`leebounds`: Lee's (2009) treatment effects bounds for non-random sample selection for Stata

Harald Tauchmann (RWI & CINCH)

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
& CINCH Health Economics Research Centre

1. June 2012

2012 German Stata Users Group Meeting, WZB, Berlin

Introduction

- Random assignment of treatment: ideal setting for estimating treatment effects
  - $\rightarrow$ Randomized trials
- Non-random sample attrition (selection) still undermines validity of econometric estimates
  - $\rightarrow$ Selection bias
- Typical examples:
  - Dropout from program
  - Denied information on outcome
  - Death during clinical trial
- Possibly severe attrition bias
- Direction of bias a priory unknown

Selection Correction Estimators

- Modeling the mechanism of sample selection/attrition
- Classical Heckman (1976, 1979) **parametric** selection correction estimator
  - Stata command **`heckman`**
  - Assumes *joint normality*
  - Exclusion restrictions beneficial
  - Identification through non-linearity – in principle – possible
  - → Parametric approach relying on strong assumptions
- **Semi-parametric** approaches (e.g. Ichimura and Lee, 1991; Ahn and Powell, 1993)
  - Assumption of joint normality not required
  - Exclusion restrictions essential
  - → Valid exclusion restrictions may not be available

Treatment Effect Bounds

- Rather than correcting point estimate of treatment effect
- Determining interval for effect size
- Correspond to extreme assumptions about the impact of selection on estimated effect

1. **Horowitz and Manski (2000) bounds**
    - No assumptions about the the selection mechanism required
    - Outcome variable needs to be bounded
    - Missing information is imputed an basis of minimal and maximal possible values of the outcome variable

    $\rightarrow$ Frequently yields very wide (i.e. hardly informative) bounds

    $\rightarrow$ Useful benchmark for binary outcome variables

Treatment Effect Bounds II

2. Lee (2009) bounds

Assumptions:

(i) Besides *random assignment of treatment*

(ii) *Monotonicity* assumption about selection mechanism

- Assignment to treatment can only affect attrition in one direction
- I.e. (in terms of sign) no heterogeneous effect of treatment on selection
- Average treatment effect for never-attriters

Intuition:

- Sample trimmed such that the share of observed individuals is equal for both groups
- Trimming either from above or from below
- Corresponds to extreme assumptions about missing information that are consistent with
  (i) The observed data and
  (ii) A one-sided selection model

Estimating Lee (2009) bounds

Let denote Y the outcome, T a binary treatment indicator, W a binary selection indicator, and i individuals. Calculate:

1. $q_T \equiv \frac{\sum_i 1(T_i=1,W_i=1)}{\sum_i 1(T_i=1)}$ and $q_C \equiv \frac{\sum_i 1(T_i=0,W_i=1)}{\sum_i 1(T_i=0)}$,
   i.e. the shares of individuals with observed Y
2. $q \equiv {}^{(q_T - q_C)}\!/_{q_T}$, if $q_T > q_C$ (If $q_T < q_C$, exchange C for T)
3. $y_q^T = G_Y^{-1}(q|T=1,W=1)$ and $y_{1-q}^T = G_Y^{-1}(1-q|T=1,W=1)$,
   i.e. qth and the $(1-q)$th quantile of observed outcome in the treatment group
4. Upper bound $\hat{\theta}^{upper}$ and lower bound $\hat{\theta}^{lower}$ as

$$
\begin{aligned}
\hat{\theta}^{upper} &= \frac{\sum_i 1\left(T_i=1,W_i=1,Y_i \geq y_q^T\right) Y_i}{\sum_i 1\left(T_i=1,W_i=1,Y_i \geq y_q^T\right)} - \frac{\sum_i 1\left(T_i=0,W_i=1\right) Y_i}{\sum_i 1\left(T_i=0,W_i=1\right)} \\
\hat{\theta}^{lower} &= \frac{\sum_i 1\left(T_i=1,W_i=1,Y_i \leq y_{1-q}^T\right) Y_i}{\sum_i 1\left(T_i=1,W_i=1,Y_i \leq y_{1-q}^T\right)} - \frac{\sum_i 1\left(T_i=0,W_i=1\right) Y_i}{\sum_i 1\left(T_i=0,W_i=1\right)}
\end{aligned}
$$

Tightening Bounds

- Lee (2009) bounds rest on comparing unconditional means of (trimmed) subsamples

→ No covariates considered

- Using covariates yields tighter bounds:
  1. Choose (discrete) variable(s) that have explanatory power for attrition
  2. Split sample into cells defined by these variables
  3. Compute bounds for each cell
  4. Take weighted average

  → Lee (2009) shows that such bounds are tighter than unconditional ones

- Researcher can generate such variables by deliberately varying the effort on preventing attrition (DiNardo et al., 2006)

