SFB 823

Lee’s treatment effect bounds for non-random sample selection - an implementation in Stata

Harald Tauchmann

Nr. 35/2013

Discussion Paper

Lee's treatment effect bounds for non-random sample selection – an implementation in Stata

Harald Tauchmann
University of Erlangen-Nuremberg,
and RWI (Rheinisch-Westfälisches Institut für Wirtschaftsforschung),
and CINCH (Centre of Health Economics Research)
Findelgasse 7/9, 90402 Nürnberg, Germany
harald.tauchmann@wiso.uni-erlangen.de

Abstract. Non-random sample selection may render estimated treatment effects biased even if assignment of treatment is purely random. Lee (2009) proposes an estimator for treatment effect bounds that limit the possible range of the treatment effect. In this approach, the lower and upper bound, respectively, correspond to extreme assumptions about the missing information, which are consistent with the observed data. As opposed to conventional parametric approaches to correcting for sample selection bias, Lee's bounds estimator rests on very few assumptions. We introduce the new Stata command `leebounds` that implements the estimator in Stata. The command allows for several options, such as tightening bounds by the use of covariates.

Keywords: non-parametric, randomized trial, sample selection, attrition, bounds, treatment effect.

1 Introduction

Random assignment of treatment provides an ideal setting for identifying treatment effects. Most prominent, randomized trials are exactly designed for generating a situation where randomness of treatment is guaranteed, ruling out any potential endogeneity bias. However, this ideal setting can easily be distorted by non-random sample attrition. Dropout from a program, denied information on the outcome variable, and death during a clinical trial may serve as examples. While treatment is purely random in the original population, this no more holds for the actual estimation sample if attrition is linked to the treatment status, potentially leading to attrition bias with perhaps unknown direction.

Parametrically correcting for attrition and selection bias has developed into a standard procedure in applied empirical research, rendering the seminal method by Heckman (1976, 1979) a work horse of applied econometrics. This procedure is implemented in Stata by the `heckman` command. Yet, this parametric approach has been criticised for relying on restrictive assumptions, in particular, joint normality, and being vulnerable to misspecification (e.g. Puhani 2000; Grasdal 2001), which has led to the development of semi-parametric approaches (e.g. Ichimura and Lee 1991; Ahn and Powell 1993). Though relying on less restrictive distributional assumptions, valid exclusion restrictions are even more essential for these estimators. More recently, bound estimators have been proposed that require only very few assumptions and do not rely on valid exclusion restrictions. These estimators, rather than correcting point estimates for potential bias, determine an interval for the true treatment effect. The interval bases on extreme assumptions about the impact of selection on estimated effect that are consistent with the data. One of such estimators is Horowitz and Manski (2000). This approach does not involve any assumption about the selection mechanism, yet it is only applicable to outcome variables that are bounded to a certain interval since missing information is imputed on basis of minimal and maximal possible values. This impedes its application to numerous problems and regularly yields very wide bounds.

The present paper introduces the new Stata command `leebounds` that facilitates the estimation of alternative bounds prosed by Lee (2009), which impose more structure on the assumed selection mechanism while allowing for outcome variables with unbounded support and often yielding more narrow bounds. The following section gives summary of Lee's bounds estimator. The syntax of `leebounds` is described in section 3. Section 4 illustrates the application of `leebounds`. Section 5 summarizes and concludes the article.

2 The Lee (2009) bounds estimator

2.1 The intuition behind the estimator

Lee (2009) proposes a bounds estimator that estimates an interval for the true value of the treatment effect in the presence of non-random sample selection. It rests on only two assumptions: random assignment of treatment and monotonicity. The latter implies that assignment to the treatment group can affect attrition in only one direction. That means that besides observations for with the outcome variable is observed irrespective of the assigned treatment status, the actual estimation sample either includes observation for which the outcome is observed because of receiving the treatment, or observation for which the outcome is observed because of not receiving the treatment, but not both simultaneously.

The intuition of the bounds estimator is to trim either the sample of the treated or the non-treated observations such that the share of observations with observed outcome is equal for both groups. Trimming is either from above or from below. This corresponds to two extreme assumption about missing information that are consistent with the observed data and a one-sided selection mechanism. That is, in the group that suffers less from attrition either the largest or the smallest values of the outcome are regarded as 'excess observations' and are excluded from the analysis. This implies that precisely the treatment effect on never attriters is subject to estimation. The present paper focusses on the practical issue of how estimates for the bounds are calculated and, in particular, on how this procedure is implemented in Stata; for more theory, we refer to Lee (2009).

