Trimming for Bounds on Treatment Effects with Missing Outcomes *

David S. Lee

UC Berkeley and NBER

September 2003

Abstract

Consider a treatment evaluation problem with missing outcomes due to sample selection: $Y^* = D\beta + X\pi_1 + U$, $Z^* = D\gamma + X\pi_2 + V$, and $Y = 1[Z^*>0]\cdot Y^*$. *(U,V)* is independent of *(D,X)*, *D* is a binary treatment, and data are only provided on *(Y,X,D)*. This paper proposes a simple and intuitive procedure for bounding β in this model (and its generalizations), when no exclusion restrictions hold, and when *Y* has unbounded support. The proposed trimming procedure yields the tightest bounds on β consistent with the observed data. The result is directly applicable to matching estimation procedures, and the analysis of randomized experiments, whenever there is non-random sample attrition.

* Department of Economics, UC Berkeley, 549 Evans Hall, #3880, Berkeley, CA 94720-3880. dslee@econ.berkeley.edu. An earlier draft, "Trimming for Bounds on Treatment Effects with Missing Outcomes," is available online as NBER Technical Working Paper #277I thank Jim Powell, David Card, Guido Imbens, Jack Porter, Josh Angrist, Ed Vytlacil, Aviv Nevo, Jonah Gelbach, Doug Miller, David Autor, and participants of the UC Berkeley Econometrics and Labor Lunches, for helpful discussions and suggestions.

1 Introduction

Consider a version of Heckman's (1979) sample selection problem where one of the exogenous regressors D is a binary treatment variable:

$$\begin{aligned} Y^* &= D\beta + X\pi_1 + U \\ Z^* &= D\gamma + X\pi_2 + V \\ Y &= 1\left[Z^* \geq 0\right] \cdot Y^* \end{aligned} \tag{1}$$

where Y^* and Z^* are latent variables, and Y is the observed outcome. (U, V) are assumed to be jointly independent of the regressors (D, X). The parameter of interest is β, whose estimation would be straightforward, were it not for the non-random sample selection that causes missing data on Y^*.

There are two general approaches to addressing this problem. One is to explicitly model the process determining selection. In some cases, this may involve assuming that data are missing at random, perhaps conditional on a set of covariates (Rubin 1976) (e.g. conditional on X and D, U is independent of V). Alternatively, it involves assuming that some of the exogenous variables determine sample selection, but do not have its own direct impact on the outcome of interest (i.e. for some elements of X, elements of π_1 are zero while corresponding elements of π_2 are nonzero). Such exclusion restrictions are utilized in parametric and semi-parametric models of the censored selection process (Heckman 1979, 1990; Ahn and Powell 1993; Andrews and Schafgans 1998; Das, Newey, and Vella 2000).

An alternative approach is motivated by researchers' reluctance to rely upon specific exclusion restrictions. Boundedness of the outcome variable's support (e.g. $\underline{Y} \leq Y \leq \overline{Y}$) is used to construct "worst-case" bounds for the treatment effect parameter – bounds that are still consistent with the data that are observed. Horowitz and Manski (2000a) use this notion to provide a general framework for constructing bounds for treatment effect parameters when outcome and covariate data are non-randomly missing in an experimental setting. Others (Balke and Pearl 1997; Heckman and Vytlacil 1999, 2000a, 2000b) have constructed such bounds to address a different problem – that of imperfect compliance of the treatment,

even when "intention" to treat is effectively randomized (Bloom 1984; Robins 1989; Imbens and Angrist, 1994; Angrist, Imbens, and Rubin 1996). A limitation of these kinds of procedures is that when the support is unbounded, construction of such bounds are impossible without some further restriction on the sample selection process (Manski 1995). As a consequence, the bounds become increasingly wide as the support becomes large.

This paper shows that it is possible to bound β in the above model (and its extensions), without the need for exclusion restrictions, and even when Y has unbounded support. The procedure takes advantage of the independence of the error terms (U, V) and a monotonicity property implicitly contained in (1), whereby assignment to treatment impacts selection probabilities only in "one direction". The approach involves "trimming" the upper or lower tails of observed outcome distributions. The procedure is valid even after relaxing several assumptions implied in (1): the linear index structure of the regressors, and the implied constant treatment effects assumption, and full independence of (U, V). It yields the tightest bounds consistent with the observed data, and can be readily applied to the above model, its generalizations, matching estimation contexts, as well as to analyses of randomized experiments with non-random drop out.

It is shown below within a heterogeneous treatment effects framework that – without bounded support conditions – no bounds can be constructed for average treatment effects for two sub-populations: 1) the individuals whose outcomes will never be observed, and 2) the individuals who are induced to drop out (or remain in the sample) because of the treatment. Thus, average treatment effects are only bounded for those inframarginal individuals whose sample selection is not affected by the treatment. In some contexts, this may be precisely the parameter of interest.

The paper is organized as follows. Section 2 describes the nature of the problem and the basic idea in a familiar latent variable sample selection framework. Section 3 shows that the procedure is valid in a heterogeneous treatment effects model with non-random sample selection. Economic examples in which the above average treatment effect is the parameter of interest are provided. Section 4 describes how baseline covariates can be used to narrow the width of the bounds. Section 5 discusses some testable implications of the key restrictions of the model for trimming, and Section 6 concludes. Throughout this

paper, the treatment variable is assumed to be dichotomous, and always observed; hence, the analysis applies to censored and not truncated samples.