Standard Errors and Confidence Intervals

- Lee (2009) derives analytic standard errors for bounds
- Allows for straightforward calculation of a 'naive' confidence interval
- Covers the *interval* $[\theta^{lower}, \theta^{upper}]$ with probability $1 - \alpha$
- Imbens and Manski (2004) derive confidence interval for the *treatment effect* itself
- Tighter than confidence interval for the interval

leebounds: Syntax

Viewer - help leebounds

File Edit History Help

help leebounds

Dialog Also See Jump To

help orderalpha

Title

leebounds — Lee (2009) treatment effect bounds

Syntax

leebounds *depvar* *treatvar* [*if*] [*in*] [*weight*], [*options*]

Outcome and treatment	Description
Model	
depvar	dependent variable
treatvar	binary treatment indicator

options	Description
Main	
select(*varname*)	selection indicator
tight(*varlist*)	covariats for tightened bounds
cieffect	compute confidence interval for treatment effect
SE/Bootstrap	
vce(_analytic_|_bootstrap_**)**	compute analytic or bootstrapped standard errors; **vce(**_analytic_**)** is the default.
Reporting	
level(_#_**)**	set confidence level; default is **level(**_95_**)**

pweights (default), **fweights**, and **iweights** are allowed, **aweights** are not allowed; see weight. Observations with negative weight are skipped for any weight type.
bootstrap is allowed, **by** and **svy** are not allowed; see prefix.

Ready CAP NUM OVR

leebounds: Saved Results

Viewer - help leebounds

File Edit History Help

help leebounds

Dialog Also See Jump To

Saved results

leebounds saves the following in **e()**:

Scalars

e(N)	number of observations
e(Nsel)	number of selected observations
e(trim)	(overall) trimming proportion
e(cells)	number of cells (only saved for **tight()**)
e(cilower)	lower bound of treatment effect-confidence interval (only saved for **cieffect**)
e(ciupper)	upper bound of treatment effect-confidence interval (only saved for **cieffect**)
e(level)	confidence level
e(N_reps)	number of bootstrap repetitions (only saved for **vce(***bootstrap***)**)

Macros

e(cmd)	**leebounds**
e(cmdline)	command as typed
e(title)	**Lee (2009) treatment effect bounds**
e(vce)	either **analytic** or **bootstrap**
e(vcetype)	**Bootstrap** for **vce(***bootstrap***)**
e(depvar)	*depvar*
e(treatment)	*treatvar*
e(select)	*varname* (only saved for **select()**)
e(covariates)	*varlist* (only saved for **tight()**)
e(trimmed)	either **treatment** or **control**
e(wtype)	either **pweight**, **fweight**, or **iweight** (only saved if weights are specified)
e(wexp)	= *exp* (only saved if weights are specified)
e(properties)	**b V**

Matrices

e(b)	*1x2* vector of estimated treatment effect bounds (*colnames* are of the form *treatvar:lower* and *treatvar:upper*)
e(V)	*2x2* variance-covariance matrix for estimated treatment effect bounds (covariance set to zero for **vce(***analytic***)**)

Functions

e(sample)	marks estimation sample

Ready CAP NUM OVR

Experimental Design

Research question: Do financial incentives aid obese in reducing bodyweight?

- Ongoing randomized trial (Augurzky et al., 2012)
- 698 obese (BMI $\geq$ 30) individuals recruited during rehab hospital stay
- Individual weight-loss target (typically 6–8% of body weight)
- Participants prompted to realize weight-loss target within four months
- Randomly assigned to on of three experimental groups:
  i. No financial incentive (control group)
  ii. 150 € reward for realizing weight-loss target
  iii. 300 € reward for realizing weight-loss target
- After four months: weight-in at assigned pharmacy

Attrition Problem

Experimental groups:

	group size	compliers	attrition
control group	233	155	33.5%
150 € group	236	172	27.1%
300 € group	229	193	15.7%
	698	520	25.5%

- Attrition rate negatively correlated with size of reward
- Plausible since (successful) members of incentive group have stronger incentive not to dropout
- Selection on success (in particular for incentive groups) likely
- Overestimation of incentive effect likely
  downward bias still possible

Simple Bivariate OLS (comparison of means)

- Outcome variable: ***weightloss*** (percent of body weight)
- Focus on comparing ***group 300 €*** with ***control group***

```
. regress weightloss group300

      Source |       SS       df       MS              Number of obs =     348
-------------+------------------------------           F(  1,   346) =   23.17
       Model |  686.575435     1  686.575435           Prob > F      =  0.0000
    Residual |  10253.2078   346  29.6335486           R-squared     =  0.0628
-------------+------------------------------           Adj R-squared =  0.0601
       Total |  10939.7832   347  31.5267528           Root MSE      =  5.4437

------------------------------------------------------------------------------
  weightloss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    group300 |   2.826111   .5871336     4.81   0.000     1.671311    3.980911
       _cons |    2.34758   .4372461     5.37   0.000     1.487585    3.207575
------------------------------------------------------------------------------
```