2.2 Estimation

Estimating treatment effect bounds as suggested by Lee (2009) is computationally straightforward. Only a raw group mean and two trimmed group means of the outcome variable need to be calculated. Let Y_i denote the outcome, T_i a binary treatment indicator, and S_i a binary selection indicator, with $S_i = 0$ indicating attriters for which Y_i is not observed. As usual, i indexes observations. The shares of observations with observed outcome in the treatment group q_T and its counterpart for the control group q_C can then be written:

$$q_T = \frac{\sum_i 1\left(T_i = 1, S_i = 1\right)}{\sum_i 1\left(T_i = 1\right)} \tag{1}$$

$$q_C = \frac{\sum_i 1\left(T_i = 0, S_i = 1\right)}{\sum_i 1\left(T_i = 0\right)}. \tag{2}$$

Here $1(\cdot)$ denotes the indicator function. To simplify notation, let us consider the case $q_T > q_C$ that is the treatment group suffers less from attrition.[1] Then

$$q = \frac{q_T - q_C}{q_T} \tag{3}$$

and $1 - q$ determines the quantiles at which the distribution of Y in the treatment group are trimmed in order to exclude extreme values of Y from the analysis. Hence

$$y_q^T = G^{-1}_{Y|T=1,S=1}(q) \tag{4}$$

$$y_{1-q}^T = G^{-1}_{Y|T=1,S=1}(1-q) \tag{5}$$

determine the marginal values y_q^T and y_{1-q}^T of the outcome that enter the trimmed means, with G_Y^{-1} denoting the inverse empirical distribution function of Y. Using this notation, estimates for the upper bound and the lower bound are calculated as

$$\hat{\theta}^{\text{upper}} = \frac{\sum_i 1\left(T_i = 1, S_i = 1, Y_i \geq y_q^T\right) Y_i}{\sum_i 1\left(T_i = 1, S_i = 1, Y_i \geq y_q^T\right)} - \frac{\sum_i 1\left(T_i = 0, S_i = 1\right) Y_i}{\sum_i 1\left(T_i = 0, S_i = 1\right)} \tag{6}$$

$$\hat{\theta}^{\text{lower}} = \frac{\sum_i 1\big(T_i = 1, S_i = 1, Y_i \leq y_{1-q}^T\big) Y_i}{\sum_i 1\big(T_i = 1, S_i = 1, Y_i \leq y_{1-q}^T\big)} - \frac{\sum_i 1\left(T_i = 0, S_i = 1\right) Y_i}{\sum_i 1\left(T_i = 0, S_i = 1\right)}. \tag{7}$$

Lee (2009) considers a purely continuous outcome variable Y. Yet, especially in survey data, variables that are inherently continuous are often imprecisely reported, resulting in ‘ties’ in the observed outcome data. Monthly disposable income may serve as an example, for which many individuals tend to report a round number, such as $ 1,000 or $ 1,500. Such ‘ties’ may violate the intuition behind (6) and (7) if the

[1] For the opposite case $q_T < q_C$, all arguments hold symmetrically, with q being defined as $(q_C - q_T)/q_C$, the control group being trimmed at y_q^C and y_{1-q}^C, respectively, and the treatment group remaining untrimmed. For $q_T = q_C$, both, the upper and the lower bound coincide with the difference in raw means.

marginal values y_q^T and y_{1-q}^T are frequent. For this reason, `leebounds` excludes the $q \cdot N_T$ (rounded down to the nearest integer) smallest – respectively largest – values of Y for the calculation of the trimmed means. Here N_T denotes the number of observations in the treatment group for which the outcome variable is observed. This means that a certain fraction of the observations that exhibit the marginal values y_q^T and y_{1-q}^T enter the trimmed means. With no ties in Y, this procedure coincides with (6) and (7).