2 Trimming in the Sample Selection Problem

The usual approach to identifying β in Equation 1 is to recognize that the expected value of Y conditional on the regressors and sample selection can be written as

$$E\left[Y|D,X,Z^{*}\geq 0\right]=D\beta+X\pi_{1}+E\left[U|D,X,V\geq -D\gamma-X\pi_{2}\right] \tag{2}$$

Parametric models (e.g. Heckman 1979) specify a particular functional form for the selection correction term, and semi-parametric approaches (e.g. Ahn and Powell 1993) attempt to "partial out" the term in a non-parametric way. A standard assumption for identification is an exclusion restriction whereby some elements of X affect selection but not the outcome (i.e. some elements of π_1 are zero, while the corresponding elements in π_2 are nonzero).

When the researcher is reluctant to assert such exclusion restrictions, an alternative approach is to utilize properties of the support of Y to bound the parameter of interest. For example, if Y is known to always lie between $\underline{Y}$ and $\overline{Y}$, it is easy to see that an upper bound on $E\left[Y^{*}|D=1,X=x\right]$ is the weighted average

$$\left[1-G(-\gamma-x\pi_{2})\right]E\left[Y|D=1,X=x,Z^{*}\geq 0\right]+G\left(-\gamma-x\pi_{2}\right)\cdot\overline{Y} \tag{3}$$

where $G\left(\cdot\right)$ is the cdf of V (see Manski 1995). An analogous lower bound for $E\left[Y^{*}|D=0,X=x\right]$ can be constructed, which leads to an upper bound on the difference, or the treatment effect β. This strategy is discussed in detail, in Horowitz and Manski (2000a), who show the approach can be particularly useful, when Y is a binary outcome.

Naturally, this "worst-case scenario" imputation procedure cannot be used when the support of Y is unbounded. Even if the support is bounded, if its range is very large, these bounds may not be practically useful – except to highlight that further restrictions on the process are needed to produce tighter bounds (see Horowitz and Manski 2000b).

This paper shows that it is nonetheless possible to bound β when Y has unbounded support. To see how this is possible, assume w.l.o.g. that $\gamma > 0$, and note that

$$\begin{aligned} E[Y|D=0, X=x, Z^* \geq 0] &= x\pi_1 + E[U|D=0, X=x, V \geq -x\pi_2] \\ &= x\pi_1 + E[U|V \geq -x\pi_2] \end{aligned} \tag{4}$$

and

$$\begin{aligned} & E[Y|D=1, X=x, Z^* \geq 0] = \beta + x\pi_1 + E[U|D=1, X=x, V \geq -\gamma - x\pi_2] \\ = \; & \frac{G(-x\pi_2) - G(-\gamma - x\pi_2)}{1 - G(-\gamma - x\pi_2)} \{\beta + x\pi_1 + E[U| -\gamma - x\pi_2 \leq V \leq -x\pi_2]\} \\ & + \frac{1 - G(-x\pi_2)}{1 - G(-\gamma - x\pi_2)} \{\beta + x\pi_1 + E[U|V \geq -x\pi_2]\} \end{aligned}$$

The latter expression highlights the fact that the data observed for the treated population is a mixture of two distributions, representing two sub-populations: 1) the individuals whose outcomes would also have been observed even if they had been in the control group ($V \geq -x\pi_2$), and 2) the individuals whose outcomes are observed but would not have been had they been in the control state ($-\gamma - x\pi_2 \leq V \leq -x\pi_2$). In essence, there is a marginal sub-population that is induced to remain in the sample *because* of the treatment D.

If it were possible to delete the individuals in the latter group from the treatment population, the remaining individuals would be comparable to the entire control group. The difference in the averages would then eliminate the term $E[U|V \geq -x\pi_2]$. Without knowing specifically which individuals to delete, the most conservative approach would be to trim the lower tail of the distribution by the proportion in the marginal group $\left(\frac{G(-x\pi_2)-G(-\gamma-x\pi_2)}{1-G(-\gamma-x\pi_2)}\right)$ – a proportion that is identified by the data. Let $\overline{E}(x)$ denote the mean of the remainder of this distribution, after trimming. Clearly, the true value of $E[Y^*|D=1, X=x, V \geq -x\pi_2]$ cannot exceed $\overline{E}(x)$.

An analogous approach that trims the upper tail of the distribution will produce a lower bound $\underline{E}(x)$ for $E[Y^*|D=1, X=x, V \geq -x\pi_2]$, which produces the bound

$$\underline{E}(x) - E[Y|D=0, X=x, V \geq -x\pi_2] \leq \beta \leq \overline{E}(x) - E[Y|D=0, X=x, V \geq -x\pi_2] \tag{5}$$

since the control group's outcomes need not be trimmed. Since the bounds can vary by X, combining all of the X-specific bounds yields the overall bounds

$$\max_x \underline{E}(x) - E[Y|D=0, X=x, V \geq -x\pi_2] \leq \beta \leq \min_x \overline{E}(x) - E[Y|D=0, X=x, V \geq -x\pi_2] \quad (6)$$

which will be weakly narrower than any individual bound conditional on x.

