CENTER FOR LABOR ECONOMICS
UNIVERSITY OF CALIFORNIA, BERKELEY
WORKING PAPER NO. 51

Trimming for Bounds on Treatment Effects with Missing Outcomes [*]

David S. Lee

UC Berkeley and NBER

March 2002

Abstract

Even if there is perfect compliance of a treatment that is randomly assigned, identification of average treatment effects is not straightforward when outcome data are missing in a non-random way. There already exist simple procedures that assume bounded support of the outcome in order to generate bounds of the effects. These procedures suggest that when the outcome has unbounded (or very large support), narrower bounds must necessarily rely on further stochastic restrictions. This paper describes how imposing a monotonicity restriction on the censored sample selection process permits a bounding analysis of treatment effects with missing outcomes, when the outcome has unbounded support. The model implies that a simple and intuitive trimming procedure will yield the tightest bounds (given the model) on treatment effects that are consistent with the observed data. The inclusion of baseline covariates in the analysis tightens the bounds further.

[*] I thank Jim Powell, David Card, Guido Imbens, Jack Porter, and Aviv Nevo for helpful discussions.

1 Introduction

In the treatment evaluation problem, even when it is stipulated that there is perfect compliance of a treatment that is “as good as randomly assigned”, identifying average causal effects for a population of interest is still not straightforward if outcome data are unobserved in a non-random way. In some cases, outcome data is “missing”, for example, due to non-response or sample attrition. In other cases, outcomes may not even be well-defined for the entire population. For example, hourly or weekly wages are not defined for the non-working (Heckman, 1974). When the process determining observability of the outcome is related to the assignment of treatment, an analysis that ignores the selection process will in general yield biased estimates of the treatment effects of interest (Heckman, 1979).

There are two general strategies for dealing with the problem. One is to explicitly model the selection process. In some cases, it may involve assuming that data are missing at random, perhaps conditional on a set of covariates (Rubin 1976). Other times, it may involve assuming the existence of exogenous variables that determine selection, but do not have its own direct impact on the outcome of interest. Such an exclusion restriction is often 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). A second strategy is to focus on information about the support of the outcome variable in order 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).

This paper shows how certain assumptions about the sample selection process allows one to “trim” observed distributions of data in order to yield informative bounds on average treatment effects, in the presence of non-random missing outcome data. The key assumptions are 1) random assignment of the treatment, and 2) a “monotonicity” condition whereby treatment assignment impacts sample selection only in “one direction”. If the treatment (control) group has a higher proportion of non-missing outcome data, these two assumptions imply that the observed outcome distribution for the treatment (or control) group is a mixture of two groups. One group possesses outcomes that can properly be contrasted to the control (treatment) group, and the other group was induced to “select into the sample” because of the assignment to treatment (control). The idea is to trim the lower or upper tails of the observed outcome distribution of the treatment (control) group by the proportion that belong to that latter group; this proportion is identified from the data. Under these assumptions, this should yield upper and lower bounds for the mean outcome for the former group.

There are three distinctive aspects of the trimming approach proposed here. First, it has the potential to produce informative bounds even when the outcome variable has unbounded (or very large) support. Second, the bounds crucially rely on particular “monotonicity” assumption, so it is not virtually “assumption-free” as is the approach of Horowitz and Manski (2000a), who only stipulate random assignment of the treatment, and utilize boundedness of supports to produce bounds on treatment effects. Third, the trimming procedure produces bounds for the average treatment effect for a very particular sub-population: those whose outcomes will be observed, irrespective of the assignment to treatment. No bounds are produced for the average treatment effect for other sub-populations of interest. Throughout this paper, the treatment variable is assumed to be dichotomous, and always observed; hence, the analysis applies only to censored and not truncated samples.

The paper is organized as follows. Section 2 describes the basic model and trimming procedure. Section 3 describes how baseline covariates can be used narrow the width of the bounds. Section 4 contrasts the trimming procedure to an imputation procedure for producing bounds. Section 5 discusses some testable implications of the key restrictions of the model for trimming, and Section 6 concludes.

2 Missing Outcomes in a Heterogeneous Treatment Effect Model

I begin by outlining conditions under which a trimming approach can produce bounds for average treatment effects for a particular sub-population of interest. Consider the random variables $(Y_1^*, Y_0^*, S_1, S_0, D)$ where Y_1^* and Y_0^* are continuous potential outcomes of interest when $D = 1$ and $D = 0$, respectively. S_1 and S_0 denote whether the outcome is observed when $D = 1$ and $D = 0$, respectively. For example, the realization $S_1 = 1, S_0 = 0$ implies that the outcome would be observed if $D = 1$, but would be missing if $D = 0$. (Y, S, D) is observed, where $Y = Y_1^* D + Y_0^* (1 - D)$ if $S = 1$, Y is missing if $S = 0$; $S = S_1 D + S_0 (1 - D)$. Y_1^* and Y_0^* are never simultaneously observed, and S_1 and S_0 are never simultaneously observed.