2.3 Tightening bounds

Estimating Lee (2009) bounds does not involve any covariates. This corresponds to the assumption of random assignment of treatment, under which the differences in conditional and unconditional expectations of Y coincide. Yet, covariates that are determined prior to treatment may be used to tighten treatment effect bounds. Covariates that have some explanatory power for attrition are used to spilt the sample into cells, and bounds are separately calculated for each cell. Finally, a weighted average of cells specific bounds is computed. The appropriate weights are the probabilities of cell membership for never attriters (Lee 2009, 1094). These probabilities are unknown. Yet, because of random assignment of treatment and monotonicity, they can consistently be estimated by $\frac{\sum_i 1(J_i=1, S_i=1, T_i=0)}{\sum_i 1(S_i=1, T_i=0)}$ for each cell J, where $J_i = 1$ indicates membership in J. Lee (2009) shows that such averaged bounds are tighter than those that doe not use any covariates (Lee 2009, 1086).[2] Tightening bounds is offered by `leebounds` as an option.

Technically, only a limited number of discrete[3] variables can be used for tightening, as the number of observations und the joint distribution of treatment status and selection must allow for estimating the bounds for each individual cell. Thus, estimation regularly fails if a large number of covariates is used. Tightening may also fail if for some cells the control group suffers relatively more from attrition, while for other cells attrition is more frequent in the treatment group. Due to sampling error, this will frequently occur, if the sample is split in too many cells.[4] `leebounds` checks for this, issues a warning if a selection pattern is detected that is heterogeneous across cells and saves a macro indicating the type of the selection pattern.

2.4 Standard errors and inference

Estimates for the treatment effect bounds are subject to sampling error. Lee (2009, 1088) provides analytic standard errors for them; we refer to the original paper for details about the calculation of standard errors. Analytical standard errors, and as an alternative bootstrapped standard errors, are implemented in `leebounds`. On basis of these standard errors, one may determine 'naive' confidence intervals that cover the interval $[\theta^{\text{lower}}, \theta^{\text{upper}}]$ with probability $1-\alpha$. Interestingly, based on Imbens and Manski (2004), Lee (2009, 1089) also derives a confidence interval for the treatment effect itself,

[2] The proof is for the population parameters, not for their sample analogues. Hence, especially for ill-suited covariates, estimated bounds may fail in getting tighter with option `tight()`.

[3] In practice, continuous variables (e.g. age) must be transformed into categorial ones (age classes).

[4] This may also provide indication for a violation of the monotonicity assumption.

i.e. the scalar parameter of ultimate interest. This interval is tighter than the combined confidence interval for θ^{lower} and θ^{upper}. It captures both, uncertainty about the bias due to non-random sample attrition and uncertainty because of sampling error. `leebounds` optionally provides estimation of the confidence interval for the treatment effect.

3 The leebounds command

`leebounds` requires Stata 11 or higher. The prefix commands `by` and `svy` are not allowed. The prefix command `bootstrap` is allowed, yet, its use is not recommended. `pweights` (default), `fweights`, and `iweights` are allowed, `aweights` are not allowed. Observations with a negative weight are skipped for any type of weight.

3.1 Syntax for leebounds

The syntax for `leebounds` reads as follows:

```
leebounds depvar treatvar [if] [in] , [ select(varname) tight(varlist)
   cieffect vce(analytic|bootstrap) level(#) ]
```

depvar is a numeric outcome variable and *treatvar* is a binary treatment indicator, which can either be numeric or a string variable. The (alphanumerically) lager value of *treatvar* is assumed to indicate treatment.

3.2 Options for leebounds

`select(`*varname*`)` specifies a binary selection indicator. *varname* may only take the value zero or one. If no selection indicator is specified, any observation with non-missing information on *depvar* is assumed to be selected while all observations with missing information on *depvar* are assumed to be not selected.

`tight(`*varlist*`)` specifies a list of covariates for computing tightened bounds. With `tight()` specified, the sample is split into cells defined by the covariates in *varlist*. Continuous variables in *varlist* will cause failure of the estimation procedure.

`cieffect` requests calculation of a confidence interval for the treatment effect.

`vce(analytic|bootstrap)` specifies whether analytic or bootstrapped standard errors are calculated for estimated bounds. `analytic` is the default. `bootstrap` allows for the suboptions `reps(#)` and `nodots`. For `vce(analytic)` the covariance for the estimated lower and upper bound is not computed. If this covariance is of relevance, one should choose `vce(bootstrap)`. Instead of specifying `vce(bootstrap)` one may alternatively use the prefix command `bootstrap`, which allows for numerous additional options. Yet `leebounds`' internal bootstrapping routine is much faster than the prefix command, allows for sampling weights by performing a weighted bootstrap, and makes also the option `cieffect` use bootstrapped standard errors.