There are several issues to consider with this trimming approach: 1) What aspects of Equation (1) are necessary for this procedure to be valid? 2) Does the procedure produce the tightest bounds consistent with the observed data? 3) In a heterogeneous treatment effects framework with Y unbounded, for which sub-populations can average treatment effects be bounded? These questions are addressed in the next section.

3 Missing Outcomes in a Heterogeneous Treatment Effect Model

This section generalizes the selection model in Equation (1) to allow for arbitrary heterogeneity in treatment effects. The discussion below begins by abstracting from the presence of covariates. This case is useful in the analysis of randomized experiments with non-random attrition.

Consider the following model of treatment effects with sample selection

Assumption S

$$\begin{aligned} & (Y_1^*, Y_0^*, S_1, S_0, D) \text{ is i.i.d. across individuals} \\ S &= S_1 D + S_0 (1-D) \\ Y &= S \cdot \{Y_1^* D + Y_0^* (1-D)\} \\ & (Y, S, D) \text{ is observed} \end{aligned} \quad (7)$$

D, S, S_0, and S_1 are all binary indicator variables. D denotes treatment status; S_1 and S_0 are "potential" sample selection indicators for the treated and control states. For example, when an individual has $S_1 = 1$ and $S_0 = 0$, this means the outcome Y will be observed if treatment is given, and will not be observed if treatment is denied. The second line highlights the fact that we observe the individual in only one or the other state. Y_1^* and Y_0^* are latent potential outcomes for the treated and control states, and the third line points out that we do not observe both outcomes for any given individual, and we do not observe

the outcome unless $S = 1$.

Assumption A

$$(Y_1^*, Y_0^*, S_1, S_0) \text{ is independent of } D \tag{8}$$

This assumption corresponds to the assumption of the independence of (U, V) in the previous section. This is a useful assumption to consider, particularly in the context of randomized experiments, since the random assignment will ensure this assumption will hold.

Furthermore, it is assumed that assignment to D, if it affects S at all, can affect S in only "one direction". This is expressed as

Assumption B

$$\Pr[S_1 = 0, S_0 = 1] = 0 \tag{9}$$

This assumption precludes the possibility that some individuals are induced to drop out of the sample because of the treatment. Note that imposing $\Pr[S_1 = 0, S_0 = 1] = 0$ rather than $\Pr[S_1 = 1, S_0 = 0] = 0$ is innocuous, as a parallel argument to that presented below is valid if the latter assumption is imposed instead. This assumption is analogous to the monotonicity assumption in studies of imperfect compliance of treatment (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996; Vytlacil 2000).

Assumptions A and B imply that the difference between the means of the sample-selected treatment and control groups is

$$\begin{aligned} & E[Y|D=1, S=1] - E[Y|D=0, S=1] \\ = \; & \frac{\Pr[S_0=0, S_1=1|D=1]}{\Pr[S=1|D=1]} E[Y_1^*|S_0=0, S_1=1] \\ & + \frac{\Pr[S_0=1, S_1=1|D=1]}{\Pr[S=1|D=1]} E[Y_1^*|S_0=1, S_1=1] \\ & - E[Y_0^*|S_0=1, S_1=1] \end{aligned} \tag{10}$$

In general, this will be biased for the average treatment effect $E[Y_1^* - Y_0^*|S_0 = 1, S_1 = 1]$. While the weights $\frac{\Pr[S_0=0,S_1=1|D=1]}{\Pr[S=1|D=1]}$ and $\frac{\Pr[S_0=1,S_1=1|D=1]}{\Pr[S=1|D=1]}$ can be identified from the observed data, $E[Y_1^*|S_0 = 0, S_1 = 1]$

and $E[Y_1^*|S_0 = 1, S_1 = 1]$ cannot be identified without further restrictions.

It is possible, however, without further restrictions to construct upper and lower *bounds* $\overline{E}$ and $\underline{E}$ such that $\underline{E} \leq E[Y_1^*|S_0 = 1, S_1 = 1] \leq \overline{E}$. It then follows that there exist bounds such that

$$\underline{E} - E\left[Y|D = 0, S = 1\right] \leq E\left[Y_1^* - Y_0^*|S_0 = 1, S_1 = 1\right] \leq \overline{E} - E\left[Y|D = 0, S = 1\right] \tag{11}$$

for the average treatment effect for this subpopulation.

$\underline{E}$ and $\overline{E}$ are constructed as in the following proposition:

Proposition 1 *Suppose Assumptions S, A, and B hold, and* $\Pr\left[S = 1|D = 0\right] \neq 0$. *Denote the observed cumulative distribution of* Y, *conditional on* $D = 1, S = 1$ *as* $F\left(y\right)$. *Then*

$$\underline{E} \equiv \frac{1}{1-p}\int_{-\infty}^{F^{-1}(1-p)} ydF\left(y\right) \leq E\left[Y_1^*|S_0 = 1, S_1 = 1\right]$$

and

$$\overline{E} \equiv \frac{1}{1-p}\int_{F^{-1}(p)}^{\infty} ydF\left(y\right) \geq E\left[Y_1^*|S_0 = 1, S_1 = 1\right]$$

where

$$p = \frac{\Pr\left[S = 1|D = 1\right] - \Pr\left[S = 1|D = 0\right]}{\Pr\left[S = 1|D = 1\right]}$$

Also, $\underline{E}$ $\left(\overline{E}\right)$ *is equal to the smallest (largest) possible value for* $E\left[Y_1^*|S_0 = 1, S_1 = 1\right]$ *that is consistent with the distribution of observed data on* (Y, S, D).