Assumption A

$$(Y_1^*, Y_0^*, S_1, S_0) \text{ is independent of } D \tag{1}$$

This assumption corresponds to the random assignment of D. It is useful to consider this assumption, as it means that any bias in identifying average treatment effects will be due to censored selection, rather than to the usual confounding problem. Furthermore, it is assumed that assignment to D, if it affects S at all, can affect S in only "one direction". This is a "monotonicity" assumption.

Assumption B

$$\Pr [S_1 = 0, S_0 = 1] = 0 \tag{2}$$

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

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

$$E [Y|D = 1, S = 1] - E [Y|D = 0, S = 1] \tag{3}$$

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

In general, this will be biased for a particular parameter of interest: $E\left[Y_1^*-Y_0^*|S_0=1,S_1=1\right]=E[Y_1^*|S_0=1,S_1=1]-E[Y_0^*|S_0=1,S_1=1]$, the average treatment effect for the subpopulation whose outcome data will be observed irrespective of treatment status. 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, $E\left[Y_1^*|S_0=0,S_1=1\right]$ and $E[Y_1^*|S_0=1,S_1=1]$ cannot be identified without some further restrictions.

However, if the observed data can yield upper and lower bounds $\overline{E}$ and $\underline{E}$ such that $\underline{E}\leq E[Y_1^*|S_0=1,S_1=1]\leq\overline{E}$ then there would exist bounds

$$\underline{E}-E\left[Y|D=0,S=1\right]\leq E\left[Y_1^*-Y_0^*|S_0=1,S_1=1\right]\leq\overline{E}-E\left[Y|D=0,S=1\right] \tag{4}$$

for the average treatment effect for this subpopulation.

The approach in this paper is to construct these bounds by trimming the lower or upper tails of the observed distribution of Y for the treatment group, by a proportion given by $\frac{\Pr[S=1|D=1]-\Pr[S=1|D=0]}{\Pr[S=1|D=1]}$: the proportion of the selected treatment group that is induced to have a non-missing value of the outcome because of the assignment to treatment.

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

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

and

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

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).

Given Assumption B, $E\left[Y_0^*|S_0=1,S_1=1\right]$ equals $E\left[Y|D=0,S=1\right]$, which can be computed from the observed data from the control group.

Corollary 2 *Given Assumptions A and B 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 $E[Y_1^*-Y_0^*|S_0=1,S_1=1]$ *that is consistent with the distribution of the observed data on* (Y,S,D).

The "monotonicity" assumption is crucial to this approach. It ensures that subpopulation of the control group for whom we observe outcomes consists only of those for whom $S_0=1,S_1=1$ – that is, those who will always have non-missing outcome data, irrespective of the assignment to treatment. The independence assumption is also important, since it is what justifies the contrast between the trimmed population of the treatment group and the control group.

An implication of Assumptions A and B is that as p vanishes, so does the sample selection bias. The intuition is simply 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 selection into the sample.

It should be noted that the bounds are only informative insofar as $\Pr[S=1|D=0]\neq 0$. Otherwise, the distribution of observed outcome data for the treatment group would be *completely* truncated, leaving no remaining data.

It should also be noted that these bounds can also be useful in typical latent-variable formulations of the sample selection process (Heckman 1979). Consider the system of equations

$$\begin{aligned} Y^* &= \beta_0+\beta_1 T+U \\ Z^* &= \gamma_0+\gamma_1 T+V \end{aligned} \tag{5}$$

where Y^* is an outcome of interest, T takes on the values 0 or 1, β_1 is the treatment effect of interest. Y is observed and equals Y^* if $Z^*\geq 0$, but is missing if $Z^*\leq 0$. It is often assumed (for example, in maximum likelihood estimation of parametric selection models) that (U,V) is independent of T. In addition, if $\gamma_1\geq 0$, then it is possible to use the bounds proposed above to assess missing outcome bias. To see this, note that this system implies $Y_1^* = \beta_0 + \beta_1 + U$, $Y_0^* = \beta_0 + U$, $S_1 = 1\left(V \geq -\gamma_0 - \gamma_1\right)$, $S_0 = 1\left(V \geq -\gamma_0\right)$, where $1\left(A\right)$ is an indicator variable that equals 1 in the event of A (0 otherwise), and $D = T$. The independence of T implies Assumption A, and if $\gamma_1 \geq 0$, then $\Pr\left(V < -\gamma_0 - \gamma_1, V \geq -\gamma_0\right) = 0$, implying that Assumption B holds also. It should be noted that the bounds proposed above can also be applied to a more general "heterogeneous treatment effect" version of the above latent-variable formulation. Since the independence and monotonicity assumptions are equivalent to a general latent index model (Vytlacil 2000), the above formulation can equivalently be re-cast in a latent-variable framework, yielding identical identification results.

