanism for y is ignored under (1.2) and only the observations with $\delta = 1$ are used; in the latter, missing y is filled in under (1.3), and all observations are used. Although (1.3) is verifiable in the given data, the ensuing imputation requires a *non-verifiable* assumption

$$E(y|x, \delta = 0) = E(z|x, \delta = 0) \qquad (1.5)$$

so that the non-identified $E(y|x, \delta = 0)$ can be replaced with the identified $E(z|x, \delta = 0)$.

Without (1.4), z would not qualify as a source for imputation, and without (1.5), the imputation fails. But, since (1.4) and (1.5) together implies $E(y|x) = E(z|x)$, the imputation approach faces an awkward situation: as far as identification goes, why bother with $E(y|x)$ at all if $E(y|x) = E(z|x)$ and $E(z|x)$ is identified? In view of this, one may wonder whether it is possible to avoid (1.2) in the ignoring approach while making a less demanding use of z than in the imputation approach. A contribution of this paper is showing that indeed it is possible to go halfway (and further) between the two extreme approaches.

A remedy for a selection problem and its consequences on the ensuing statistical inference depend on the problem at hand, and there are cases where we do not need the full force of the either approach above. For example, suppose, for some subpopulation characterized by x,

$y = 1$ if vote for a particular party candidate and 0 otherwise,
$\delta = 1$ if turnout for voting and 0 otherwise,
$z = 1$ if support the party overall and 0 otherwise.

Here we may be interested only in whether $E(y|x) > 0.5$ or not. For this, there is no need to know $E(y|x)$ exactly; bounding $E(y|x)$ from below would be enough. Of course, bounding $E(y|x)$ will not work for every selection problem; it will work only when the inferential goal is modest as in the voting example.

If bounding $E(y|x)$ is the goal, then instead of (1.2), its "50% weaker" one-sided version may do, say,

$$E(y|x, \delta = 1) \leq E(y|x, \delta = 0) \quad \text{((decreasing) `}\textit{monotonicity in } \delta\text{')}. \tag{1.6}$$

Since $E(y|x, \delta = 0)$ is not identified, instead of (1.6), we may test for

$$E(z|x, \delta = 1) \leq E(z|x, \delta = 0) \tag{1.7}$$

that is identified. This motivates *"inequality imputation (or monotonicity imputation)"*: take a test for (1.7) as a test for (1.6), making a less demanding use of z than the usual imputation. Whereas the ignoring approach assumes the non-identified equality (1.2) and the usual imputation approach assumes the non-identified equality (1.5)—call the usual imputation "equality imputation"—inequality imputation assumes the non-identified implication '(1.7) $\Longrightarrow$ (1.6)'.

For an error term e and a covariate w, often the orthogonality $E(ew) = 0$ is assumed, which is relaxed to $E(ew) \geq 0$ in Manski and Pepper (2000). Using $E(ew) = 0$ is analogous to doing equality imputation, whereas using $E(ew) \geq 0$ is analogous to doing inequality imputation. While inequalities such as (1.6) and $E(ew) \geq 0$ are just assumed in the literature, we will try to justify or verify them in this paper—an uncharted territory in the literature.

Going further, when y and z are binary, bivariate inequalities such as

$$P(y=1, z=1|x, \delta=1) \leq P(y=1, z=1|x, \delta=0) \qquad (1.8)$$

will be examined to derive sharper bounds on $P(y=1|x)$. Another motivation to look at (1.8) is that, differently from the univariate monotonicity (1.6), (1.8) is testable, for it implies an identified condition

$$P(y=1, z=1|x, \delta=1) \leq P(z=1|x, \delta=0). \qquad (1.9)$$

Throughout, our approach will be nonparametric without specifying the functional form of $E(y|x)$, nor the distribution of $y|x$.

Section 2 presents various bounds on $E(y|x)$ in the literature to show how monotonicities in δ help bounding $E(y|x)$. Section 3 introduces 'orthant dependence' and common-factor models to show when the monotonicities hold (or are plausible); proxy variables do not appear yet. In Section 4, now with a proxy available, the monotonicities become testable and some bounds obtained in Section 2 get sharpened by taking advantage of bivariate monotonicities such as (1.8). Section 5 analyzes the 1996 US presidential election illustrating our approach with y, δ, z defined as in the above voting example, to find that the Republican candidate (Bob Dole) is likely to have lost anyway, even if everybody had voted,. Finally, Section 6 concludes.

2. How Monotonicity Helps Bounding E(y|x)

In this section, we briefly review part of the bounding approaches that we need for later sections; see Horowitz and Manski (2000), Manski (2003), and the references therein for the literature. Suppose that y is bounded by known lower and upper bounds y_L and y_H; this assumption will be relaxed near the end of this section. Then, the linear transformation $(y-y_L)/(y_H-y_L)$ is bounded by 0 and 1. Hence, without loss of generality, assume

$$0 \leq y \leq 1.$$

Suppose

$$(\delta_i, \delta_i y_i, x_i'), \;\; i = 1, ..., N, \;\; \text{are observed, } \;\; \text{iid across } i, \tag{2.1}$$

where δ_i is a binary random variable and x_i is a regressor vector; y_i is observed only if $\delta_i = 1$ while x_i is observed always. Our interest is on $E(y|x_o)$ $(= E(y|x = x_o))$, not on $E(y|x_o, \delta = 1)$ that is easily identified.

Omitting the subscript i in view of the iid assumption, we get

$$E(y|x) = E(y|x, \delta = 0) \cdot P(\delta = 0|x) + E(y|x, \delta = 1) \cdot P(\delta = 1|x). \tag{2.2}$$

Here, all but $E(y|x, \delta = 0)$ is identified, handling of which determines the identification of $E(y|x)$. To avoid $P(\delta = 1|x) = 0$, assume

$$0 < P(\delta = 1|x) \quad \text{for a.e. } x; \tag{2.3}$$

if $0 < P(\delta = 1|x)$ only for some x, then the bounds to be discussed hereafter hold only for those x satisfying $0 < P(\delta = 1|x)$.