Intuitively, we know that a fraction p of the observed distribution of Y for the selected, treated group needs to be deleted, in order to allow a valid comparison to the outcomes of the control group. $\underline{E}$ is constructed by calculating means after truncating the upper tail of distribution, and $\overline{E}$ is computed by truncating the lower tail – by the proportion p.

Given Assumption B, $E\left[Y_0^*|S_0 = 1, S_1 = 1\right]$ equals $E\left[Y|D = 0, S = 1\right]$; no trimming is necessary for the control group.

Corollary 2 *Given Assumptions S, A, and B and* $\Pr\left[S = 1|D = 0\right] \neq 0$

$$\underline{E} - E\left[Y|D = 0, S = 1\right] \leq E\left[Y_1^* - Y_0^*|S_0 = 1, S_1 = 1\right] \leq \overline{E} - E\left[Y|D = 0, S = 1\right]$$

where the lower bound (upper bound) is equal to the smallest (largest) possible value for the average treatment effect, $E[Y_1^* - Y_0^*|S_0 = 1, S_1 = 1]$, *that is consistent with the distribution of the observed data on* (Y, S, D).

Both Assumptions A and B are crucial to the result. Monotonicity ensures that the subpopulation of the control group for whom we observe outcomes consists only of those for whom $S_0 = 1, S_1 = 1$,

those whose sample selection is not affected by the treatment. Without monotonicity, the control and treatment groups could, in principle, consist solely of the sub-populations for whom $S_0 = 1, S_1 = 0$ and $S_0 = 0, S_1 = 1$, respectively. This would imply no "overlap" between the two sub-populations, making it impossible to make a comparison that could be interpreted as a causal effect. The independence assumption is also important, since it is what justifies the contrast between the trimmed treatment group and the control group. Thus the assumptions of the model in Equation 1 that are crucial to the result are the independence of (U, V) and the latent index, threshold-crossing structure of sample selection (i.e. monotonicity).

Below are two economic examples in which the parameter $E\left[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1\right]$ is of economic interest. In the first, the monotonicity condition could be expected to hold, and in the second, economic reasoning suggests that monotonicity would probably not hold.

Example 1 *Labor supply with a Negative Income Tax, experimental variation in tax rate*

Consider a static labor supply setting, where we are interested in the *intensive* margin response of hours of work to a change in marginal tax rates. Subjects are randomized into treatment and control groups. Both groups are given the same guaranteed income subsidy of G, which is taxed away at rates t_t and $t_c (> t_t)$, for the treated and control, respectively. Suppose we are interested in the average treatment effect of the tax rate on Y, the natural logarithm of hours worked. Since the interest is in the intensive margin labor supply response, the focus is naturally on the population that will work positive hours, irrespective of treatment status. It is natural to expect that the treatment (a higher effective wage) to induce some individuals to work, causing a potential sample selection bias.

Under the assumption of optimizing behavior given a complete, transitive, and strictly monotone preference relation over leisure and consumption (l and c), any consumer who would work positive hours facing tax rate t_c would work positive hours facing t_t. To see this, consider any individual who works positive hours under t_c. Denote the optimal hours as $\exp\left(Y_0^*\right) = h_0 > 0$. The bundle of consumption and leisure $\left(G + w\left(1 - t_t\right) h_0, T - h_0\right)$, (where T is total time available), which is a bundle that is feasible given the treatment, is strictly preferred to $\left(G + w\left(1 - t_c\right) h_0, T - h_0\right)$, which itself is preferred to (G, T)

(the bundle attained by not working) by hypothesis. By transitivity, (G, T) cannot be the optimal choice for the consumer facing t_t.

Thus, in this economic context, the monotonicity assumption (**B**) is rationalized by optimizing behavior given a fairly standard preference relation. The trimming procedure described above can be used to generate bounds on the percentage change in hours of labor supply induced by a marginal tax rate reduction, accounting for the presence of non-random sample selection that results from labor supply behavior on the extensive margin of employment.

Example 2 *Labor supply with a Negative Income Tax, experimental variation in tax rate* and *guaranteed subsidy*

Consider the same setting as above, except that in addition to different tax rates, different levels of the guaranteed subsidy $G_t > G_c$ are offered to the treatment and control groups, respectively. Again, consider the control group individual who optimally chooses positive hours of work by choosing the combination $(G_c + w(1 - t_c) h_0, T - h_0)$. Without further information about preferences, we cannot rule out the possibility that (G_t, T) is strictly preferred by this individual, and that it would have been the optimal choice under the treatment assignment. In other words, we cannot rule out the possibility that treatment *induces* some individuals to stop working. We also cannot rule out that the treatment induces other individuals to work positive hours (in other words, that $(G_t + w(1 - t_t) h_1, T - h_1)$, $h_1 > 0$, is preferred to (G_t, T) (which, in turn, is preferred to (G_c, T)). In this example, economic reasoning *cannot* be used to justify Assumption **B**.

An important implication of Assumptions A and B is that as p vanishes, so does the sample selection bias.[1] The intuition is that if $p = 0$, then under the monotonicity assumption, the population with observed outcome data – whether in the treatment or control group – is comprised of individuals whose *sample selection* was unaffected by the assignment to treatment (those for whom $S_0 = 1$, and $S_1 = 1$). These individuals can be thought of as the "always-takers" sub-population (Angrist, Imbens, and Rubin 1996), except that "taking" is not the taking of the treatment, but rather selection into the sample.