`level(#)` as usual sets the level of confidence. One may change the reported confidence level by retyping `leebounds` without arguments and only specifying the option `level(#)`. However, this affects only the confidence interval for the bounds, but not for the confidence interval requested with `cieffect`.

3.3 Saved results for leebounds

`leenounds` saves the following results to `e()`:

Scalars			
`e(N)`	number of observations	`e(cilower)`	lower bound of treatment effect-confidence interval (only for option `cieffect`)
`e(Nsel)`	number of selected obs.	`e(ciupper)`	upper bound of treatment effect-confidence interval (only for option `cieffect`)
`e(trim)`	(overall) trimming proportion	`e(level)`	level of confidence
`e(cells)`	number of cells (only for option `tight()`)	`e(N_reps)`	number of bootstrap repetitions (only for option `vce(bootstrap)`)
Macros			
`e(cmd)`	leebounds	`e(select)`	*varname* (only for option `select()`)
`e(cmdline)`	command as typed	`e(cellsel)`	cell-specific selection pattern, `homo`, or `hetero` (only for option `tight()`)
`e(title)`	Lee (2009) treatment effect bounds	`e(covariates)`	*varlist* (only for option `tight()`)
`e(vce)`	either `analytic` or `bootstrap`	`e(trimmed)`	either treatment or control
`e(vcetype)`	Bootstrap for `vce(bootstrap)`	`e(wtype)`	either `pweight`, `fweight`, or `iweight` (if weights are specified)
`e(depvar)`	*depvar*	`e(wexp)`	`= exp` (if weights are specified)
`e(treatment)`	*treatvar*	`e(properties)`	`b V`
Matrices			
`e(b)`	1×2 vector of estimated treatment effect bounds (column names are *treatvar:lower*, *treatvar:upper*)	`e(V)`	2×2 variance-covariance matrix for estimated treatment effect bounds (covariance set to zero for `vce(analytic)`)
Functions			
`e(sample)`	marks estimation sample		

4 Example for using the leebounds command

We use Stata's `cancer.dta` example data set for a simple purely illustrative application. We analyse whether being treated with an active ingredient (`drug == 2 | drug == 3`), compared to being treated with a placebo (`drug == 1`), has an effect on survival time (`studytime`). We treat the data as if information on survival time were only available for those, who died during the study (`died == 1`). This is not entirely correct, as for those who did not die (`died == 0`), after all we know that they survived at least for the rest of the study period. Yet, for illustration, we regard them as attritters without any (valid) information on the outcome `studytime`.

```
. sysuse cancer.dta, clear
(Patient Survival in Drug Trial)

. gen activedrug = (drug == 2 | drug == 3)

. leebounds studytime activedrug, select(died)

Lee (2009) treatment effect bounds

Number of obs.                  =   48
Number of selected obs.         =   31
Trimming porportion             =   0.5489

------------------------------------------------------------------------------
   studytime |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
activedrug   |
       lower |   2.866667   3.909154     0.73   0.463    -4.795134    10.52847
       upper |       14.3   3.163771     4.52   0.000     8.099123    20.50088
------------------------------------------------------------------------------
```

The output displays that 48 individuals participated in the trial. Out of these 31 died during the study, while 17 survived. The latter are regarded as not selected as we have no precise information about survival time. The trimming proportion corresponds to q, see equation (3). The value 0.5489 indicates that the control group is trimmed by more than half, as the survival rate is much higher among individuals who were treated with an active drug. Correspondingly, the estimated treatment effect bounds are pretty wide ranging from 2.87 to 14.30 months gain in survival time. Taking standard errors into account, the lower bound even does not significantly deviate from zero. For obtaining a confidence interval for the treatment effect (see section 2.3), one can choose the `cieffect` option:

```
. leebounds studytime activedrug, select(died) cieffect

Lee (2009) treatment effect bounds

Number of obs.                  =   48
Number of selected obs.         =   31
Trimming porportion             =   0.5489
Effect 95% conf. interval       : [-3.5633  19.5039]

------------------------------------------------------------------------------
   studytime |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
activedrug   |
       lower |   2.866667   3.909154     0.73   0.463    -4.795134    10.52847
       upper |       14.3   3.163771     4.52   0.000     8.099123    20.50088
------------------------------------------------------------------------------
```