Substituting

$$0 \le E(y|x, \delta = 0) \le 1 \tag{W}$$

into (2.2), we get the 'worst-case bound'

$$E(y|x, \delta = 1) \cdot P(\delta = 1|x) \le E(y|x) \le P(\delta = 0|x) + E(y|x, \delta = 1) \cdot P(\delta = 1|x); \tag{B-W}$$

the size of the bound is

$$P(\delta = 0|x). \tag{S-W}$$

"W" is for Worst case, "B" is for Bound, and "S" is for Size.

Suppose $E(y|x, \delta)$ is monotonic in δ for some x:

$$(0 \le) \; E(y|x, \delta = 0) \le E(y|x, \delta = 1). \tag{M$_\text{U}$}$$

Substituting W and M_U into (2.2), for those x satisfying M_U, we get

$$E(y|x, \delta = 1) \cdot P(\delta = 1|x) \le E(y|x) \le E(y|x, \delta = 1); \tag{B-M$_\text{U}$}$$

the size of the bound is

$$E(y|x, \delta = 1) \cdot P(\delta = 0|x) \le \text{S-W}. \tag{S-M$_\text{U}$}$$

For our empirical example later, $\delta = 1$ if vote, and $y = 1$ if vote for Bob Dole (the Republican candidate in the 1996 US presidential election); $\mathrm{M_U}$ means the higher proportion of Dole-supporters among the voters than among the non-voters.

If the monotonicity $\mathrm{M_U}$ holds with the opposite inequality

$$(1 \geq)\ E(y|x, \delta = 0) \geq E(y|x, \delta = 1), \qquad (\mathrm{M_L})$$

then we get

$$E(y|x, \delta = 1) \leq E(y|x) \leq P(\delta = 0|x) + E(y|x, \delta = 1) \cdot P(\delta = 1|x); \qquad (\text{B-M}_\mathrm{L})$$

the size of the bound is

$$\{1 - E(y|x, \delta = 1)\} \cdot P(\delta = 0|x) \leq \text{S-W}. \qquad (\text{S-M}_\mathrm{L})$$

Suppose now that y is either bounded with unknown bounds or unbounded with $E|y| < \infty$. In this case, no worst-case bound is available, but $\mathrm{M_U}$ and $\mathrm{M_L}$ still yield at least, respectively,

$$E(y|x) \leq E(y|x, \delta = 1) \ \text{ and } \ E(y|x, \delta = 1) \leq E(y|x). \qquad (2.4)$$

If y is bounded only from one side with the bound known, then the worst-case bound is one-sided, and combining the known bound with $\mathrm{M_U}$ or $\mathrm{M_L}$ may yield two-sided bounds. These cases are omitted, however, for they are straightforward to derive.

Manski and Pepper (2000) use $\mathrm{M_U}$ where δ is a treatment and both "potential" responses (say, y_0 and y_1) for $\delta = 0$ and $\delta = 1$ appear as y in the framework of treatment effect analysis (see, e.g., Lee (2004)). This setup is more restrictive than our sample selection framework where only one response appears in $\mathrm{M_U}$. In the empirical example of Lee and Melenberg (1998), $\mathrm{M_U}$ or $\mathrm{M_L}$ more or less halved the worst-case bound, which was a drastic improvement. Thus it is interesting to know when the monotonicities would hold and what kind of models would allow them. Answers to these questions are given in the following section.

3. Conditions and Models for Monotonicity: Orthant Dependence

In the preceding section, we saw bounds for $E(y|x)$ under two monotonicities in δ. In this section, primitive conditions and specific models for the monotonicities are presented; no proxy variable appears in this section. The main result of this section is that the monotonicities hold for the binary response model in (3.4) and for the linear model (3.5) under 'orthant dependence'.

3.1 Orthant dependence

Define 'positive lower orthant dependence (PLOD)' between two continuously distributed random variables u_0 and u_1 as

$$\begin{aligned} P(u_0 &< c_0, u_1 < c_1) \geq P(u_0 < c_0) \cdot P(u_1 < c_1) \quad \forall c_0, c_1 \qquad (3.1) \\ &\iff P(u_1 < c_1 | u_0 < c_0) \geq P(u_1 < c_1) \quad \text{under } P(u_0 < c_0) > 0. \end{aligned}$$

PLOD is equivalent to 'positive upper orthant dependence (PUOD)'

$$\begin{aligned} P(u_0 &> c_0, u_1 > c_1) \geq P(u_0 > c_0) \cdot P(u_1 > c_1) \quad \forall c_0, c_1 \qquad (3.2) \\ &\Leftrightarrow P(u_1 > c_1 | u_0 > c_0) \geq P(u_1 > c_1) \quad \text{under } P(u_0 > c_0) > 0. \end{aligned}$$

Owing to the equivalence, PLOD and PUOD are called simply 'positive orthant dependence (POD)'. The condition $P(u_0 < c_0) > 0$ ($P(u_0 > c_0) > 0$) is not really a restriction, for if $P(u_0 < c_0) = 0$ ($P(u_0 > c_0) = 0$), then (3.1) ((3.2)) holds trivially. POD, which implies non-negative covariance, originates in Lehmann (1966). See Tong (1980), Dharmadhikari and Joag-dev (1988), and Joe (1997) for various other definitions and assertions in this section. Denuit and Scaillet (2001) propose tests for POD and present applications in insurance, but their tests are not applicable when the sample selection problem is present.

If the middle inequalities are reversed in PLOD and PUOD, then we get, respectively, negative lower orthant dependence (NLOD) and negative upper orthant dependence (NUOD); NLOD is equivalent to NUOD, and NLOD and NUOD are simply called 'negative orthant dependence (NOD)'. The equivalencies (PLOD $\iff$ PUOD, and NLOD $\iff$ NUOD) do not hold, however, for more than two random variables. POD and NOD are possibly the weakest concepts of dependence, matching closely what we have in mind when we loosely state two variables are positively or negatively related.