[1] A vanishing p corresponds to individuals with the same value of the sample selection correction term. See, for example, (Heckman and Robb 1986; Heckman 1990; Ahn and Powell 1993; Angrist 1997).

One example of a practical implication of this is that when analyzing randomized experiments, if the "drop-out" rates in the treatment and control groups are similar, and if the monotonicity condition is believed to hold, then a comparison of the treatment and control means is a valid estimate of an average treatment effect. Note that p here is proportional to the *difference* in the fraction that are sample selected between the treatment and control groups, and so this observation should not be confused with "identification at infinity" in Heckman (1990), in which the bias term vanishes as the fraction that is sample selected tends to unity.

It is instructive to highlight the primary features of the proposed trimming procedure that distinguish it from existing bounds approaches in the literature. First, the model and procedure proposed here can produce finite bounds when the outcome has unbounded support. This should be contrasted to a method that addresses missing outcomes by essentially assigning the values of upper and lower bounds of support to missing data to bound parameters of interest (Horowitz and Manski 1998, 2000a).

This advantage of trimming, however, does not come without a cost. The second distinctive feature (and disadvantage) of the model proposed above is that it relies crucially on an unverifiable assumption about the selection process. For example, the model assumes that *every* control (treatment) group individual who reported an outcome would have reported outcome if they had been assigned treatment (to the control group) – a conjecture that simply cannot be verified one way or another. The appropriateness of this "monotonicity" assumption may or may not be "plausible" depending on the particular application, as illustrated in the economic examples above.

4 Trimming with Covariates

The trimming procedure can easily be extended to the case where other covariates are available to the researcher. In the context of randomized experiments, these covariates are typically used to assess whether or not the randomization "failed", and if successful randomization is not rejected by the data the covariates are often included in the analysis to reduce the sampling variability of the estimates. With non-experimental matching methods, it is sometimes assumed that treatment is "as good as randomly assigned",

conditional on the covariates.

Suppose there exists a vector of baseline covariates X, where each element has discrete support, so that this vector can take on one of a finite number of discrete values. Focus on the values $\{x_1, \ldots, x_J\}$, such that for each $j = 1, \ldots, J$, $\Pr(X = x_j | D = 0, S = 1) \neq 0$.

Assumption C

$$(Y_1^*, Y_0^*, S_1, S_0, X) \text{ is independent of } D \tag{12}$$

Assumption C would hold if D were randomly assigned, and X were pre-determined, relative to the point of random assignment.

Under this assumption, an upper (lower) bound for $E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1]$ can be constructed by trimming the lower (upper) tails of distributions of Y, conditional on $D = 1$ and X, by a proportion given by $p_j = \frac{\Pr[S=1|D=1,X=x_j] - \Pr[S=1|D=0,X=x_j]}{\Pr[S=1|D=1,X=x_j]}$. The overall mean of the truncated distributions of the sub-groups of the treated is computed by averaging across values of X.

Proposition 3 *Suppose Assumptions S, B, and C hold, covariates X are observed, and* $\Pr[S = 1|D = 0] \neq 0$. *Denote the observed cumulative distribution of* Y, *conditional on* $D = 1$, $S = 1$, *and* $X = x_j$, *as* $F(y|x_j)$. *Then*

$$\underline{E}^* \equiv \sum_{j=1}^{J} \Pr[X = x_j | S = 1, D = 0] \frac{1}{1 - p_j} \int_{-\infty}^{F^{-1}(1-p_j|x_j)} y dF(y|x_j)\, dy \leq E[Y_1^* | S_0 = 1, S_1 = 1]$$

and

$$\overline{E}^* \equiv \sum_{j=1}^{J} \Pr[X = x_j | S = 1, D = 0] \frac{1}{1 - p_j} \int_{F^{-1}(p_j|x_j)}^{\infty} y dF(y|x_j)\, dy \geq E[Y_1^* | S_0 = 1, S_1 = 1]$$

where

$$p_j = \frac{\Pr[S = 1|D = 1, X = x_j] - \Pr[S = 1|D = 0, X = x_j]}{\Pr[S = 1|D = 1, X = x_j]}$$

Also, $\underline{E}^* \left(\overline{E}^*\right)$ *is equal to the smallest (largest) possible value for* $E[Y_1^* | S_0 = 1, S_1 = 1]$ *that is consistent with the distribution of observed data on* (Y, S, D, X)

Corollary 4 *Given Assumptions S, B, and C, and observed covariates X, and* $\Pr[S = 1|D = 0] \neq 0$

$$\underline{E}^* - E[Y|D = 0, S = 1] \leq E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1] \leq \overline{E}^* - E[Y|D = 0, S = 1]$$

where the lower bound (upper bound) is the smallest (largest) possible value for the average treatment effect, $E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1]$, *that is consistent with the distribution of the observed data on* (Y, S, D, X).

Intuitively, Assumption C implies that the assumptions used to justify the trimming procedure will also justify trimming, conditional on X. Given bounds for $E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, X = x_j]$, it is

possible to average across values of X to produce bounds for $E\left[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1\right]$.

The motivation for this modified trimming procedure is that using the covariates in this way will lead to tighter bounds on the treatment effect parameter of interest.

