Notes: Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects, David Lee, 2009, RESTUD

Notes David Reinstein

Introduction

> even with the aid of a randomized experiment, the impact of a training program on wages is difficult to study because of sample selection, a pervasive problem in applied microeconometric research

- Intuitive trimming procedure for bounding average treatment effects in the presence of sample selection…
- Requires neither exclusion restrictions nor a bounded support for the outcome of interest."
- (Also) applicable to "nonrandom sample selection/attrition", as well as to the 'conditional on positive'/hurdle/mediation effect discussed here

> analyses and evaluations typically focus on "reduced form impacts on total earnings, a first-order issue for cost-benefit analysis. Unfortunately, exclusively studying the effect on total earnings leaves open the question of whether any earnings gains are achieved through raising individuals hypothesis wage rates (price affects or hours of work (quantity effects).

Important methodological point to constantly bring up: "even a randomized experiment cannot guarantee the treatment and control individuals will be comparable conditional on being employed."

Claims that standard "parametric or semi-parametric methods for correcting sample selection require exclusion restrictions that have little justification in this case." Notes that most of the baseline variables could affect employment probabilities or have a direct impact on wage rates.

Summary of the method: "…amounts to first identifying the excess number of individuals who were induced to be selected (employed) because of the treatment and then trimming the upper and lower tails of the outcome… distribution by this number, yielding a worst-case scenario bound."

Uses same assumptions as in "conventional models for sample selection"

1. regressor of interest is independent of the errors in the outcome and selection models selection equations – this is ensured by random assignment.
2. "the selection equation can be written as a standard latent variable binary response model"

– what meaningful restriction does this impose?

He proves this procedure "yields the tightest bounds for the average treatment effect that are consistent with the observed data."

The bounds estimator is shown to be $\sqrt{(n)}$ consistent and asymptotically normal with an intuitive expression for its asymptotic variance which depends on the variance of the trimmed outcome and the trimming threshold, an estimated quantifiable; (and an added term accounting for the estimation of which quantile to trim on)

– *DR, Note, charity data: We can make confidence statements over the bounds themselves. Will this procedure be easy to bring into our code?* – In our (charity) experiment we in fact do have upper bounds on the outcome variable. Could this yield even greater efficiency?

– *DR, Note for the Netherlands data*: it is not immediately clear how this could be adapted to instrumental variables; we shall see. Can we recover something meaningful from the reduced form model they are? Can it be applied to the (instrumental variables) estimates to disentangle the impact of changing courses from the impact of the institution itself?

In Lee's paper, the estimate seems to give very narrow and informative bounds even though they have a great deal of people who do not earn any wages as a share of the population, about 54%. These are much narrower than the bounds proposed by Horowitz and M then what those bounds produce.

– ask @Gerhard whether his Horowitz/Manski estimator incorporated the natural bounds on the outcome.

The National Job Corps Study and Sample Selection [prior approaches]

In the experiment discussed here those in the control group were embargoed rom the program for three years but could join afterwards, thus "when I use the phrase 'effect of the program' I am referring to this reduced-form treatment effect", i.e., the intent to treat effect.

– "some subpopulations were randomized into the program group with differing, but known probabilities. Thus analyzing the data requires the use of design weights." – *Note:* (@NL) this bears some resemblance to our Dutch data situation, and we can probably use examples from analyses of these programs. – Note also that they impute means of the baseline variables with their means; this seems to be an accepted practice.

Lee notes that he focuses exclusively on the "sample selection on wages caused by employment" and not the attrition/nonresponse problem, to focus attention on this, but they could have used it for the other as well.

– *DR:* (@NL) Note again that their desire to separate the employment hours and wage effects of the program is very similar to our desire to separate out different margins of the impact of winning an institution. …Namely the impact on completing a course or starting a course versus other impacts and the impact of entering a specialization versus remaining impacts. …Similar decompositions for the geography outcomes.

```
- To do: check whether any papers cite Lee using an IV approach, extending the technique and the estimation
```

> "the problem of nonrandom sample selection is well-understood in the training literature; … may be one of the reasons why most evaluations of job-training programs focus on total earnings, including zeros for those without a job, rather than on wages conditional on employment" "of the 24 studies referenced in a survey … (Heckman et al.)… Most examine annual, quarterly, or monthly earnings without discussing the sample selection problem examining rage rates."