Subtract the second part of (3.1) from one to get

$$P(u_1 > c_1 | u_0 < c_0) \leq P(u_1 > c_1) \iff P(u_1 > c_1 | -u_0 > -c_0) \leq P(u_1 > c_1). \qquad (3.3)$$

Hence the POD of (u_0, u_1) is equivalent to the NOD of $(-u_0, u_1)$. Analogously, the NOD of (u_0, u_1) is equivalent to the POD of $(-u_0, u_1)$.

3.2 Threshold-crossing binary-response

Consider a typical threshold-crossing binary response selection model:

$$(2.1), \quad \delta_i = 1[x_i'\alpha + \varepsilon_i > 0] \quad \text{and} \quad y_i = 1[x_i'\beta + u_i > 0] \tag{3.4}$$

where α and β are parameter vectors and ε_i and u_i are continuously distributed error terms. $\mathrm{M_U}$ is equivalent to

$$P(y=1|x, \delta=0) \leq P(y=1|x).$$

For (3.4), this condition is, under the NOD of $(-\varepsilon, u)|x$,

$$P(u > -x'\beta|x, \varepsilon < -x'\alpha) = P(u > -x'\beta|x, -\varepsilon > x'\alpha) \leq P(u > -x'\beta|x).$$

Hence, (the NOD of $(-\varepsilon, u)|x \iff$) the *POD of* $(\varepsilon, u)|x$ *is sufficient for* M_{U} *to hold in the model (3.4)*, and necessary as well if the support of $(x'\alpha, x'\beta)$ is R^2.

As for $\mathrm{M_L}$, it is equivalent to

$$P(y=1|x) \leq P(y=1|x, \delta=0),$$

and for (3.4), this is

$$P(u > -x'\beta|x) \leq P(u > -x'\beta|x, \varepsilon < -x'\alpha) = P(u > -x'\beta|x, -\varepsilon > x'\alpha).$$

Hence, (the POD of $(-\varepsilon, u)|x \iff$) *the NOD of* $(\varepsilon, u)|x$ *is sufficient for* M_{L} *to hold in the model (3.4)*, and necessary as well if the support of $(x'\alpha, x'\beta)$ is R^2. For the model (3.4), POD and NOD provide a fairly complete characterization of $\mathrm{M_U}$ and $\mathrm{M_L}$.

3.3 Linear-model unbounded-response

For an unbounded y, the bounds in the literature consider only a known bounded transformation of y, say $T(y)$, to get bounds for $T(y)$, which is however rather artificial. In this subsection, we will show that one-sided bounds analogous to $\mathrm{M_U}$ and $\mathrm{M_L}$ hold for an unbounded y under POD and NOD if $E|y| < \infty$.

Suppose, we have

$$(2.1), \quad \delta_i = 1[x_i'\alpha + \varepsilon_i > 0] \quad \text{and} \quad y_i = x_i'\beta + u_i, \quad E|y| < \infty. \tag{3.5}$$

Observe

$$\begin{aligned} E(y|x) &= \int_0^\infty S(t|x)dt - \int_{-\infty}^0 F(t|x)dt \quad \text{where} \\ F(t|x) &\equiv P(y \le t|x) = P(u \le t - x'\beta|x), \quad S(t|x) \equiv 1 - F(t|x) = P(u > t - x'\beta|x). \end{aligned}$$

If $(\varepsilon, u)|x$ is POD, then

$$\begin{aligned} P(u &< t - x'\beta|x) \le P(u < t - x'\beta|x, \varepsilon < -x'\alpha) \qquad (3.6) \\ &\iff P(u < t - x'\beta|x) \le P(u < t - x'\beta|x, -\varepsilon > x'\alpha) = P(u < t - x'\beta|x, \delta = 0). \end{aligned}$$

Also, subtract this from one to get

$$P(u > t - x'\beta|x) \ge P(u > t - x'\beta|x, -\varepsilon > x'\alpha) = P(u > t - x'\beta|x, \delta = 0). \qquad (3.7)$$

Use (3.6) and (3.7), respectively, for $F(t|x)$ and $S(t|x)$ to get

$$E(y|x, \delta = 0) \le E(y|x) \iff E(y|x, \delta = 0) \le E(y|x, \delta = 1).$$

Doing analogously, if $(\varepsilon, u)|x$ is NOD, then

$$E(y|x, \delta = 0) \ge E(y|x) \iff E(y|x, \delta = 1) \le E(y|x, \delta = 0).$$

Hence, the *POD (NOD) of $(\varepsilon, u)|x$ is sufficient for M_U (M_L) in the model (3.5).* Differently from the bounded y case with known bounds, however, $\mathrm{M_U}$ and $\mathrm{M_L}$ give only one-sided bounds in (2.4).

Intermediate cases such as a known lower bound and no finite upper bound are omitted, for they can be easily derived by combining the results of Subsection 3.2 and 3.3.

3.4 Common factor models and regressor dependence

POD and NOD are primitive conditions that may be taken for $\mathrm{M_U}$ and $\mathrm{M_L}$ without much hesitation. But one may further inquire what implies POD or NOD. This subsection presents 'common factors' as further primitive conditions for POD and NOD (and thus for $\mathrm{M_U}$ and $\mathrm{M_L}$).

One of the simplest models of POD is a "common factor" model: with μ denoting a common random variable, suppose

$$u_0 = \mu + v_0 \text{ and } u_1 = \mu + v_1 \text{ where } v_0 \text{ and } v_1 \text{ are iid and } \mu \text{ is independent of } (v_0, v_1); \qquad (3.8)$$

the iid assumption will be relaxed shortly. The variables u_0 and u_1 become NOD if μ in u_1 is replaced by $-\mu$. For instance, if (u_0,u_1) follows a normal distribution with $COV(u_0, u_1) \geq 0$ and has the same marginal distribution, then they can be always written as (3.8).