A curious difference between the above system and typical formulations of censored selection models, is that there are no exclusion restrictions utilized here. Typically, in order to achieve identification of β_1 that does not rely upon functional form assumptions about the joint distribution of U and V, researchers posit the existence of an additional variable that directly impacts sample selection, but is not included in the outcome equation (Heckman 1990; Ahn and Powell 1993; Andrews and Schafgans 1998; Das, Newey, and Vella 2000). The trimming approach proposed here does not utilize an exclusion restriction.

3 Trimming Using Baseline Covariates

In randomized experiments, researchers often possess "baseline" characteristics of both the treatment and control subjects. These covariates are typically used to assess whether or not the randomization "failed", and if such a failure is not rejected by the data the covariates are often included in the analysis to reduce residual variation. These covariates can be used in a modified trimming method that will lead to tighter bounds on $E\left[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1\right]$ than that constructed without the covariates. I suppose that there is no missing data on these baseline covariates, in contrast to the bounds analysis of Horowitz and Manski (2000a).

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 $j = 1, \ldots, J$, $\Pr\left(X = x_j | D = 0, S = 1\right) \neq 0$.

Assumption C

$$(Y_1^*, Y_0^*, S_1, S_0, X) \text{ is independent of } D \tag{6}$$

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 $E\left[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1\right]$ 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 B and C hold, and* $\Pr\left[S = 1|D = 0\right] \neq 0$ *for each* $j = 1, \ldots, J$*. Denote the observed density and cumulative distribution of* Y*, conditional on* $D = 1$ *(and* $S = 1$*) and* $X = x_j$*, as* $f\left(y|x_j\right)$ *and* $F\left(y|x_j\right)$*, respectively. Then*

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

and

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

where

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

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, X)

Corollary 4 *Given Assumptions B and C and* $\Pr\left[S = 1|D = 0, X = x_j\right] \neq 0$ *for each* $j = 1, \ldots, J$

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

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

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

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 B and C hold 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 aggregate proportion that is eventually truncated remains fixed. An implication of the Proposition 5 is that in general, using more baseline covariates will lead to producing tighter bounds on $E\left[Y_1^* - Y_0^* | S_0 = 1, S_1 = 1\right]$.

It is interesting to relate these trimming bounds to the estimand that would result from a "matching on observables" approach to addressing missing outcome bias. Matching on the baseline covariates would dictate computing the quantity $\sum_{j=1}^{J} \Pr\left[X = x_j | S = 1, D = 0\right] \{E\left[Y | D = 1, S = 1, X = x_j\right] - E\left[Y | D = 0, S = 1, X = x_j\right]\}$. A comparison with the comparable quantity in the Corollary above makes it clear that this quantity will lie strictly in between the upper and lower "trimming" bounds.

4 A Comparison of Trimming to Using Bounded Support Conditions

There are three distinctive features of the trimming bounds proposed here. First, the model and procedure is appropriate for situations in which the outcome has unbounded, or very large support. This should be contrasted to a method that deals with 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). A limitation to the latter approach is that unbounded supports for the outcome variable will often imply that there will be no informative bounds for parameters of interest (Manski 1995).

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.

A third distinctive feature (and limitation) is that the bounds are only appropriate for average treatment effects for a particular sub-population: those individuals whose outcomes will be observed, irrespective of the assignment to treatment. Thus, the model and procedure ignores the average treatment effects for the sub-population that was "induced" to yield valid outcome data because of the treatment, $E\left[Y_1^* - Y_0^* | S_0 = 0, S_1 = 1\right]$. It also ignores the average treatment effects for the sub-population that will always have missing outcomes, irrespective of the treatment status, $E\left[Y_1^* - Y_0^* | S_0 = 0, S_1 = 0\right]$. However, when supports of the outcomes are unbounded for these latter sub-populations, it is not obvious that informative bounds could ever be constructed for these two parameters of interest.

Given these differences, a comparison between the "trimming" bounds and "imputation"-type bounds may not be meaningful in one respect, since the bounds are for different parameters of interest. (For ease of exposition, I refer to bounds generated through bounded support conditions as "imputation"-type bounds, since in many cases it is equivalent to generating worst-case scenarios by assigning upper and lower bounds of support to missing data). Nonetheless, it is informative to focus on a situation where both procedures can be applied, in order to examine the conditions under which one set of bounds will be either wider or tighter than the other, at a purely *mechanical* level. In the comparison that follows, I focus on the case where Y_1^* and Y_0^* are binary outcomes, taking on the values 0 or 1. I also abstract from the use of covariates.