– *DR:* (@NL)Note that this is relevant to our question of whether to exclude zeros in log models, etc. While there will be less unemployment in our data, it still may be a relevant influence made have a strong effect on the estimates.

...previous conventional approaches to the sample selection problem (skip if desired). One may explicitly model the process determining selection, such as in Heckman (1979) …

Separate equations for the wage and the propensity to be employed, where employment occurs if the latter crosses a particular threshold, in which case a wage is observed. It is reasonable to think that the treatment variable can have effects on both terms..

"sample selection bias can be seen as specification error in the conditional expectation…"

The expected wage conditional on treatment exogenous variables and the selection into working (that is the underlying propensity to work variable exceeding zero) his status is equal to the true effect of the treatment an adjustment for the differences in the observable's exogenous variables and a bias term representing the expectation of the idiosyncratic unobservables given the treatment and the exogenous variables exceeding the value necessary to induce work participation. The unobservable term needs to exceed the prediction based on the observable term for the entire term to exceed zero inducing labor force participation.

One may assume the data are missing at random, perhaps conditional on a set of covariates (Rubin, 1976; essentially assuming the error terms in each equation are independent of one another, here "employment status is unrelated to the determination of wages"… This "is strictly inconsistent with standard models of labor supply that account for the participation decision (Heckman, 1974)."

A more common assumption is that some exogenous variables "determine sample selection but do not have their own direct impact on the outcome of interest…. Exclusion restrictions are used in parametric and semi-parametric models…"

but "there may not exist credible 'instruments… excluded from the outcome equation"

– *DR, aside:* We can return to (our) previous papers to impose these Lee bounds! One example would be the Siskel and Ebert your reviews paper and perhaps incorporating us with subsequent approaches, considering the "selection to review" equation.

Second approach "the construction of worst-case scenario bounds of the treatment effect"

"Impute missing data with either the largest or smallest possible values to compute the largest and smallest possible treatment effects consistent with the data" as in Horwitz and Manski (2000a) who provide a general framework for this.

- Particularly useful with binary outcomes.

This cannot be used when the support is unbounded. … note in their replication example they are actually using the equivalent of the bottom 5th percentile and the top 95th percentile. Strictly using a procedure would provide even wider bounds.

Lee considers his approach to be a hybrid of the two previous general approaches.

…end of section 2.. .a statement of the Horwitz upper bound for the treatment effect; very intuitive: "what if everyone in the treatment who dropped out would have had the largest possible wage and everyone in the control group that drop out would've had the smallest possible wage; this will give the upper bound." Switching this the other way around will give a lower bound.

DR, an aside thought: (@NL) Something akin to the Horwitz and M approach (or maybe Lee bounds) could be applied to our issue of swapping into institutions directly. Suppose we only focus on those who *actually* complied: those assigned to an institution who also went to that institution. Our concern was that this would under-represent those who had particularly strong institutional preferences. Suppose you are interested in looking at the impact of winning the lottery (for once preferred institution) itself, as that was our most simply identifiable outcome.

… Let's consider evaluating a treatment effect for those who happened to swap in. Those who swapped in might be assigned a counter-factual outcome of the lowest value of the lifetime income among those who did not get their

institution of choice. Similarly, the small group who swapped out might be assigned a counterfactual outcome (had they no swapped out) representing the highest outcome value for those who did get their institution of choice. This should give us an upper bound on the treatment effect for these two groups of what we might call non-compliers. Making the opposite assumptions, precisely that those who swapped *into* their institution of choice would've had a very good counterfactual outcome (if they had not got their institution of choice) that comes from the highest outcomes for those who didn't get their institution of choice (and also reversing this for those who swapped out of their preferred institution) would give us a lower bound for the treatment effect for this group. We can then combine these bounded treatment effects for these non-compliers with the treatment effect for the compliers to get a measure of the average treatment effect with bounds for this sort of behavior.

This will also allow us to come up with estimates with bounds *without* having to use the instrumental variable strategy which has issues of its own.

Section 3: identification of bounds on treatment effects; the main meat of the model

He starts with a simple example. He begins with a model with a treatment indicator and no other covariates, and a continuous outcome variable, but notes that this will clearly apply to discrete outcome variables and will also apply conditional on controls.