Proposition 5 *If Assumptions S, B, and C hold, and given covariates X, and $\Pr\left[S = 1 | D = 0\right] \neq 0$, then $\underline{E}^* \geq \underline{E}$ and $\overline{E}^* \leq \overline{E}$.*

Intuitively, this is true because a lower-tail truncated mean of a distribution will always be larger than the average of lower-tail truncated means of sub-groups of the population, provided that the proportion of the entire population that is eventually truncated remains fixed. An implication of Proposition 5 is that in general, using more baseline covariates will lead to tighter bounds on $E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1]$.

Finally, consider the weaker assumption

$$(Y_1^*, Y_0^*, S_1, S_0) \text{ is independent of } D, \text{ conditional on } X \tag{13}$$

This weakens Assumption C by allowing D to be unconditionally correlated with the potential outcomes. It is assumed that conditional on X, treatment-control contrasts yield average treatment effects conditional on X. This weaker assumption is the basis of matching estimators for evaluation studies. Note that in the latent variable model in Equation 1, this weaker assumption implies that X need not be independent of (U, V).

It is easy to see that Proposition 1 can be applied for each value of X. So for each value of X, there will be an upper and lower bound for $E[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1, X = x_j]$. Therefore, upper and lower bounds for any weighted average of these effects, can be constructed by computing corresponding weighted averages of the upper and lower bounds.

5 Testable Implications

While it is clear that the assumptions of the model proposed above are fundamentally unverifiable, it is important to examine whether the restrictions generate any testable implications, however weak they might be. As is well known, the independence assumption (C), which corresponds to random assignment, has the implication that the baseline pre-determined characteristics X be distributed identically between the

treatment and control groups.

The monotonicity assumption (B) is restrictive enough to generate a testable restriction. In particular, Assumption **B** implies that there exists no j, such that $\Pr[S=1|D=1,X=x_j] < \Pr[S=1|D=0,\ X=x_j]$. Essentially, the monotonicity restriction is inconsistent with the existence of j' and j'' such that $\Pr[S=1|D=1,X=x_{j'}] < \Pr[S=1|D=0,\ X=x_{j'}]$ while at the same time $\Pr[S=1|D=1,X=x_{j''}] > \Pr[S=1|D=0,X=x_{j''}]$.

Finally, suppose $\Pr[S=1|D=1] = \Pr[S=1|D=0]$. As mentioned earlier, in this case, Assumptions B and C imply that there is no sample selection bias, and that a simple contrast between $E[Y|D=1,S=1] - E[Y|D=0,S=1]$ is valid for identifying a meaningful causal parameter. $0 = \Pr[S=1|D=1] - \Pr[S=1|D=0] = \sum_j^J \{\Pr[X=x_j|D=0]\ (\Pr[S=1|D=1,X=x_j] - \Pr[S=1|\ D=0,X=x_j])\}$ because of Assumption C. Since $\Pr[X=x_j|D=0] > 0$ for all $j=1,\ldots,J$, and Assumption B implies that $\Pr[S=1|D=1,X=x_j] - \Pr[S=1|D=0,X=x_j] \geq 0$ for $j=1,\ldots,J$, then it must be true that $\Pr[S=1|D=1,X=x_j] - \Pr[S=1|D=0,X=x_j] = 0$ for $j=1,\ldots,J$. It can then be shown, using Assumption C and Bayes' Rule, that this implies $\Pr[X=x_j|S=1,D=1] = \Pr[X=x_j|S=1,D=0]$ for $j=1,\ldots,J$. Therefore, if $\Pr[S=1|D=1] = \Pr[S=1|D=0]$, then Assumptions B and C imply that the distributions of the baseline covariates between the selected treatment group and the selected control group are identical, which is testable given the observed data.

6 Conclusions and Extensions

In many situations, researchers may be willing to entertain the possibility that treatments are "as good as randomly assigned" but are at the same time considerably less confident about the underlying process that determines whether outcomes are missing. A potentially useful alternative to specifying exclusion restrictions is a bounding analysis that generates "worst-case" sample selection biases. In the context of outcomes with essentially unbounded support, existing nonparametric bounding approaches (e.g. Horowitz and Manski 1998, 2000a) of unbounded outcomes immediately suggest there will be no finite bounds on average treatment effects. This can be informative in the sense that it suggests that *any* finite bounds on

treatment effects in this context will necessarily be a consequence of some further stochastic restriction on the data generating process (Horowitz and Manski 2000b). The question then becomes Which restrictions have relatively large benefits and/or small costs?

This paper has proposed a simple and intuitive trimming procedure that is justified under the added restriction of monotonicity of the censored selection process. The main benefit from imposing the monotonicity restriction is that it allows one to generate finite bounds even when the outcome variable has unbounded support. The main cost of the restriction is that such a behavioral assumption may or may not be plausible, depending on the particular context of the selection problem. This paper has described two economic contexts: one in which the monotonicity assumption could be considered plausible, and another where economic reasoning suggests the assumption is unwarranted.