The most readily comparable procedure in this context is that of Horowitz and Manski (2000a). Their more general framework would imply (in this special case of only missing outcome data with no covariates) lower and upper bounds, respectively, of

$$\begin{aligned} &\Pr\left[Y = 1 | D = 1, S = 1\right] \Pr\left[S = 1 | D = 1\right] \\ &- \Pr\left[Y = 1 | D = 0, S = 1\right] \Pr\left[S = 1 | D = 0\right] - \Pr\left[S = 0 | D = 0\right] \end{aligned} \tag{7}$$

and

$$\begin{aligned} &\Pr\left[Y = 1 | D = 1, S = 1\right] \Pr\left[S = 1 | D = 1\right] + \Pr\left[S = 0 | D = 1\right] \\ &- \Pr\left[Y = 1 | D = 0, S = 1\right] \Pr\left[S = 1 | D = 0\right] \end{aligned} \tag{8}$$

In essence, the procedure can be thought of as a construction of "worst-case" bounds by imputing the missing outcome data with all 1's or all 0's appropriately.

The discreteness of Y implies that there will not be a proper *density* function for the outcomes of the treatment group with valid outcome data, nor will there be a corresponding one-to-one cumulative distribution function. However, the trimming procedure described above can be modified appropriately, as long as some care is taken in noting the location of the "steps" in the cdf. In this binary outcome case, it can be shown that the lower and upper bounds for $E\left[Y_1^* - Y_0 | S_0 = 1, S_1 = 1\right]$ will be, respectively,

$$\max\left[0, \frac{\Pr\left[S=1|D=1\right]}{\Pr\left[S=1|D=0\right]}\left\{\Pr\left[Y=1|D=1,S=1\right] - \frac{\Pr\left[S=1|D=1\right]-\Pr\left[S=1|D=0\right]}{\Pr\left[S=1|D=1\right]}\right\}\right] - \Pr\left[Y=1|D=0,S=1\right] \tag{9}$$

and

$$\min\left[1, \frac{\Pr\left[Y=1|D=1,S=1\right]}{\left(\frac{\Pr[S=1|D=0]}{\Pr[S=1|D=1]}\right)}\right] - \Pr\left[Y=1|D=0,S=1\right] \tag{10}$$

The lower bound is simply computed by reducing the fraction $Y=1$ by the proportion in the excess group (those with $S_0=0, S_1=1$), which effectively assumes that the trimmed group had $Y_1^*=1$, and inflating by a factor of $\frac{\Pr[S=1|D=1]}{\Pr[S=1|D=0]}$, reflecting that the denominator for computing the fraction $Y=1$ has diminished due to the trimming. This quantity will equal zero if all of the 1's in the group are trimmed. For the upper bound, assuming that the "trimmed" group all had $Y_1^*=0$ requires inflating $\Pr\left[Y=1|D=1,S=1\right]$ by $\frac{\Pr[S=1|D=0]}{\Pr[S=1|D=1]}$ to reflect that the denominator for computing the fraction $Y=1$ has diminished due to the trimming. This quantity will equal 1 if all of the 0's in the group are trimmed.

In the case where all the 1's and 0's would be trimmed in computing the lower and upper trimming bounds, then the width of those bounds would be 1. The width of the imputation bounds is $\Pr\left[S=0|D=0\right] + \Pr\left[S=0|D=1\right]$. So the comparison of the widths of the two bounds simply amounts to comparing the quantity $\Pr\left[S=0|D=0\right] + \Pr\left[S=0|D=1\right]$ to 1.

In the polar opposite case, where trimming the lower tail only eliminates 0's and trimming the upper tail only eliminates 1's (so that the $\min$ and $\max$ are not strictly binding at 1 and 0, respectively), then with some re-arranging of terms it can be shown that the imputation bounds will be narrower than the trimming bounds if and only if

$$\Pr\left[S=1|D=1\right] > \Pr\left[S=1|D=0\right]\frac{3-\Pr\left[S=1|D=0\right]}{1+\Pr\left[S=1|D=0\right]} \tag{11}$$

and

$$\Pr\left[S=1|D=1\right] > 1-\Pr\left[S=1|D=0\right] \tag{12}$$

The second constraint exists because the width of the trimming bound can never be larger than 1. Focusing on the upper triangular portion of Figure I, region A represents the combinations of $\Pr\left[S=1|D=0\right]$ and $\Pr\left[S=1|D=1\right]$ such that the imputation bounds will be narrower than the trimming bounds. The vertex of the region is $\Pr\left[S=1|D=0\right]=\frac{1}{3}$ and $\Pr\left[S=1|D=1\right]=\frac{2}{3}$.

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, if $\Pr\left[S_0=1,S_1=0\right]=0$, it implies that there exists no j, such that $\Pr\left[S=1|D=1,X=x_j\right] < \Pr\left[S=1|D=0,X=x_j\right]$. Essentially, the monotonicity restriction is inconsistent with the existence of j' and j'' such that $\Pr\left[S=1|D=1,X=x_{j'}\right] < \Pr\left[S=1|D=0,X=x_{j'}\right]$ while at the same time $\Pr\left[S=1|D=1,X=x_{j''}\right] < \Pr\left[S=1|D=0,X=x_{j''}\right]$.