Nest, he brings forward the statement… from the earlier selection models. In each case the latent variable must overcome a hurdle for the outcome to be observed and in fact *the hurdle differs depending on the impact of the treatment itself.* In general *when the errors in the selection and outcome equations are correlated the difference in these means differs from the actual treatment effect.* In other words through a slightly complicated story, when those who have unobservables that make them more likely to work also tend to have unobservables that would make them likely to earn more the standard difference in outcomes between control and treatment will *not* describe the true treatment effect.

A key insight seems to be that we could identify the treatment effect if we could estimate the expected outcome given treatment *and* given that the unobservable component in the selection equation would lead to an observable outcome had the person *not* been given treatment. If so, we could subtract the observed mean control outcome from the above to yield the true treatment effect (for those who would be observed always). However, we obviously do not observe this because we only observe the outcomes for those who are treated where the selection equation *is* in fact positive and not "where the selection equation *would have* been positive had they not been treated."

However, the insight here is that this term can in fact be bounded. We *do* observe these outcomes for the treated people (note we are assuming without loss of generality that the treatment raises the probability of selection for this discussion) but we don't know exactly which ones they are. In other words, we observe outcomes for more people in the treatment group than we need; we wish we could figure out what is the subset of these that would have *also* been observed had they not been treated, so we could compare like-to-like. The observed treatment mean is a weighted average of the thing we are seeking (to difference from the control) and "the mean for a subpopulation of marginal individuals… that are induced to be selected into the sample *because* of the treatment"

This then gets us the upper bound for the term expressing the treatment outcome for those who would have been observed even if they had been in the control. The upper bound for this is the expected outcome for those in the treated group (who are observed of course) and who are in quantile-p or above of the outcome, where this p is the share of the treated population that are in the marginal group we referred to that were only induced to be selected into the sample because of the treatment.

In other words the worst case scenario is that the smallest share p values of Y are in the marginal group and the largest one (which is share 1-pone are in the inframarginal group. We don't know which observations are inframarginal and which ones are marginal.

p: the share of marginal individuals and (1-p) the share of inframarginal individuals (the latter is group we want the average outcome for). The highest could be would be the average outcomefor the largest (1-p) share of this group. We are looking for the expectation given that they are at or above at will at or above percentile p within this group.

In other words we trim the lower tail of the Y distribution by the portion p, (so what remains is the 1-p share) to get the upper bound for the inframarginal groups mean. We can then subtract the mean for the control group to get an upper bound for the treatment effect.

To compute this "trimming proportion p": this p is equal to the share of the treated group whose outcome is observed minus the share of the control group whose outcome is deserved is observed, divided by the share of the treatment group where the outcome is observed. Something like the *increased likelihood of observation that is driven by the treatment, as a share of the total number as a share of the probability of observation in the treatment group.*

The average observed outcome for the treatment group is including too many observations; we need to difference out the share of observations that are observed only because the treatment caused them to be observed; this share is certainly no larger than the increased probability of observation in the treatment group as a share of the probability of observations the treatment group.

Another much simpler way of saying this is "trimming the data by the known proportion of excess individuals" in the treatment group. (To gain bounds on the mean for the inframarginal group which we can then difference from the control-group mean get the treatment effect).

Perhaps some intuition for why this improves on the Horwitz model: we don't need to assume that those observed in the treatment group that wouldn't have been observed in the control would've had the highest possible outcomes. No, we only need to assume (to get the upper bound) that these came from the highest *distribution* because they had to come from somewhere. These were the people in the upper tail of the relevant group but they couldn't *all* have been the individual highest achiever.

The model is extended to heterogeneity and heteroscedasticity. This begins with the independence of treatment assignment the "potential sample selection indicators" for either treatment or control, in other words whether that individual will have an observed outcome under treatment and whether that the individual would have an observed outcome under control, and the latent potential outcomes.

Experimental or random assignment ensures that each of the potential outcomes (and the correspondence to observability under each treatment) is independent of the actual assigned treatment.

The second assumption is monotonicity: treatment assignment can only affect sample selection in one direction.

– DR: For our (substitution) experiments, it is in fact not clear to me whether this should necessarily be the case, as some (less generous?) people may be induced to leave because of having been asked to donate, while potentially other (more generous people) might be induced to return given that they were asked to donate. (This proposed nonmonotonicity implies that the 'asked twice' sample tends to weed out the less generous, which would lead to a bias *against* substitution, strengthening the case for our result.) - DR, aside: However, even though the paper doesn't say it, I suspect this assumption could be weakened and you would still get some similar bounds. To put it another way, I would imagine that these bounds could be adjusted based on some reasonable ad hoc assumptions about the share of the population who is affected in either direction.