This interval, is narrower than the combined confidence intervals for the bounds. One may allow for a less strikt level of confidence by specifying `level(90)` and opt for bootstrapped rather than analytic standard errors:

```
. set seed 13052007

. leebounds studytime activedrug, sel(died) cie level(90) vce(boot, reps(250))
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
```

```
Lee (2009) treatment effect bounds

Number of obs.                   =   48
Number of selected obs.          =   31
Trimming porportion              =   0.5489
Effect 90% conf. interval        : [-1.9390  18.1498]

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
   studytime |      Coef.   Std. Err.      z    P>|z|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
activedrug   |
       lower |   2.866667   3.749864     0.76   0.445    -3.301311    9.034644
       upper |       14.3    3.00403     4.76   0.000     9.358811    19.24119
------------------------------------------------------------------------------
```

Bootstrapped standard errors are similar to their analytical counterparts. Even the 90-percent confidence interval for the treatment effect overlaps the value of zero. Finally we try to tighten the bounds by the use of a covariate. The only available is `age`, which we have to transform into a categorial variable. Here we choose three age categories such that each category has roughly the same number ob observations:

```
. _pctile age, percentiles(33 66)

. gen agecat = recode(age,r(r1),r(r2),100)

. leebounds studytime activedrug, select(died) cieffect tight(agecat)

Tightened Lee (2009) treatment effect bounds

Number of obs.                   =   48
Number of selected obs.          =   31
Number of cells                  =   3
Overall trimming porportion      =   0.5489
Effect 95% conf. interval        : [ 0.1028  19.6897]

------------------------------------------------------------------------------
   studytime |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
activedrug   |
       lower |          7   4.155293     1.68   0.092    -1.144225    15.14423
       upper |   12.55556    4.29805     2.92   0.003     4.131531    20.97958
------------------------------------------------------------------------------
```

Tightening yields much narrower bounds for the treatment effect. Indeed, with specifying the `tight()` option, the 95-percent effect confidence interval does not include the value zero.

5 Summary and conclusions

In this article, the new command `leebounds` was introduced that implements Lee (2009) treatment effect bounds for data with random assignment of treatment that suffer from non-random sample selection. In addition to calculating point estimates for the bounds, the command accommodates the calculation of confidence intervals for the treatment effect and tightened bounds based on covariates. `leebounds` complements the contributions of Beresteanu and Manski (2000) and Palmer et al. (2011), who have already made other bounds estimators available to Stata users.

6 Acknowledgements

This work has been supported in part by the Collaborative Research Center "Statistical Modelling of Nonlinear Dynamic Processes" (SFB 823) of the German Research Foundation (DFG). The author is grateful to Arndt R. Reichert, Domenico Depalo, and participants of the 2012 German Stata Users Group meeting for many valuable comments. This article is meant for being submitted to the Stata Journal.

7 References

Ahn, H., and J. L. Powell. 1993. Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism. *Journal of Econometrics* 58: 3–29.

Beresteanu, A., and C. F. Manski. 2000. Bounds for STATA: Draft Version 1.0. Northwestern University.

Grasdal, A. 2001. The performance of sample selection estimators to control for attrition bias. *Health Economics* 10: 385–398.

Heckman, J. J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models. *Annals of Economics and Social Measurement* 5: 475–492.

———. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47: 153–161.

Horowitz, J. L., and C. F. Manski. 2000. Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data. *Journal of the American Statistical Association* 95: 77–84.

Ichimura, H., and L. Lee. 1991. Semiparametric least squares Estimation of Multiple Index Models: Single Equation Estimation. vol. 5 of *International Symposia in Economic Theory and Econometrics*, 3–32. Cambridge University Press.

Imbens, G., and C. F. Manski. 2004. Confidence Intervals for Partially Identified Parameters. *Econometrica* 72: 1845–1857.

Lee, D. S. 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies* 76: 1071–1102.

Palmer, T. M., R. R. Ramsahai, V. Didelez, and N. A. Sheehan. 2011. Nonparametric bounds for the causal effect in a binary instrumental-variable model. *The Stata Journal* 11: 345–367.

Puhani, P. 2000. The Heckman Correction for Sample Selection and its Critique. *Journal of Economic Surveys* 14: 53–68.