In (3.4) and (3.5), there is no sample selection problem if ε and u are independent given x. One easy way for ε and u to be related is through an additive omitted variable. In this regard, (3.8) is attractive: μ is the omitted variable linking $u_0 = \varepsilon$ and $u_1 = u$. In a given model for (3.4) and (3.5), one may have a good candidate for μ and know to which direction μ affects δ and y; if the directions are the same, then POD is plausible; otherwise, NOD is plausible. POD in fact holds for a generalized version of (3.8) allowing for nonlinearity of u_j in μ and v_j, which is shown in the following after dependence concepts stronger than POD and NOD are introduced.

Define 'positive regression dependence (PRD)' of u_1 on u_0 as

$$P(u_1 > c_1 | u_0 = c_0) \text{ is non-decreasing in } c_0 \ \forall c_1. \tag{3.9}$$

The PRD of u_1 on u_0, which is also said to be "u_1 is stochastically increasing in u_0", implies POD of u_0 and u_1. If

$$P(u_1 > c_1 | u_0 = c_0) \text{ is non-increasing in } c_0 \ \forall c_1,$$

then u_1 is "negatively regression-dependent on u_0 (NRD)". One example for PRD (NRD) is a bivariate normal distribution with a positive (negative) correlation.

Suppose

$$\begin{aligned} &u_0 = g_0(\mu, v_0) \quad \text{and} \quad u_1 = g_1(\mu, v_1) \quad \text{where } \mu, v_0, v_1 \text{ are independent,} \\ &\text{and } g_j(\cdot, \cdot) \text{ is such that } u_j \text{ is PRD on } \mu, \ j = 0, 1. \end{aligned} \tag{3.10}$$

Theorem 5.3.1 in Tong (1980) proves that POD holds for (3.10) which includes (3.8) as a special case; v_0 and v_1 in (3.10) are not required to follow the same distribution as in (3.8). The common factor models appear in Lee (1999) in a bivariate context related to (3.4).

In summary of this section, common factor models imply POD or NOD, which in turn implies monotonicities in δ for $E(y|x, \delta)$. But if we do not have an idea on the direction of the influence of the common factor on δ and y, this finding is not of much help. For this, an alternative is looking for a proxy variable z for y, which is explored in the following section. Of course, there is no "free lunch": we need either a restriction such as M_U or an extra variable such as z.

4. Bivariate Monotonicities for Bivariate Binary Responses

In the preceding sections, we derived bounds for $E(y|x)$ under monotonicities in δ in univariate response cases and then provided primitive conditions and models implying the monotonicities. In this section, bivariate responses with y and z are considered where z is fully observed along with x regardless of δ:

$$(\delta, \delta y, z, x') \text{ is observed.} \tag{4.1}$$

As mentioned already, we may do *inequality imputation* with z—test for the inequality (1.7) to accept/reject the inequality (1.6)—which is one of main proposals of this paper. But since this is straightforward to do, this section focuses on bivariate monotonicities of the type (1.8), because they can be tested and then used to get sharper bounds on $E(y|x)$. We will let y and z binary in this section so that the resulting bounds can be used for the empirical example for voting in Section 5. The bounds derived for binary y and z may be extended to generic y and z, using $1[y < t]$ and $1[z < t]$ and then doing analogously to Subsection 3.3.

Although we entertain the availability of z, the new variable z does not necessarily mean that we need extra information relative to the univariate response cases, because a component of x can be pulled out of x to be used as z. This statement is, however, subject to the caveat that the subpopulation characterized by x changes by losing the component.

Assume

$$0 < P(z = 1|x, \delta = 0) < 1 \quad \text{for a.e. } x; \tag{4.2}$$

if this holds only for some x, the bounds below hold only for those x. In our empirical example later, $y = 1$ if vote for Dole, $z = 1$ if Republican, and $\delta = 1$ if vote; z is observed for everybody. In Subsection 4.1 and 4.2, we examine bounds on the joint probabilities $P(y = 1, z = 1|x)$ and $P(y = 1, z = 0|x)$, respectively, using bivariate monotonicities. In Subsection 4.3, the bounds on the joint probabilities are combined to render improved bounds on $P(y = 1|x)$.

It should be also mentioned that, although we use z as a proxy variable for y and explore bounds on $P(y = 1, z = 1|x)$ and $P(y = 1, z = 0|x)$ only for $P(y = 1|x)$, there are cases where we may be genuinely interested in $P(y = 1, z = 1|x)$ and $P(y = 1, z = 0|x)$. An example is that y and z are two measures of a job-training program success; e.g., y is the dummy variable for being employed one year after (thus observed only for those still tracked after one year) while z is finding a job immediately after the program (thus observed for all trainees). In such cases, Subsection 4.1 and 4.2 are of interest on their own. To avoid cluttering notations, "$|x$" will be omitted in this section other than in the subsection headings; qualifiers such as "for some x for which this condition holds" will be omitted as well.