– @NL: I'm coming to think that our Dutch data problems are more things involving "hurdle models". Can this technique also be applied to such hurdle models?

Next proposition 1a states that given these assumptions we can derive sharp lower and upper bounds for the average treatment effect (conditional on 'would be observed in both states'). Note that for this estimator if the probability of observation is greater under the treatment we need to trim the treatment groups outcome distribution and if the probability of observation is greater under the control we need to trim the control group's outcome distribution.

- DR, aside comment: we seem to be throwing out a bit of the data in these estimates, which would suggest that something more efficient could be generated.

(The stated bounds you can estimate are exactly the same as the bounds from the previous specification, at least as I had interpreted the way they would be produced.)

Their remark 2 notes that an implication is that as P_0, that as the "difference between the relative probability of observation of an outcome under treatment versus control" tends to zero, i.e., as the probability of having an observed outcome (or the conditional probability of this) is the same for treatment and control) then there is no sample selection bias.

Their estimate convergences to the estimate he calls an estimate for the "always takers subpopulation… except that taking… is selection into the [outcome-observed] sample."

So, a very vanilla estimator is acceptable if we find the same conditional probability of selection for each group, under monotonicity, which, for this case, we can test (see Remark 4 below).

– (DR: To me this suggests that there might be something wrong going on here. Intuitively, If I simply observe the same rate of attrition in the treatment and control groups this *shouldn't* be enough to tell me that attrition did not matter, as it could occur differentially for both groups, but it seems to be a result here; this is probably due to the assumption of monotonicity of the selection/observation term, as well as the random/exogenous assignment to each group.)

Remark 3 discusses the importance of monotonicity for the bounds, saying this assumption is "minimally sufficient" (I think it would be better to say minimally sufficient for these particular bounds that he computed). To demonstrate this he gives an extreme example. Without monotonicity it could be (note: this would seem like a very unlikely outcome!) that every observation in the control group comes from the population in the treatment group that would *not* have been observed had they been treated and every observation in the treatment group happens to come from the set of people that would *not* have been observed had they been in the control group. These two "subpopulations do not overlap, so the difference in the means could not be interpreted as a causal effect."

– DR, aside : there must be some way to impose some restrictions on this even allowing for this non-monotonicity. (He notes that this can be improved upon somewhat by thinking about the total the idea that the total masses of unobserved that would've been observed in the other group can't be greater than the share that is not observed in the other treatment group, but this doesn't seem like a particularly fruitful route as it in most reasonable cases will still allow for very wide bounds.)

Remark 4 suggests that if we can assume (or somehow observe?) that the conditional probabilities of selection are the same for treatment and control, we can *test whether monotonicity in fact holds* and the simple difference in means will be an appropriate estimate of the treatment effect. Here, the assumption implies that everyone in the treatment or control group would have been observed under the opposite treatment as well. This in fact implies that the distribution of the exogenous variables should be the same in the treatment and control groups conditional on being selected. This seems fairly intuitive, we look at whether selection seems to be occurring in different ways are on different margins for the two groups treatment versus control.

Apparently for this test to have *power* we need that the subpopulations of "noncompliant errors in opposite directions" (quotation mine) must have *distinct* distributions of baselines exogenous characteristics. If these were the same then whether or not monotonicity holds the test doesn't tell us anything.

– DR: *I wonder if anyone uses this test for Monotonicity under non-differential selection?*

Another relevant note that he bundles in this remark is that the technique here only yields estimates *for those who would be with an observed outcome for either treatment or control.* One could *additionally* try to bound this as an estimate for the entire population using the Horwitz and Manski bounds for this latter thing. However, in many contexts there are reasons that the bounded estimates they mainly use are the relevant ones, such as "the impact of the program on wage rates for those whose employment status was not affected by the program."

- DR: In our substitution experiment case, the substitution patterns for those for whom attrition was not affected by the first-round-charity treatment
- @NL: E.g., the impact of an institution on income for those whose choice to remain in the course was not affected by their institutional assignment

---

"Narrowing bounds using covariates"

All of the above could be done conditional on a particular set of baseline characteristics such as gender or race. The average treatment effect could be estimated separately for each. (Note: and perhaps combined in a fruitful way?)

