Econometrics III: Applied Methods

NovaSBE, 2026

Problem Set 2: IV

Exercise 1: IV and Spillovers

This exercise explores instrumental variables estimation when there may be spillover effects between treated and untreated individuals. You will learn to extend the potential outcomes framework to account for general equilibrium effects.

This exercise is borrowed from “Do Labor Market Policies have Displacement Effects?”, by B. Crépon et al., *Quarterly Journal of Economics*, 2013.

We consider a labor market program that consists in counseling the unemployed on their labor market search. We want to evaluate the impact of this program on the propensity to find a job some time after the program (say 6 months). A randomized controlled trial has taken place within several unemployment agencies: half of the unemployed in each of the agencies were randomly offered the program. They could refuse or enter counseling.

We call $Z = 0, 1$ a dummy variable that is equal to 1 if the unemployed has been randomly assigned to the program, and 0 otherwise. We call $T = 0, 1$ a dummy variable that is equal to 1 if the unemployed has effectively entered the counseling program, and 0 otherwise. We call $y = 0, 1$ an outcome variable that is equal to 1 if the unemployed has found a job after 6 months, and 0 otherwise. Z, T and y are all observed in the data.

1. In this context, what do we call ‘compliers’ (C), ‘always takers’ (A), ‘never takers’ (N), ‘defiers’ (D)? Define them formally and intuitively.

2. The following table presents the joint distribution of Z and T in the data. We assume there are no defiers in this population. What is the proportion of ‘compliers’, ‘always takers’, ‘never takers’ in the population of unemployed? Explain.

   Joint distribution of assignment and actual treatment

	$Z = 1$	$Z = 0$
$T = 1$	0.22	0.00
$T = 0$	0.28	0.50

3. Use the following table to compute the ‘Intention to treat’ parameter. Explain.

   Proportion in employment ($y = 1$) depending on assignment to treatment

$Z = 1$	0.377
$Z = 0$	0.360

4. Explain in words what is the Local Average Treatment Effect (LATE), and estimate it.

5. We now assume that there can be spillovers: the unemployed within each unemployment agency compete for the available job openings. There is thus a possibility that the unemployed that were actually helped by a counselor had better access to jobs, at the expense of the non-treated. In that case, would this experiment overestimate or underestimate the relevance of the counseling policy? Explain.

   To overcome that issue, an additional randomization took place. Within a large set of employment agencies, agencies that took part into the program were determined randomly. In those agencies, the program was implemented as described above (individual randomization); the data so far was from those agencies only. In control agencies, the program was not implemented at all. We call $P = 0, 1$ a dummy variable that is equal to 1 if the unemployed person belongs to an agency that has been randomized into the program, and 0 if she belongs to a control agency.

6. We define counterfactual outcomes as $y(t, p)$: each counterfactual depends on whether the person is treated ($t = 0, 1$) and whether she belongs to a treated agency ($p = 0, 1$) (thus is exposed to a spillover). In this design, what values of $y(t, p)$ are observed and for whom?

7. Formally give the expression of the LATE estimated within treated agencies (in question 4) *in terms of this counterfactual.* If needed, use the notations π_C, π_A and π_N for the proportions of compliers, always takers and never takers, respectively, in the population.

8. Using $y(t, p)$, define the spillover effect on the individuals that are not treated. Suggest a way to estimate it and write down explicitly on what populations this parameter would be estimated (compliers, etc.).

9. We could define the full effect of the treatment as $y(1, 1) - y(0, 0)$. We wonder if we can estimate this parameter on a well-defined population. To answer this question, decompose $E(y|Z = 1, P = 1) - E(y|P = 0)$ into direct effect and spillover effect and comment your findings.

Exercise 2: LATE when treatment takes a finite number of values

This exercise extends the LATE framework to settings where treatment intensity varies. You will derive the weighted average interpretation of IV when treatment is multi-valued.

This exercise is taken from Angrist and Imbens (1995)[1]. It derives an extension to the LATE when treatment takes a finite number of discrete values.

We consider a treatment variable (say education) that can take 3 values: $S = 0, 1, 2$. Outcome (say earnings) is a function $y(S)$ of schooling. We have a dummy instrument (say quarter of birth), $Z = 0, 1$. Call S_0 the counterfactual schooling level when an individual is subject to $Z = 0$ and S_1 her counterfactual schooling level when she is subject to $Z = 1$.

We assume

Independence: $S_0, S_1, y(0), y(1), y(2) \perp Z$

Monotonicity: Either $S_1 \geq S_0$ or $S_1 \leq S_0$ for everyone (we will assume $S_1 \geq S_0$)

[1] "Two-stage least squares estimation of average causal effects in models with variable treatment intensity", *Journal of the American Statistical Association*, 1995, 90(430), 431–442.