4.1 $\mathbf{P(y_1{=}1,y_2{=}1|x)}$

Observe

$$P(y=1,z=1)=P(y=1,z=1|\delta=0)\cdot P(\delta=0)+P(y=1,z=1|\delta=1)\cdot P(\delta=1);\quad (4.3)$$

all but $P(y=1,z=1|\delta=0)$ is identified. Differently from univariate cases, the worst-case bound uses, *not* $0\le P(y=1,z=1|\delta=0)\le 1$, but

$$0\le P(y=1,z=1|\delta=0)\le P(z=1|\delta=0).\qquad \text{(W')}$$

W' yields

$$\begin{aligned}P(y &= 1,z=1|\delta=1)\cdot P(\delta=1)\ \le\ P(y=1,z=1)\ \le &\text{(B-W')}\\ P(z &= 1|\delta=0)\cdot P(\delta=0)+P(y=1,z=1|\delta=1)\cdot P(\delta=1);&\\ P(z &= 1|\delta=0)\cdot P(\delta=0)\ \text{ is the size of B-W'.}&\text{(S-W')}\end{aligned}$$

Suppose, the *bivariate version of monotonicity in* δ holds:

$$P(y=1,z=1|\delta=0)\le P(y=1,z=1|\delta=1).\qquad (\mathrm{M_U}')$$

For our empirical example, $\mathrm{M_U}$' is that there are more Dole-supporting Republicans among voters than among non-voters. Combine $\mathrm{M_U}$' with W' to get

$$0\le P(y=1,z=1|\delta=0)\le \min\{P(y=1,z=1|\delta=1),\ P(z=1|\delta=0)\}.$$

This yields

$$\begin{aligned}P(y=1,z=1|\delta=1)\cdot P(\delta=1)\ \le P(y=1,z=1)\ \le\qquad &(\text{B-M}_\text{U}')\\ \min\{P(y=1,z=1|\delta=1),\ P(z=1|\delta=0)P(\delta=0)+P(y=1,z=1|\delta=1)P(\delta=1)\};&\\ \text{the size is }\min\{P(y=1,z=1|\delta=1)P(\delta=0),\ \text{S-W'}\}\le \text{S-W'}.\qquad &(\text{S-M}_\text{U}^{'})\end{aligned}$$

The appendix shows a primitive condition for the monotonicity in the bivariate response case for the threshold-crossing model; the condition is "dependence through stochastic ordering" that generalizes orthant dependence.

Suppose $\mathrm{M_U}$' holds with the opposite inequality:

$$P(y=1,z=1|\delta=0) \geq P(y=1,z=1|\delta=1). \tag{M$_\mathrm{L}$'}$$

For $\mathrm{M_L}$', it is necessary to have

$$P(y=1,z=1|\delta=1) \leq P(z=1|\delta=0) \tag{M$_\mathrm{L}$'-ID}$$

that is identified. Combine $\mathrm{M_L}$' with W' to get

$$P(y=1,z=1|\delta=1) \leq P(y=1,z=1|\delta=0) \leq P(z=1|\delta=0).$$

This yields

$$\begin{gathered} P(y=1,z=1|\delta=1) \ \leq \ P(y=1,z=1) \ \leq \qquad \text{(B-M}_\mathrm{L}\text{')} \\ P(z=1|\delta=0)\cdot P(\delta=0) + P(y=1,z=1|\delta=1)\cdot P(\delta=1); \\ \text{the size is } \{P(z=1|\delta=0) - P(y=1,z=1|\delta=1)\}P(\delta=0) \leq \text{S-W'}. \qquad \text{(S-M}_\mathrm{L}\text{')} \end{gathered}$$

$\mathrm{M_L}$'-ID assures S-$\mathrm{M_L}$'≥ 0. Although $\mathrm{M_U}$' and $\mathrm{M_L}$' are impossible to verify, the implication $\mathrm{M_L}$'-ID of $\mathrm{M_L}$' is. Hence, *checking out M_L'-ID will help see which one of M_U' and M_L' is the more plausible.*

4.2 $\mathbf{P(y_1{=}1,y_2{=}0|x)}$

Turning to $P(y=1,z=0)$, analogous bounds hold, which are presented here without discussion. Observe

$$\begin{gathered} 0 \leq P(y=1,z=0|\delta=0) \leq P(z=0|\delta=0); \qquad \text{(W")} \\ P(y=1,z=0|\delta=1)\cdot P(\delta=1) \ \leq \ P(y=1,z=0) \ \leq \qquad \text{(B-W")} \\ P(z=0|\delta=0)\cdot P(\delta=0) + P(y=1,z=0|\delta=1)\cdot P(\delta=1); \\ P(z=0|\delta=0)\cdot P(\delta=0). \qquad \text{(S-W")} \end{gathered}$$

Suppose

$$P(y=1,z=0|\delta=0) \leq P(y=1,z=0|\delta=1). \tag{M$_\mathrm{U}$"}$$

Combine $\mathrm{M_U}$" with W" to get

$$0 \ \leq \ P(y=1,z=0|\delta=0) \ \leq \ \min\{P(y=1,z=0|\delta=1), \ P(z=0|\delta=0)\}.$$

This yields

$$P(y=1,z=0|\delta=1)\cdot P(\delta=1) \ \le\ P(y=1,z=0) \ \le \qquad \text{(B-M}_\text{U}\text{”)}$$
$$\min\{P(y=1,z=0|\delta=1),\ P(z=0|\delta=0)P(\delta=0)+P(y=1,z=0|\delta=1)P(\delta=1)\};$$
$$\min\{P(y=1,z=0|\delta=1)P(\delta=0),\ \text{S-W”}\} \le \text{S-W”}. \qquad \text{(S-M}_\text{U}\text{”)}$$

Suppose that the monotonicity $\mathrm{M_U}$” holds with the opposite equality:

$$P(y=1,z=0|\delta=0) \ge P(y=1,z=0|\delta=1), \qquad \text{(M}_\text{L}\text{”)}$$

for which it is necessary to have

$$P(y=1,z=0|\delta=1) \le P(z=0|\delta=0) \qquad \text{(M}_\text{L}\text{”-ID)}$$

that is identified. Combine $\mathrm{M_L}$” with W” to get

$$P(y=1,z=0|\delta=1) \le P(y=1,z=0|\delta=0) \le P(z=0|\delta=0).$$

This yields

$$\begin{aligned} P(y &= 1,z=0|\delta=1) \ \le\ P(y=1,z=0) \ \le && \text{(B-M}_\text{L}\text{”)} \\ P(z &= 0|\delta=0)\cdot P(\delta=0)+P(y=1,z=0|\delta=1)\cdot P(\delta=1); && \\ \{P(z &= 0|\delta=0)-P(y=1,z=0|\delta=1)\}\cdot P(\delta=0) \le \text{S-W”}. && \text{(S-M}_\text{L}\text{”)} \end{aligned}$$