One can alternately use covariates to reduce the width of these bounds. To give intuition, we can imagine a baseline covariate that perfectly predicts an individual's wage. Because treatments are randomly assigned the maintained assumptions will still hold conditionally on this X. The results the methods can be applied separately for each value of this covariate, and for each such value the trimming procedure will actually have no impact on the estimate.

DR: I think this is the "estimate and sum things up in a weighted way" procedure I thought about a moment ago.

Proposition 1B gives the balance estimator for a model involving exogenous variables. Essentially, this computes the corresponding bounds estimator at each X, where the differential selection probability is computed for that particular X, the upper quantile value of the outcome is given conditional on the same X and on being in the treatment group. These are then integrated (or summed up) weighted by the distribution or the cdf of this covariate in the control group. These bounds will necessarily be sharper than the balance without controls.

Section 4: estimation and inference

The asymptotic variance depends on components reflecting the variance of the trimmed distribution, the variance of the estimated trimming threshold, and the variance in the estimate of "how much of the distribution to trim" (the relative selection probability differential).

Equation 6 formally defines the estimator

Estimated bounds consistent for 'true bounds' under standard conditions

Two ways to compute CI's – CI's for the 'true bounds' or CI's for the TE itself. A 95% CI for the former will contain the latter with even greater probability.

Imbens and Manski '04 can be used to derive the latter which are 'more apppropriate here' since the object of interest is the TA and not the 'region of all rationalizable treatment effects. These are built off of a transformation of the estimate UB and LB and max estimated sd of each of these.

- the latter are reported by the 'cie' option in 'leebounds'

Generalisation to monotonicity (without knowing direction of impact of treatment on selection)…

> As an overall procedure, it is asymptotically valid to estimate p, and if positive, trim the treatment group and conduct inference as discussed in Subsections 4.1 and 4.2. And if negative… [do similar]

> though coverage rates for confidence intervals are asymptotically correct, a large discontinuity in the asymptotic variance suggests coverage rates may be inaccurate when sample sizes are small and p0 is "close" to zero … A simple, conservative approach to combining the trimmed and untrimmed intervals is to compute their union

Section 5: Empirical Results

Table 4 gives a step-by-step that is a good way of seeing and understanding the construction of the estimator, and where the 'action' is, in treimming, in components of the SE, etc.

Intervals are 1/14 the width of the equivalent Horowitz/Manski bounds

5.2 using covariates to narrow bounds

> Any baseline covariate will do, as will any function of all the baseline covariates. In the analysis here, a single baseline covariate—which is meant to be a proxy for the predicted wage potential for each individual—is constructed from a linear combination of all observed baseline characteristics. This single covariate is then discretized, so that effectively five groups are formed according to whether the predicted wage is within intervals defined by $\$6 \cdot 75$, \$7, $\$7 \cdot 50$, and $\$8 \cdot 50$.

- @Substitution: this is essentially what I propose we do, but using Ridge Regressions or something similar

  To compute the bounds for the overall average…the group-specific bounds must be averaged, weighted by the proportion (sPr Group J|S0=1,S1=1)

  The estimated asymptotic variance for these overall averages is the sum of (1) a weighted average of the group-specific variances and (2) the (weighted-) mean squared deviation of the group-specific estimates from the overall mean. This second term takes into account the sampling variability of the weights

→ result: 11% narrower bounds

Interesting; possibly do similar for @NL-ed:

> By statistically ruling out any effect more negative than $-0 \cdot 037$, this suggests that after 4 years, the Job Corps enabled program group members to offset at least 35% (and perhaps more) of the potential $0 \cdot 058$ loss in wages due to lost labour market experience that could have been caused by the program

Section 6: Conclusions: implications and applications

Interesting intuitive argument:

> Another reason to interpret the evidence as pointing to positive wage effects is that the lower bound is based on an extreme and unintuitive assumption—that wage outcomes are perfectly negatively correlated with the propensity to be employed. From a purely theoretical standpoint, a simple labour supply model suggests that, all other things equal, those on the margin of being employed will have lowest wages not the highest wages (i.e., the "reservation wage" will be the smallest wage that draws the individual into the labour force). In addition, the empirical evidence in Table 2 suggests that there is positive selection into employment: those who are predicted to have higher wages are more likely to be employed (i.e., U and V are positively correlated). If this is true, it seems relatively more plausible to trim the lower rather than the upper tail of the distribution to get an estimate of the treatment effect.