Finally, suppose $\Pr\left[S=1|D=1\right]=\Pr\left[S=1|D=0\right]$. As mentioned earlier, in this case, Assumptions B and C imply that there is no sample selection bias, and that a simple contrast between $E\left[Y|D=1,S=1\right]-E\left[Y|D=0,S=1\right]$ is valid for identifying a meaningful causal parameter. $0=\Pr\left[S=1|D=1\right]-\Pr\left[S=1|D=0\right]=\sum_j^J\{\Pr\left[X=x_j|D=1\right]\left(\Pr[S=1|D=1,X=x_j]-\Pr[S=1|\ D=0,X=x_j]\right)\}$because of Assumption C. Assumption B implies that $\Pr\left[S=1|D=1,X=x_j\right]-\Pr\left[S=1|D=0,X=x_j\right]=0$ for $j=1,\ldots,J$. It can then be shown, using Assumption C and Bayes' rule, that this implies $\Pr\left[X=x_j|S=1,D=1\right]=\Pr\left[X=x_j|S=1,D=0\right]$ 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 Conclusion

This paper has explored how treatment effect parameters can be bounded if a monotonicity assumption is imposed on the censored selection process. It has been shown that a simple and intuitive trimming procedure yields such bounds, and the use of baseline covariates will in general narrow their width. The main benefit from imposing the monotonicity restriction is that it allows one to generate 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. Existing nonparametric bounding approaches (e.g. Horowitz and Manski 1998, 2000a) of unbounded outcomes immediately suggest there will be no finite bounds on 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 issue then becomes Which restrictions have relatively large benefits and/or small costs? Viewed from this perspective, the benefits of being able to analyze outcomes with unbounded (or extremely large) support may, in some contexts, outweigh the cost of imposing the behavioral assumption that this trimming procedure requires.

Appendix A.

Lemma 6 *Suppose the probability density $f^*(y)$ is a mixture of two probability densities, $m^*(y)$ and $n^*(y)$ such that $f^*(y) = p^* m^*(y) + (1-p^*) n^*(y)$, where $p^* \in [0,1)$ is fixed. Let $F^*(y)$ be the cumulative distribution function corresponding to $f^*(y)$. Consider the truncated density $g^*(y)$ which is equal to $\frac{1}{1-p^*} f^*(y)$ on $\left[F^{*-1}(p^*), \infty\right]$, 0 otherwise. Then $\int_{-\infty}^{\infty} y g^*(y)\, dy \geq \int_{-\infty}^{\infty} y n^*(y)\, dy$.*

Proof of Lemma 6. First consider $p^* \in (0,1)$. Let $N^*(y)$ be the cumulative distribution function corresponding to $n^*(y)$. Compare the truncated density, $g^*(y)$ for $y \geq F^{*-1}(p)$ (0 otherwise), to an arbitrarily chosen $n^*(y)$ (that is not identical to the truncated density). $\int_{-\infty}^{\infty} y g^*(y)\, dy - \int_{-\infty}^{\infty} y n^*(y)\, dy$ $= \int_{F^{*-1}(p^*)}^{\infty} y \left(\frac{1}{1-p^*}\right) f^*(y)\, dy - \int_{-\infty}^{\infty} y n^*(y)\, dy = \int_{F^{*-1}(p)}^{\infty} y \left[\left(\frac{1}{1-p^*}\right) f^*(y) - n^*(y)\right] dy - \int_{-\infty}^{F^{*-1}(p^*)} y \cdot$ $n^*(y)\, dy$. Multiplying both sides by $\frac{1}{N^*(F^{*-1}(p^*))}$ yields $\frac{1}{N^*(F^{*-1}(p^*))} \left\{\int_{-\infty}^{\infty} y g^*(y)\, dy - \int_{-\infty}^{\infty} y n^*(y)\, dy\right\}$ $= \frac{1}{N^*(F^{-1}(p^*))} \int_{F^{*-1}(p^*)}^{\infty} y \left[\left(\frac{1}{1-p^*}\right) f^*(y) - n^*(y)\right] dy - \frac{1}{N^*(F^{*-1}(p^*))} \int_{-\infty}^{F^{*-1}(p)} y n^*(y)\, dy$. By definition $n^*(y) = \frac{f^*(y) - p^* m^*(y)}{1-p^*}$, so for any y on $\left[F^{*-1}(p^*), \infty\right]$, $n^*(y) \leq \frac{1}{1-p} f^*(y)$. If $n^*(y) \neq g^*(y)$, then it can be shown that $\frac{1}{N^*(F^{*-1}(p^*))} \left[\left(\frac{1}{1-p^*}\right) f^*(y) - n^*(y)\right]$ defined on $\left[F^{*-1}(p^*), \infty\right]$ and $\frac{1}{N^*(F^{-1}(p))} n^*(y)$ defined on $\left[-\infty, F^{*-1}(p^*)\right]$ are each proper probability densities that integrate to 1. The support of the former is strictly above the support of the latter. Therefore, $\frac{1}{N^*(F^{*-1}(p^*))} \left\{\int_{-\infty}^{\infty} y g^*(y)\, dy - \int_{-\infty}^{\infty} y n^*(y)\, dy\right\}$ > 0. If $n^*(y) = g^*(y)$, then $\int_{-\infty}^{\infty} y g^*(y)\, dy = \int_{-\infty}^{\infty} y n^*(y)\, dy$.