Owing to $\mathrm{M_L}$”-ID, S-$\mathrm{M_L}$”≥0. Although $\mathrm{M_U}$” and $\mathrm{M_L}$” are impossible to verify, the implication $\mathrm{M_L}$”-ID of $\mathrm{M_L}$” is. *Checking out M_L”-ID will help see which one of M_U” and M_L” is the more plausible.*

4.3 Combining bivariate inequalities

After testing for $\mathrm{M_L}$’-ID and $\mathrm{M_L}$”-ID, one can combine the bounds on $P(y=1,z=1|x)$ and $P(y=1,z=0|x)$ to get a bound on $P(y=1|x)$. This may yield an improved bound for $P(y=1|x)$ relative to B-W—but not always. For instance, using B-W’ and B-W” in the preceding subsections yields only B-W; no improvement here. But some combinations of $\mathrm{M_L}$’, $\mathrm{M_L}$”, $\mathrm{M_U}$’, and $\mathrm{M_U}$” do give better bounds. One example is ($\mathrm{M_U}$’, $\mathrm{M_L}$”); to shorten our exposition, we present only this combination in the following paragraph, which will then be used for our empirical analysis later.

Under (M_U',M_L"), combining B-M_U' and B-M_L" renders

$$P(y=1|\delta=1)\cdot P(\delta=1)+P(y=1,z=0|\delta=1)\cdot P(\delta=0) \ \leq \ P(y=1) \ \leq$$
$$\min\{P(\delta=0)+P(y=1|\delta=1)P(\delta=1), \qquad \text{(B-M}_{\text{UL}}\text{)}$$
$$P(z=0|\delta=0)P(\delta=0)+P(y=1|\delta=1)P(\delta=1)+P(y=1,z=1|\delta=1)P(\delta=0)\}.$$

The lower bound is sharper than that of B-M_U and B-W which is only the first term in the lower bound of B-M_{UL}. The upper bound is at least as good as that of B-M_L and B-W which is the first term in the min function. Hence, B-M_{UL} is sharper than at least B-W.

5. 1996 US Presidential Election

In this section, we apply our proposals to the 1996 US presidential election with three major candidates: Bill Clinton, Bob Dole, and Ross Perot. The US presidential election follows an electoral college system, not the usual popular vote. In the US history, it happened a couple of times (including the one between Bush vs. Gore) that a candidate lost the election despite winning the popular vote. We will however pretend that the US presidential election follows the popular vote. One reason is that, if the election is so close that we have to worry about the difference between the two systems, then statistical methods are unlikely to provide meaningful answers; recall the "statistical dead-heat" before election in Bush vs. Gore. Another reason is that our sample size is not large enough to go down to the state level, which is necessary to consider the electoral college system.

The following is the voting proportions using the data of $N = 1670$ in the National Election Studies (Warren et al. (1999)):

No vote	Clinton	Dole	Perot
0.33	0.36	0.26	0.05

This data with $N = 1670$ excludes those who voted for the other candidates; only 3% of the original data with $N = 1714$ voted this way. Differently from the 1992 election, Perot's presence was not much.

Given the small 10% difference between Clinton and Dole and the large 33% no vote rate, an intriguing question from Dole's viewpoint would be whether he could have won if the voter turnout had been better. Some Dole supporters might have chosen not to vote, for they thought that Dole would lose anyway with or without their votes. Although the choice is not observed for non-voters, there was no missing in the dummy variable for being Republican, which is likely to be highly positively correlated with voting for Dole. Thus, set

$y = 1$ if vote for Dole among the three candidate.
$z = 1$ if Republican.

We will examine two questions: Could Dole have won if

Q1: everybody had voted, with Perot out,
Q2: everybody had voted, with Perot in;

although Perot took only 5% of the votes, this may matter if this imaginary election with everybody voting is a close one. The variable δ will be set in two different ways:

for Q1, $\delta = 1$ if vote for Clinton or Dole,
for Q2, $\delta = 1$ if vote for Clinton, Dole, or Perot;

in the former, voting for Perot is treated as no-vote.

All bounds given below carry four numbers, say $a < b < c < d$, where b and c are the lower and upper bound estimate for $E(y)$, respectively, and a is the lower 95% confidence interval (CI) bound for b whereas d is the upper 95% CI bound for c. The differences $b - a$ and $d - c$ are $2 \sim 3\%$ in most cases and do not alter any conclusion based only on b and c in the following.

For B-M$_{\text{UL}}$ in Subsection 4.3, the upper bound involves a min function of two estimators. For this, we derived the asymptotic joint normal distribution of the two estimators and then obtained a 95% CI using

$$0.975 = P(\min(a_N, b_N) < \lambda) = 1 - P(\min(a_N, b_N) > \lambda) = 1 - P(a_N > \lambda, b_N > \lambda)$$

for two estimators a_N and b_N and a constant λ; λ gives the upper 95% CI bound for the upper bound of B-M$_{\text{UL}}$.

Examining Q1, we obtained the following bounds without using z:

B-W:	0.24	0.26	0.64	0.66	
B-M$_{\text{U}}$:	0.24	0.26	0.42	0.45	(5.1)
B-M$_{\text{L}}$:	0.39	0.42	0.64	0.66	

Only B-M$_{\text{U}}$ gives the definite conclusion that Dole would have lost anyway with Perot out, which is natural in view of what M$_{\text{U}}$ means (i.e., more Dole supporters among the voters than among the non-voters).

Turning to using z, 95% CI's for $E(z|\delta = 0)$ and $E(z|\delta = 1)$ are respectively

$$0.19 \pm 0.03 \quad \text{and} \quad 0.33 \pm 0.03;$$

clearly, $E(z|\delta = 0) \leq E(z|\delta = 1)$ holds. Doing inequality imputation, we take M$_{\text{U}}$ and the resulting B-M$_{\text{U}}$: Dole would have lost anyway with Perot out.