- Highly significant inventive effect
- Roughly three percentage points

Heckman (two-step) Selection Correction Estimator

- Exclusion restriction: ***nearby_pharmacy***
  (assigned pharmacy within same ZIP-code area as place of residence)
- Captures cost of attending weight-in, no direct link to weight loss
- No further controlls
- Two-step estimation

Heckman (two-step) Selection Correction Estimator II

```
. heckman weightloss group300, select(group300 nearby_pharmacy) twostep

Heckman selection model -- two-step estimates     Number of obs      =       462
(regression model with sample selection)          Censored obs       =       114
                                                  Uncensored obs     =       348

                                                  Wald chi2(1)       =      1.37
                                                  Prob > chi2        =    0.2415
```

weightloss	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weightloss						
group300	3.126055	2.669154	1.17	0.242	-2.105391	8.357501
_cons	1.716602	5.493513	0.31	0.755	-9.050485	12.48369
select						
group300	.5777289	.1312605	4.40	0.000	.3204631	.8349947
nearby_phar~y	.1358984	.1344283	1.01	0.312	-.1275763	.399373
_cons	.3406349	.1201113	2.84	0.005	.1052211	.5760487
mills						
lambda	1.158006	10.04912	0.12	0.908	-18.5379	20.85392
rho	0.21123					
sigma	5.4821209					

- Similar point estimate as for OLS
- Large S.E.s $\rightarrow$ insignificant incentive effect
- Low explanatory power of *nearby_pharmacy* (if regional characteristics are not controlled for)

Lee Bounds

```
. leebounds weightloss group300

Lee (2009) treatment effect bounds

Number of obs.                    =   462
Number of selected obs.           =   348
Trimming porportion               =   0.2107
```

weightloss	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
group300						
lower	.983459	.6431066	1.53	0.126	-.2770069	2.243925
upper	4.783921	.6677338	7.16	0.000	3.475187	6.092655

- Bounds cover OLS and Heckman point estimate
- Fairly wide interval
- Lower bound does not significantly differ from zero

Lee Bounds with Effect Confidence Interval

```
. leebounds weightloss group300, cie

Lee (2009) treatment effect bounds

Number of obs.                   =   462
Number of selected obs.          =   348
Trimming porportion              =   0.2107
Effect 95% conf. interval        : [-0.0744  5.8822]
```

weightloss	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
group300						
lower	.983459	.6431066	1.53	0.126	-.2770069	2.243925
upper	4.783921	.6677338	7.16	0.000	3.475187	6.092655

- Effect confidence interval covers zero

Tightened Lee Bounds

- Variable *nearby_pharmacy* used for tightening bounds
- Following the suggestion of DiNardo et al. (2006)

```
. leebounds weightloss group300, cie tight(nearby_pharmacy)

Tightened Lee (2009) treatment effect bounds

Number of obs.                  =    462
Number of selected obs.         =    348
Number of cells                 =    2
Overall trimming porportion     =    0.2107
Effect 95% conf. interval       : [-0.0595  5.8448]
```

weightloss	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
group300						
lower	1.000043	.6441664	1.55	0.121	-.2625003	2.262585
upper	4.727485	.6792707	6.96	0.000	3.396139	6.058831

- Bounds just marginally tighter
- Effect confidence interval still covers zero

Tightened Lee Bounds II

Further covariates for tightening bounds:

i. *age50* (indicator for age ≤ 50)

ii. *woman* (indicator for sex)

```
. leebounds weightloss group300, cie tight(nearby_pharmacy age50 woman)

Tightened Lee (2009) treatment effect bounds

Number of obs.                   =   462
Number of selected obs.          =   348
Number of cells                  =   8
Overall trimming porportion      =   0.2107
Effect 95% conf. interval        : [ 0.0608  5.3804]
```

weightloss	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
group300						
lower	1.282951	.7429877	1.73	0.084	-.1732782	2.73918
upper	4.065244	.7995777	5.08	0.000	2.498101	5.632388

- Bounds substantially tighter
- Effect confidence interval does not covers zero
- Confirms existence of incentive effect
- Size of (potential) attrition bias remains somewhat unclear

References

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism, *Journal of Econometrics* **58**: 3–29.

Augurzky, B., Bauer, T. K., Reichert, A. R., Schmidt, C. M. and Tauchmann, H. (2012). Does money burn fat? Evidence from a randomized experiment, *mimeo* .

DiNardo, J., McCrary, J. and Sanbonmatsu, L. (2006). Constructive Proposals for Dealing with Attrition: An Empirical Example, *University of Michigan Working Paper* .

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.

Heckman, J. J. (1979). Sample selection bias as a specification error, *Econometrica* **47**: 153–161.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data, *Journal of the American Statistical Association* **95**: 77–84.

Ichimura, H. and Lee, L. (1991). Semiparametric least squares estimation of multiple index models: Single equation estimation, Vol. 5 of *International Symposia in Economic Theory and Econometrics*, Cambridge University Press, pp. 3–32.

Imbens, G. and Manski, C. F. (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.