Now consider $p^* = 0$. Then $g^*(y) = f(y) = n^*(y)$, so $\int_{-\infty}^{\infty} y g^*(y)\, dy = \int_{-\infty}^{\infty} y n^*(y)\, dy$.

Proof of Proposition 1. Assumption 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 density of Y_1^*, conditional on $D=1$, $S_0 = 0$, $S_1 = 1$, and $n(y)$ denotes the density of Y_1^*, conditional on $D=1$, $S_0 = 1$, $S_1 = 1$. By Assumption A, $n(y)$ is also the density of Y_1^*, conditional on $S_0 = 1, S_1 = 1$. By Lemma 6, $\overline{E} \equiv \frac{1}{1-p} \int_{F^{-1}(p)}^{\infty} y f(y)\, dy \geq \int_{-\infty}^{\infty} y n(y)\, dy = E\left[Y_1^* | S_0 = 1, S_1 = 1\right]$. To show that $\overline{E}$ equals the maximum possible value for $E\left[Y_1^* | S_0 = 1, S_1 = 1\right]$ that is consistent with the distribution of the observed data on (Y, S, D), note first that the observed data can be completely described by $f(y)$, the density of Y conditional on $S = 1, D = 0$, and the probability function $\Pr[S = s, D = d]$, $s, d = 0, 1$. Set $n(y)$ equal to the density $\frac{1}{1-p} f(y)$ defined on $\left[F^{-1}(p), \infty\right]$, and $m(y)$ equal to the density

$\frac{1}{p} f(y)$ defined on $\left[-\infty, F^{-1}(p)\right]$ where $p \equiv \frac{\Pr[S=1|D=1]-\Pr[S=1|D=0]}{\Pr[S=1|D=1]} = 1 - \frac{\left(1+\frac{\Pr[S=0,D=1]}{\Pr[S=1,D=1]}\right)}{\left(1+\frac{\Pr[S=0,D=0]}{\Pr[S=1,D=0]}\right)}$; there is only one p consistent with the probability function $\Pr[S=s, D=d]$, $s, d = 0, 1$. These choices for $n(y)$ and $m(y)$ are consistent with $f(y)$ sastisfying $f(y) = pm(y) + (1-p)n(y)$, and consistent with any density of Y conditional on $S=1, D=0$. Then $E[Y_1^*|S_0=1, S_1=1]$ will equal $\frac{1}{1-p}\int_{F^{-1}(p)}^{\infty} yf(y)\,dy \equiv \overline{E}$.

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

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 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 $\frac{1}{1-p_j}\int_{F^{-1}(p_j|x_j)}^{\infty} yf(y|x_j)\,dy \geq E[Y_1^*|S_0=1, S_1=1, X=x_j]$ 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 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), note first that the data can be completely described by $f(y|x_j)$, the density of Y conditional on $S=1, D=0, X=x_j$, the probability function $\Pr[S=s, D=d|X=x_j]$, $s, d = 0, 1$, and the probability function $\Pr[X=x_j]$, $j = 1, \ldots, J$. Since Assumptions A and B hold conditionally on X, by the Proposition 1, $\frac{1}{1-p_j}\int_{F^{-1}(p_j|x_j)}^{\infty} yf(y|x_j)\,dy$ is the largest possible value for $E[Y_1^*|S_0=1, S_1=1, X=x_j]$ consistent with the observed data on (Y, S, D), conditional on $X=x_j$, for each $j = 1, \ldots, J$. It follows that $\sum_{j=1}^{J} \Pr[X=x_j|S_0=1, S_1=1]\, \frac{1}{1-p_j}\int_{F^{-1}(p_j|x_j)}^{\infty} yf(y|x_j)\,dy$ is the largest possible value for $E[Y_1^*|S_0=1, S_1=1]$. $\Pr[X=x_j|S_0=1, S_1=1] = \Pr[X=x_j|S=1, D=0]$, by Assumptions B and C, and the latter quantity can be expressed as $\frac{\Pr[S=1,D=0|X=x_j]\Pr[X=x_j]}{\sum_{k=1}^{J}\Pr[S=1,D=0|X=x_j]\Pr[X=x_k]}$. Thus, the largest possible value for $E[Y_1^*|S_0=1, S_1=1]$ that is consistent with the observed data on (Y, S, D), conditional on X, and the probability function $\Pr[X=x_j]$ is $\sum_{j=1}^{J}\{\Pr[X=x_j|S=1, D=0] \cdot \frac{1}{1-p_j}\int_{F^{-1}(p_j|x_j)}^{\infty} yf(y|x_j)\,dy\} \equiv \overline{E}^*$.