For the bivariate monotonicities, we obtained the following for the two identified conditions M$_{\text{L}}$'-ID and M$_{\text{L}}$"-ID:

$$\begin{array}{lcccc} \text{M}_{\text{L}}\text{'-ID:} & 0.29 & 0.32 & 0.15 & 0.19 \\ \text{M}_{\text{L}}\text{"-ID:} & 0.13 & 0.15 & 0.75 & 0.81 \end{array} \tag{5.2}$$

M$_{\text{L}}$'-ID is soundly rejected while M$_{\text{L:}}$"-ID is not. Hence, we take M$_{\text{U}}$' and M$_{\text{L}}$", which yields B-M$_{\text{UL}}$:

$$\begin{array}{lcccc} \text{B-M}_{\text{UL}}\text{:} & 0.28 & 0.31 & 0.64 & 0.66 \end{array} \tag{5.3}$$

The lower bound is sharper than that of B-M$_{\text{U}}$, whereas the upper bound is the same as that of B-M$_{\text{L}}$. Although B-M$_{\text{UL}}$ is sharper than B-W, it is still inconclusive.

Examining Q2, we obtained the following bounds without using z:

$$\begin{array}{rcccc} \text{B-W:} & 0.24 & 0.26 & 0.59 & 0.62 \\ \text{B-M}_{\text{U}}\text{:} & 0.24 & 0.26 & 0.39 & 0.42 \\ \text{B-M}_{\text{L}}\text{:} & 0.36 & 0.39 & 0.59 & 0.62 \end{array} \tag{5.4}$$

As in (5.1), only B-M$_{\text{U}}$ gives the definite answer that Dole would have lost anyway with Perot in, although all bounds shift downward compared with (5.1).

Turning to using z, 95% CI's for $E(z|\delta = 0)$ and $E(z|\delta = 1)$ are respectively

$$0.18 \pm 0.04 \quad \text{and} \quad 0.32 \pm 0.02;$$

$E(z|\delta = 0) \leq E(z|\delta = 1)$ still holds. Doing inequality imputation, we take M$_{\text{U}}$ and the resulting B-M$_{\text{U}}$: Dole would have lost anyway with Perot in.

M$_{\text{L}}$'-ID and M$_{\text{L}}$"-ID are

$$\begin{array}{lcccc} \text{M}_{\text{L}}\text{'-ID:} & 0.27 & 0.30 & 0.15 & 0.18 \\ \text{M}_{\text{L}}\text{"-ID:} & 0.12 & 0.14 & 0.76 & 0.82 \end{array} \tag{5.5}$$

As in (5.2), $\mathrm{M_L}$'-ID is soundly rejected while $\mathrm{M_L}$"-ID is not; as before, we take $\mathrm{M_U}$' and $\mathrm{M_L}$". Combining $\mathrm{M_U}$' and $\mathrm{M_L}$", we get B-$\mathrm{M_{UL}}$ that is sharper than B-W:

$$\text{B-M}_{\text{UL}}: \quad 0.27 \quad 0.30 \quad 0.59 \quad 0.62 \tag{5.6}$$

Although this does not give a definite answer, the number 0.30 is below the critical number 0.5 by 0.2 whereas the number 0.59 is above 0.5 by only 0.09. Recalling Q2, it is unlikely that Dole would have won with Perot in, even if everybody had voted.

It is certainly possible that, for some sub-population with $x = x_0$, some of the above bounds get sharper, yielding definitely positive outcomes for Dole. This would then mean that Dole could have focused his campaign on this sub-population to have had at least a closer race.

6. Conclusions

In this paper, we examined monotonicity conditions useful for bounding regression functions in models with sample selection or non-response problems. When no proxy variable is available that is fully observed but related to the response variable, a weak primitive condition (orthant dependence) was presented for monotonicities in binary-response models as well as in linear models. When a proxy variable is available, 'inequality imputation' is suggested where the monotonicity for the response variable is tested using the proxy variable instead. We also explored monotonicities in bivariate cases; in some cases, bivariate monotonicities led to an improved bound for the regression function. The regression-function bounds were applied to the US 1996 presidential election to show that Dole would have lost anyway even if everybody had voted; undoubtedly, our methodology can be easily applied to other elections and votings.

APPENDIX: Primitive Conditions for Monotonicity in Bivariate Responses

To simplify notations, suppose $(\delta_i, \delta_i y_{i1}, y_{i2}, x_i')$ is observed (thus $y_1 = y$ and $y_2 = z$) and

$$\delta_i = 1[x_i'\beta_0 + u_{i0} > 0] \quad \text{and} \quad y_{ij} = 1[x_i'\beta_j + u_{ij} > 0], \quad j = 1, 2 \tag{a.1}$$

where β_j, $j = 0, 1, 2$, are parameter vectors, and u_{i0}, u_{i1}, and u_{i2} are continuously distributed error terms.

The extension of PUOD to the trivariate case is, under $P(u_0 > c_0) > 0$,

$$P(u_0 \quad > \quad c_0, u_1 > c_1, u_2 > c_2) \geq P(u_0 > c_0)\ P(u_1 > c_1)\ P(u_2 > c_2) \ \ \forall c_0, c_1, c_2 \tag{a.2}$$

$$\iff \quad P(u_1 > c_1, u_2 > c_2 | u_0 > c_0) \geq P(u_1 > c_1)\ P(u_2 > c_2). \tag{a.3}$$

What we need is however not this, but

$$P(u_1 > c_1, u_2 > c_2 | u_0 > c_0) \geq P(u_1 > c_1, u_2 > c_2) \quad \forall c_0, c_1, c_2 \quad \text{under } P(u_0 > c_0) > 0. \tag{a.4}$$

If u_1 and u_2 are POD, then the lower bound in (a.4) is sharper than that in (a.3). For the trivariate version of (3.10) with $u_2 = g_2(\mu, v_2)$ added, (a.4) holds owing to Theorem 5.3.1 in Tong (1980). In the following, we explore general conditions for (a.4).