The main result is particularly relevant to the analysis of randomized experiments that have been corrupted by non-random drop out. In these cases, independence is a justifiable assumption. The results described above imply that if the sample selection process is modeled with a latent-index threshold-crossing structure, the trimming bounds proposed above are the *tightest* consistent with the observed data. This implies that any additional restriction must necessarily (weakly) tighten the bounds further.[2]

The following are potentially useful avenues for future research. First, it would be interesting to apply the proposed trimming procedure to appropriate applied contexts, and to compare the bounds to estimates obtained from other parametric and semi-parametric modeling approaches and other bounding procedures. Second, since the number of baseline covariates may be so large as to create a "small cell" problem, it would be helpful to generalize the procedure to utilize continuous covariates. Third, it seems possible to generalize the procedure in various directions. For example, it could be extended to apply to 1) the case of an endogenous regressor of interest with a valid instrument (or imperfect compliance of a treatment whose "intention-to-treat" is randomized), 2) the case of a continuous treatment variable, or 3)

[2] For example, in an analysis of the STAR class-size experiment, Krueger and Whitmore (2001) make two assumptions: 1) monotonicity of the sample selection process, and 2) those induced to be selected score lower than those who are sample selected irrespective of treatment. Since there is an extra restriction here (the second one) any consequent sharp bound cannot be larger than the estimates obtained from what they call a "linear truncation" procedure. Apparently, the last set of caluculations in the paper are not sharp bounds.

the case of more than one sample selection process (e.g. sample attrition as distinct from the labor force participation decision). Finally, it would be interesting to explore what additional plausible assumptions, beyond the monotonicity restriction, would lead to tighter bounds on average treatment effects.

Appendix A.

Lemma 6 *Suppose Y is a mixture of two random variables. Let its cdf $F^*(y) = p^* M^*(y) + (1-p^*) N^*(y)$, where $M^*(y)$ and $N^*(y)$ are both cdfs, and $p^* \in [0,1)$ is fixed. Consider the cdf of the truncated version of $F^*(y)$, $G^*(y) = \max\left[0, \frac{F^*(y)-p^*}{1-p^*}\right]$. Then $\int_{-\infty}^{\infty} y dG^*(y) \geq \int_{-\infty}^{\infty} y dN^*(y)$. $\int_{-\infty}^{\infty} y dG^*(y)$ is a sharp upper bound for $\int_{-\infty}^{\infty} y dN^*(y)$.*

Proof of Lemma 6. See Horowitz and Manski (1995), Corollary 4.1.

Proof of Proposition 1. Assumption A and B implies that $p = \frac{\Pr[S=1|D=1]-\Pr[S=1|D=0]}{\Pr[S=1|D=1]} = \frac{\Pr[S_0=0,S_1=1|D=1]}{\Pr[S=1|D=1]}$. p is strictly less than 1 by assumption. Assumption B also implies that $F(y) = pM(y) + (1-p)N(y)$, where $M(y)$ denotes the cdf of Y_1^*, conditional on $D=1$, $S_0=0$, $S_1=1$, and $N(y)$ denotes the cdf of Y_1^*, conditional on $D=1$, $S_0=1$, $S_1=1$. By Assumption A, $N(y)$ is also the cdf of Y_1^*, conditional on $S_0=1, S_1=1$. By Lemma 6, $\overline{E} \equiv \frac{1}{1-p}\int_{F^{-1}(p)}^{\infty} y dF(y) \geq \int_{-\infty}^{\infty} y dN(y) = E[Y_1^*|S_0=1, S_1=1]$.

To show that $\overline{E}$ equals the maximum possible value for $E[Y_1^*|S_0=1,S_1=1]$ that is consistent with the distribution of the observed data on (Y,S,D), it must be shown that 1) conditional on p, $\overline{E}$ is a sharp upper bound, and 2) p is uniquely determined by the data. 1) follows from 6. 2) is true because the data yield a unique probability function $\Pr[S=s, D=d]$, $s,d=0,1$, which uniquely determines p.

An argument parallel to that made above can be made for $\underline{E}$. Q.E.D.

Proof of Proposition 3. Given Assumption C, this implies that Assumption A holds, conditionally on X. It is given that for each j, $\Pr[X=x_j|D=0,S=1] \neq 0$. So $\Pr[S=1|D=0] \neq 0$ implies, using Bayes' Rule, that $\Pr[S=1|D=0,X=x_j] \neq 0$ for all $j=1,\ldots,J$. Thus, by the Proposition 1, it can be shown that $\frac{1}{1-p_j}\int_{F^{-1}(p_j|x_j)}^{\infty} y dF(y|x_j) \geq E[Y_1^*|S_0=1,S_1=1,X=x_j]$ for $j=1,\ldots,J$. It follows that $\overline{E}^* \geq \sum_{j=1}^{J} \Pr[X=x_j|S=1,D=0] E[Y_1^*|S_0=1,S_1=1,X=x_j]$. The latter quantity equals $\sum_{j=1}^{J}\{\Pr[X=x_j|S_0=1,S_1=1]E[Y_1^*|S_0=1,S_1=1,X=x_j]\} = E[Y_1^*|S_0=1,S_1=1]$ by Assumptions B and C.

To show that $\overline{E}^*$ is equal to the largest possible value for $E[Y_1^*|S_0=1,S_1=1]$ that is consistent with the distribution of observed data on (Y,S,D,X), it must be shown that 1) conditional on $X=x_j$, p_j and

$S=1$, $D=1$, $\frac{1}{1-p_j}\int_{F^{-1}(p_j|x_j)}^{\infty} ydF\left(y|x_j\right)$ is the sharp upper bound for $E\left[Y_1^*|S_0=1,S_1=1,X=x_j\right]$, and 2) the p_j's are uniquely determined by the data. 1) follows from applying 6 conditionally on X. 2) follows because the data yields a unique probability function $\Pr\left[S=s,D=d|X=x_j\right], s,d=0,1$, which yields unique values of p_j.