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

Proof of Proposition 5. Assumption B implies that $p_j \geq 0$ for $j = 1, \ldots, J$. Let $g\left(y|x_j\right) = \frac{1}{1-p_j} f\left(y|x_j\right)$ on $\left[F^{-1}\left(p_j|x_j\right), \infty\right]$, 0 otherwise. Let $h\left(y|x_j\right) = \frac{1}{p_j} f\left(y|x_j\right)$ on $\left[-\infty, F^{-1}\left(p_j|x_j\right)\right]$, 0 otherwise. Then $f\left(y\right) = \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 1\right] f\left(y|x_j\right) = \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 1\right] p_j h\left(y|x_j\right) + \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 1\right]\left(1 - p_j\right) g\left(y|x_j\right)$. Let $\widehat{p} = \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 1\right] p_j$. Then $f\left(y\right)$ can be re-written as $\widehat{p} m^*\left(y\right) + \left(1 - \widehat{p}\right) n^*\left(y\right)$, where $m^*\left(y\right) = \frac{1}{\widehat{p}} \sum_{j=1}^{J} \{\Pr\left[X = x_j|S = 1, D = 1\right] \cdot p_j h\left(y|x_j\right)\}$ and $n^*\left(y\right) = \frac{1}{1-\widehat{p}} \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 1\right]\left(1 - p_j\right) g\left(y|x_j\right)$. It is given that for each j, $\Pr\left[X = x_j|D = 0, S = 1\right] \neq 0$. So $\Pr\left[S = 1|D = 0\right] \neq 0$ implies that $\Pr\left[S = 1|D = 0, X = x_j\right] \neq 0$ for all $j = 1, \ldots, J$. So $p_j \in [0, 1)$ for $j = 1, \ldots, J$, and thus $\widehat{p}$ lies on $[0, 1)$.

Consider first $\widehat{p} \in (0, 1)$. Since $m^*\left(y\right)$ and $n^*\left(y\right)$ are both probability densities that integrate to 1, Lemma 6 applies: $\frac{1}{1-\widehat{p}} \int_{F^{-1}(\widehat{p})}^{\infty} y f\left(y\right) dy \geq \int_{-\infty}^{\infty} y n^*\left(y\right) dy$. The definition of p_j implies that $\widehat{p} = \sum_{j=1}^{J} \{\Pr[X = x_j|\ S = 1, D = 1] \left(1 - \frac{\Pr[S=1,D=0,X=x_j] \Pr[D=1,X=x_j]}{\Pr[D=0,X=x_j] \Pr[S=1,D=1,X=x_j]}\right)\}$. Simplifying, and by assumption C, $\widehat{p} = 1 - \sum_{j=1}^{J} \frac{\Pr[S=1,D=0,X=x_j] \Pr[D=1]}{\Pr[S=1,D=1] \Pr[D=0]} = 1 - \frac{\Pr[S=1,D=0] \Pr[D=1]}{\Pr[S=1,D=1] \Pr[D=0]} = 1 - \frac{\Pr[S=1|D=0]}{\Pr[S=1|D=1]} = p$. Then $n^*\left(y\right) = \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 1\right] \frac{1-p_j}{1-p} g\left(y|x_j\right)$. Using definitions of p and p_j, this is equal to $\sum_{j=1}^{J} \{\Pr[X = x_j|S = 1, D = 1] \frac{\frac{\Pr\left[S=1,D=0,X=x_j\right] \Pr\left[D=1,X=x_j\right]}{\Pr\left[D=0,X=x_j\right] \Pr\left[S=1,D=1,X=x_j\right]}}{\frac{\Pr[S=1,D=0] \Pr[D=1]}{\Pr[D=0] \Pr[S=1,D=1]}} g\left(y|x_j\right)\}$. Applying Assumption C, and simplifying, yields $\sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 1\right] \frac{\Pr[S=1,D=0,X=x_j] \Pr[S=1,D=1]}{\Pr[S=1,D=1,X=x_j] \Pr[S=1,D=0]} g\left(y|x_j\right)$. Simplifying further yields $\sum_{j=1}^{J} \frac{\Pr[X=x_j,S=1,D=0]}{\Pr[S=1,D=0]} g\left(y|x_j\right) = \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 0\right] g\left(y|x_j\right)$. Therefore, $\overline{E} = \frac{1}{1-p} \cdot \int_{F^{-1}(p)}^{\infty} y f\left(y\right) dy \geq \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 0\right] \int_{-\infty}^{\infty} y g\left(y|x_j\right) dy = \sum_{j=1}^{J} \{\Pr[X = x_j|S = 1, D = 0] \frac{1}{1-p_j} \int_{F^{-1}(p_j|x_j)}^{\infty} y f\left(y|x_j\right) dy\} \equiv \overline{E}^*$.