A multivariate version of PRD is ‘positive dependence through stochastic ordering (PDS)’ in Joe (1997, p.21): $\{u_0, u_1, ..., u_J\}$ is PDS, if for any $m \in \{0, 1, ..., J\}$ and c_j’s,

$$P(u_j > c_j,\ j \neq m,\ j = 0, 1, ..., J | u_m = c_m) \text{ is non-decreasing in } c_m. \tag{a.5}$$

PDS implies the multivariate PUOD and PLOD (Joe (1997, p.27)). We can define ‘negative dependence through stochastic ordering (NDS)’ analogously: for any $m \in \{0, 1, ..., J\}$ and c_j’s,

$$P(u_j > c_j,\ j \neq m,\ j = 0, 1, ..., J | u_m = c_m) \text{ is non-increasing in } c_m. \tag{a.6}$$

NDS implies the multivariate NUOD and NLOD where the trivariate NUOD is (NLOD is analogous and so omitted)

$$P(u_0 > c_0, u_1 > c_1, u_2 > c_2) \leq P(u_0 > c_0) \cdot P(u_1 > c_1) \cdot P(u_2 > c_2) \quad \forall c_0, c_1, c_2. \tag{a.7}$$

An example of PDS (NDS) is a multivariate standard normal distribution with all non-negative (non-positive) correlations (Joe (1997, p.34)).

For our purpose, PDS needs to hold only for $m = 0$. Now, with $c_0 > c_n$,

$$\begin{aligned} P(u_j &> c_j,\ j=1,...,J|u_0 > c_0) \geq P(u_j > c_j,\ j=1,...,J|u_0 > c_n) \qquad (\text{a.8}) \\ &\to P(u_j > c_j,\ j=1,...,J) \quad \text{as } c_n \to -\infty. \end{aligned}$$

In the trivariate case, this is exactly (a.4).

M$_\text{L}$' for (a.1) is

$$P(u_1 > -x'\beta_1, u_2 > -x'\beta_2|x) \leq P(u_1 > -x'\beta_1, u_2 > -x'\beta_2|x, -u_0 > x'\beta_0). \qquad (\text{a.9})$$

For this, it is sufficient that $(-u_0, u_1, u_2)|x$ is PDS only with respect to $-u_0$. M$_\text{L}$" for (a.1) is

$$\begin{aligned} &P(u_1 > -x'\beta_1, u_2 < -x'\beta_2|x) \leq P(u_1 > -x'\beta_1, u_2 < -x'\beta_2|x, -u_0 > x'\beta_0) \qquad (\text{a.10}) \\ \iff\ &P(u_1 > -x'\beta_1, -u_2 > x'\beta_2|x) \leq P(u_1 > -x'\beta_1, -u_2 > x'\beta_2|x, -u_0 > x'\beta_0) \end{aligned}$$

for which it is sufficient that $(-u_0, u_1, -u_2)|x$ is PDS only for $-u_0$.

As for M$_\text{U}$' for (a.1), we need

$$P(u_1 > -x'\beta_1, u_2 > -x'\beta_2|x) \geq P(u_1 > -x'\beta_1, u_2 > -x'\beta_2|x, -u_0 > x'\beta_0). \qquad (\text{a.11})$$

For this, it is sufficient that $(-u_0, u_1, u_2)|x$ is NDS only for $-u_0$. M$_\text{U}$" for (a.1) is

$$\begin{aligned} &P(u_1 > -x'\beta_1, u_2 < -x'\beta_2|x) \geq P(u_1 > -x'\beta_1, u_2 < -x'\beta_2|x, -u_0 > x'\beta_0) \qquad (\text{a.12}) \\ \iff\ &P(u_1 > -x'\beta_1, -u_2 > x'\beta_2|x) \geq P(u_1 > -x'\beta_1, -u_2 > x'\beta_2|x, -u_0 > x'\beta_0) \end{aligned}$$

for which it is sufficient that $(-u_0, u_1, -u_2)|x$ is NDS only for $-u_0$.

REFERENCES

Denuit, M. and O. Scaillet, 2001, Nonparametric tests for positive quadrant dependence, unpublished paper.

Dharmadhikari, S. and K. Joag-dev, 1988, Unimodality, convexity and applications, Academic Press.

Horowitz, J.L. and C.F. Manski, 2000, Nonparametric analysis of randomized experiments with missing covariate and outcome data, Journal of the American Statistical Association 95, 77-84.

Joe, H., 1997, Multivariate models and dependence concepts, Chapman & Hall.

Lee, M.J., 1999, Nonparametric estimation and test for quadrant correlations in multivariate binary response models, Econometric Reviews 18, 387-415.

Lee, M.J., 2004, Micro-econometrics for policy, program, and treatment effects, Oxford University Press, forthcoming.

Lee, M.J. and B. Melenberg, 1998, Bounding quantiles in sample selection models, Economics Letter 61, 29-35.

Lehmann, E.L., 1966, Some concepts of dependence, Annals of Mathematical Statistics 37, 1137-1153.

Manski, C.F., 2003, Partial identification of probability distributions, Springer-Verlag.

Manski, C.F. and J.V. Pepper, 2000, Monotone instrumental variables: with an application to the returns to schooling, Econometrica 68, 997-1010.

Rosenbaum, P.R. and D.B. Rubin, 1983, The central role of the propensity score in observational studies for causal effects, Biometrika 70, 41-55.

Tong, Y.L., 1980, Probability inequalities in multivariate distributions, Academic Press.

Warren, E.M., D.R. Kinder, and S.J. Rosenstone, 1999, National Election Studies 1996, Center for Political Studies, University of Michigan, U.S.A.