An argument parallel to that made above can be made for $\underline{E}^*$.Q.E.D.

Proof of Proposition 5. $F\left(y\right)$, the distribution of Y conditional on $D=1$, $S=1$ can always be written as $\sum_{j=1}^{J}\Pr[X=x_j|\ S=1,D=1]\cdot p_j\min\left[\frac{F(y|x_j)}{p_j},1\right]+\sum_{j=1}^{J}\Pr[X=x_j|\ S=1,D=1]\cdot(1-p_j)$ $\max\left[\frac{F(y|x_j)-p_j}{1-p_j},0\right]\}$. It needs to be shown that the second term is equal to $(1-p)\,N^*\left(y\right)$, where $N^*\left(y\right)$ is the distribution $\sum_{j=1}^{J}\Pr\left[X=x_j|S=1,D=0\right]\max\left[\frac{F(y|x_j)-p_j}{1-p_j},0\right]$, which is simply the distribution from which $\overline{E}^*$ is calculated in Proposition 5, and p is the proportion from Proposition 1. If this is true, the Lemma 6 applies, implying $\overline{E}$ is a sharp upper bound for $\overline{E}^*$.

It therefore is sufficient to show that $\Pr[X=x_j|\ S=1,D=1]\cdot(1-p_j)=(1-p)\Pr\left[X=x_j|S=1,D=0\right]$ for all $j=1,\ldots,J$. Re-arranging and using independence, it can be shown that $\Pr[X=x_j|\ S=1,D=1]$ $\cdot\,(1-p_j)=\frac{\Pr[X=x_j,S=1,D=1]}{\Pr[S=1,D=1]}\cdot\frac{\Pr[S=1|D=0,X=x_j]}{\Pr[S=1|D=1,X=x_j]}=\frac{\Pr[X=x_j,S=1,D=1]}{\Pr[S=1,D=1]}\cdot\frac{\Pr[S=1,D=0,X=x_j]\Pr[D=1,X=x_j]}{\Pr[S=1,D=1,X=x_j]\Pr[D=0,X=x_j]}=$ $\frac{\Pr[S=1,D=0,X=x_j]\Pr[D=1]}{\Pr[S=1,D=1]\Pr[D=0]}=\frac{\Pr[S=1,D=0]\Pr[D=1]}{\Pr[S=1,D=1]\Pr[D=0]}\cdot\frac{\Pr[X=x_j,S=1,D=0]}{\Pr[S=1,D=0]}=\frac{\Pr[S=1|D=0]}{\Pr[S=1|D=1]}\cdot\frac{\Pr[X=x_j,S=1,D=0]}{\Pr[S=1,D=0]}$ $=(1-p)\cdot\frac{\Pr[X=x_j,S=1,D=0]}{\Pr[S=1,D=0]}=(1-p)\cdot\Pr\left[X=x_j|S=1,D=0\right]$. Q.E.D.

References

[1] Andrews, D., and Schafgans, M. (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497-517.

[2] Ahn, H. and Powell, J. (1993), "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3-29.

[3] Angrist, J., Imbens, G., and Rubin, D. (1996) "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-445.

[4] Angrist, J., (1997) "Conditional Independence in Sample Selection Models," *Economics Letters*, 54, 103-112.

[5] Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects from Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171-1177.

[6] Bloom, H. (1984), "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 8, 225-246.

[7] Das, M., Newey, W. K., and Vella, F. (2000), "Nonparametric Estimation of Sample Selection Models", mimeo.

[8] Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.

[9] Heckman, J. J. and R. Robb (1986), "Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes," in H. Wainer, ed., *Drawing inferences from self-selected samples* (Springer, New York).

[10] Heckman, J. J. (1990), "Varieties of Selection Bias," *American Economic Review Papers and Proceedings*, 80, 313-318.

[11] Heckman, J. J., and Vytlacil, E., (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

[12] Heckman, J. J., and Vytlacil, E., (2000a), "Local Instrumental Variables," *National Bureau of Economic Research Technical Working Paper #252*.

[13] Heckman, J. J., and Vytlacil E., (2000b), "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," *National Bureau of Economic Research Technical Working Paper #259*.

[14] Horowitz, Joel L & Manski, Charles F, 1995. "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, Vol. 63 (2) pp. 281-302

[15] Horowitz, J. L., and Manski, C. F. (1998), "Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37-58.

[16] Horowitz, J. L., and Manski, C. F. (2000a) "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data" *Journal of the American Statistical Association*, 95, 77-84.

[17] Horowitz, J. L., and Manski, C. F. (2000b) Rejoinder: "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data" *Journal of the American Statistical Association*, 95, 87.

[18] Imbens, G., and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62 (4): 467-476.

[19] Krueger, Alan, and Whitmore, D. (2001) "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR", *Economic Journal*, January 2001, 1-28.

[20] Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press.

[21] Robins, J. (1989), "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, DC: U.S. Public Health Service.

[22] Rubin, D. (1976), "Inference and Missing Data" *Biometrika*, 63, 581-592.

[23] Vytlacil, E. (2000), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result" *mimeo*.