Now consider the case $p = 0$. This means $\Pr\left[S = 1|D = 1\right] = \Pr\left[S = 1|D = 0\right]$. Given Assumption B and C, this implies that $\Pr\left[S_0 = 0, S_1 = 1\right] = 0$. Analogously, if $p_j > 0$ for any $j = 1, \ldots, J$, then $\Pr\left[S = 1|D = 1, X = x_j\right] > \Pr\left[S = 1|D = 0, X = x_j\right]$ which would imply that $\Pr[S_0 = 0, S_1 = 1|X = x_j] \neq 0$. Therefore, $p_j = 0$ for all $j = 1, \ldots, J$. If $p_j = 0$, then $\Pr\left[S = 1|D = 1, X = x_j\right] = \Pr\left[S = 1|D = 0, X = x_j\right]$. Assumption C and re-arranging terms yields $\Pr\left[X = x_j|S = 1, D = 1\right] = \Pr\left[X = x_j|S = 1, D = 0\right]$, for all $j = 1, \ldots, J$. $\overline{E} \equiv \int_{-\infty}^{\infty} y f\left(y\right) dy = \sum_{j=1}^{J} \{\Pr[X = x_j|S = 1, D = 1] \int_{-\infty}^{\infty} y f\left(y|x_j\right) dy\} = \sum_{j=1}^{J} \Pr\left[X = x_j|S = 1, D = 0\right] \int_{-\infty}^{\infty} y f\left(y|x_j\right) dy = \overline{E}^*$.

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

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

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

[7] Heckman, J. J. (1974), “Shadow Prices, Market Wages, and Labor Supply,” *Econometrica*, 42, 679-693.

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

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

[10] 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.

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

[12] 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*.

[13] 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.

[14] 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.

[15] 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.

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

[17] Manski, C. F. (1989), “Anatomy of the Selection Problem,” *Journal of Human Resources*, 24, 343-360.

[18] Manski, C. F. (1990), “Nonparametric Bounds on Treatment Effects,” *American Economic Review Papers and Proceedings*, 80, 319-323.

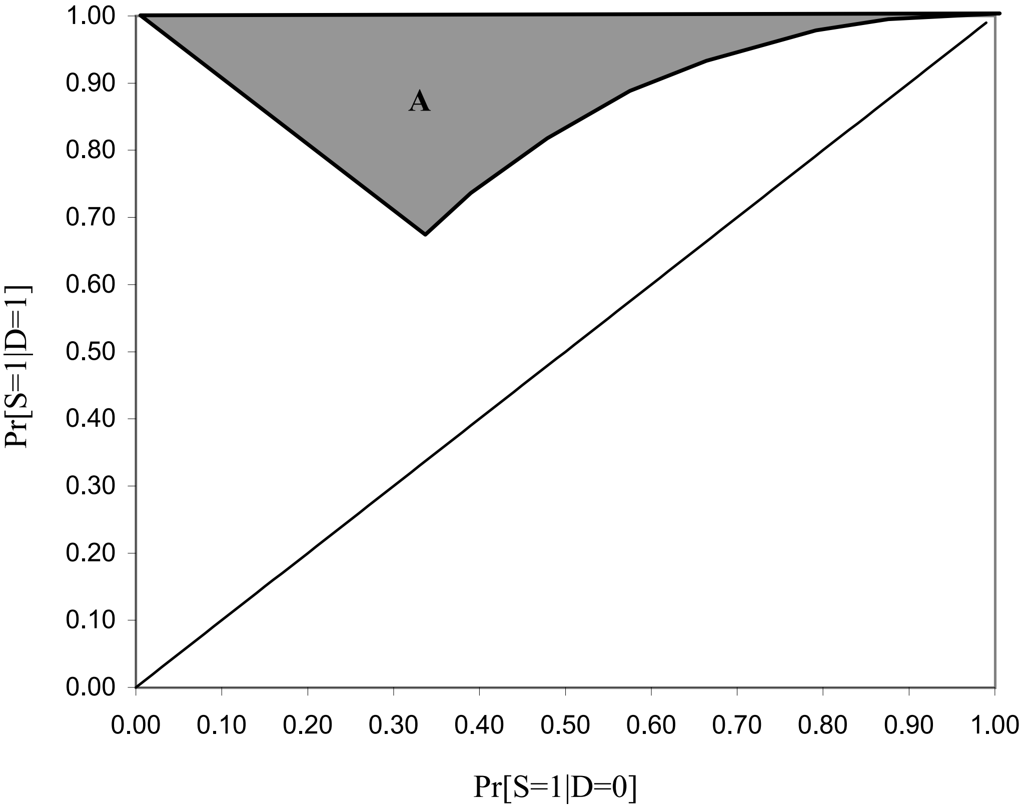
[19] Manski, C. F. (1995), *Identification Problems in the Social Sciences*, Cambridge, MA: Harvard University Press.

[20] 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.

[21] Rubin, D. (1976), “Inference and Missing Data” *Biometrika*, 63, 581-592.

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

Figure I: Comparison between Trimming and Imputation Bounds



Note: In region A, Imputation Bounds are narrower than Trimming Bounds