TITLE: Monotonicity Conditions and Inequality Imputation for Sample-Selection and Non-Response Problems.

NAME: Lee, Myoung-jae

ADDRESS: Institute of policy and planning sciences, University of Tsukuba, Tsukuba, Ibaraki, 305-8573, Japan (until Aug 31, 2001). Dept. Economics, Sungkyunkwan University, 3-53 Myongryun-dong, Chongro-gu, Seoul 110-745, Korea (from Sep.1 and onward).

Suppose $(d_i; d_i y_{1i}; y_{2i})$, $i = 1; :::; N$, are observed where d_i is a binary "selection" indicator, y_{1i} is a binary response variable of interest, and y_{2i} is a binary variable related to y_{1i}. For example, in an U.S. presidential election, $d_i = 1$ if vote, $y_{1i} = 1$ if vote for the Republican candidate, and $y_{2i} = 1$ if the voter is Republican. We are interested in $P(y_1 = 1)$; but only $P(y_1 = 1jd = 1)$ is identi…ed. In

$$P(y_1 = 1) = P(y_1 = 1jd = 0)P(d = 0) + P(y_1 = 1jd = 1)P(d = 1); \quad (1)$$

$P(y_1 = 1jd = 0)$ is not identi…ed, but can be imputed using y_2 if

$$P(y_1 = 1jd = 0) = P(y_2 = 1jd = 0); \quad (2)$$

based on the notion that y_2 is related (i.e., similar) to y_1. But (2) is too strong an assumption; I propose two alternatives using y_2 to overcome the selection (or non-response) problem.

The …rst is a weak form of imputation, called "inequality imputation". Suppose

$$P(y_1 = 1; y_2 = 1jd = 0) \le P(y_1 = 1; y_2 = 1jd = 1) \quad (3)$$

$$) \quad P(y_2 = 1jd = 0) \le P(y_1 = 1; y_2 = 1jd = 1): \quad (4)$$

(4) is testable; if (4) is accepted, adopt (3); otherwise, adopt "monotonicity in ±":

$$P(y_1 = 1; y_2 = 1jd = 0) \cdot P(y_1 = 1; y_2 = 1jd = 1): \quad (5)$$

"Inequality Imputation" is going from (5) to the monotonicity in ± involving only y_1 :

$$(0 \cdot) \; P(y_1 = 1jd = 0) \cdot P(y_1 = 1jd = 1); \quad (6)$$

because "y_1 and y_2 are similar"; compare this to the equality imputation $y_1 = y_2$, which implies (5)=(6). Using (6) in (1) yields the following bound on $P(y_1 = 1)$:

$$P(y_1 = 1jd = 1)P(d = 1) \cdot P(y_1 = 1) \cdot P(y_1 = 1jd = 1): \quad (7)$$

The second alternative is that, without going for (6) or its like resulting from inequality imputattion, (3) or (5) implies a bound on $P(y_1 = 1; y_2 = 1)$, and combining this bound with an analogous bound for $P(y_1 = 1; y_2 = 0)$ leads to a bound on $P(y_1 = 1)$: For my data, this approach yields the following lower and upper bound for $P(y_1 = 1)$:

$$P(y_1 = 1|d = 1)P(d = 1) + P(y_1 = 1; y_2 = 0|d = 1)P(d = 0); \quad (8)$$

$$P(y_1 = 1|d = 1)P(d = 1) + \min\{P(d = 0);$$

$$P(y_2 = 0|d = 0)P(d = 0) + P(y_1 = 1; y_2 = 1|d = 1)P(d = 0)\}:$$

In a 1996 U.S. Presidential Election data (N=1670, from the National Election Studies) on Bill Clinton, Bob Dole and Ross Pero, $P(d = 0) = 0:33$, $P(y_1 = 1) = 0:26$ and P(vote for Clinton) $= 0:36$; one may wonder if Dole could have won had everybody voted. Testing for (4) leads to taking (5), and an analogous test when $y_1 = 0$ leads to taking (5) with ¸ and $y_1 = 0$: The …rst approach, inequality imputation, renders the following bound for (7) (the sampling error expands the bound by about 2-3 % both ways and is thus immaterial):

$$0:26 \cdot P(y_1 = 1) \cdot 0:39:$$

The second approach yields (again allow 2-3% margin) the following bound for (8):

$$0:30 \cdot P(y_1 = 1) \cdot 0:59:$$

In the former, we get a de…nite conclusion that Dole would have lost anyway, while in the latter, we can only say that there is more room (0.2) below 0.5 than above 0.5 (0.09), meaning that it is likely that Dole would have lost as well

RESUMÉ: associate professor at University of Tsukuba.