1. What does the monotonicity assumption mean?
2. We use the notation $I(S \geq 1)$ and $I(S \geq 2)$, for dummies equal to one if $S \geq 1$ and $S \geq 2$ respectively. Express y, the observed outcome, as a function of these dummies and the counterfactual outcomes.
3. Show that:

$$\begin{aligned} E(y|Z=1) - E(y|Z=0) = & E\big([y(1) - y(0)] \times [I(S_1 \geq 1) - I(S_0 \geq 1)]\big) \\ & + E\big([y(2) - y(1)] \times [I(S_1 \geq 2) - I(S_0 \geq 2)]\big) \end{aligned}$$

   What assumption did you require?
4. Show that:

$$\begin{aligned} & E\big([y(1) - y(0)] \times [I(S_1 \geq 1) - I(S_0 \geq 1)]\big) \\ = & E\big(y(1) - y(0) \,|\, I(S_1 \geq 1) - I(S_0 \geq 1) = 1\big) \times P\big(I(S_1 \geq 1) - I(S_0 \geq 1) = 1\big) \end{aligned}$$

   What assumption did you require?
5. Explain why $P\big(I(S_1 \geq 1) - I(S_0 \geq 1) = 1\big) = P(S_1 \geq 1 > S_0)$. What kind of individuals does $(S_1 \geq 1 > S_0)$ characterize?
6. We assume with no proof that:

$$E(S|Z=1) - E(S|Z=0) = P(S_1 \geq 1 > S_0) + P(S_1 \geq 2 > S_0)$$

   Show that:

$$\frac{E(y|Z=1) - E(y|Z=0)}{E(S|Z=1) - E(S|Z=0)} = \sum_{j=1}^{2} \omega_j E\big(y(j) - y(j-1) \,|\, S_1 \geq j > S_0\big)$$

   where

$$\omega_j = \frac{P(S_1 \geq j > S_0)}{\sum_{k=1}^{2} P(S_1 \geq k > S_0)}$$

7. Interpret this expression.

Exercise 3: Bounding Treatment Effects with Sample Selection

This exercise introduces partial identification and bounds when outcomes are missing due to sample selection. You will implement Manski and Lee bounds using real experimental data.

This exercise introduces you to the problem of **sample selection** (also called attrition) in treatment effect estimation. When the outcome is only observed for a selected subsample, and treatment affects selection, naive comparisons can be biased. We will derive bounds on the treatment effect that are valid despite this selection problem.

Note: This exercise applies the bounding approach from Lee (2009, "Training, Wages, and Sample Selection", Review of Economic Studies). Lee uses a slightly different NSW sample and includes covariates to sharpen bounds. Your results may differ but should be in a similar range.

Background: The NSW Job Training Program

We use data from the National Supported Work (NSW) Demonstration, a randomized job training program conducted in the mid-1970s in the United States. The program provided guaranteed employment for 9–18 months to disadvantaged workers, along with close supervision and peer support. Participants were randomly assigned to treatment or control.

The dataset `lalonde_nsw.csv` contains 445 observations with the following variables:

- `treat`: Treatment indicator (1 = received NSW training, 0 = control)
- `re78`: Real earnings in 1978 (our outcome of interest)
- `age`, `education`, `black`, `hispanic`, `married`, `nodegree`: Baseline covariates
- `re74`, `re75`: Pre-program earnings (1974, 1975)

The Sample Selection Problem

A natural question is: what is the effect of training on *wages*? However, wages are only observed for those who are employed. Define:

- $S_i \in \{0, 1\}$: Employment status (1 if employed, i.e., `re78` > 0)
- Y_i^*: Latent wages (what the person *would* earn if employed)
- $Y_i = S_i \cdot Y_i^*$: Observed outcome (wages if employed, missing/zero otherwise)

The problem: if treatment affects both employment (S) and wages (Y^*), then comparing wages only among the employed conflates these two effects. The "always-employed" in treatment may differ systematically from the "always-employed" in control.

Potential Outcomes Framework

Let $S_i(1)$ and $S_i(0)$ denote employment status under treatment and control. Let $Y_i^*(1)$ and $Y_i^*(0)$ denote potential wages. We observe:

$$S_i = D_i \cdot S_i(1) + (1 - D_i) \cdot S_i(0)$$

where D_i is treatment assignment. We can classify individuals:

Type	$S_i(0)$	$S_i(1)$	**Description**
Always-employed (EE)	1	1	Employed regardless of treatment
Compliers (NE)	0	1	Employed only if treated
Defiers (EN)	1	0	Employed only if untreated
Never-employed (NN)	0	0	Not employed regardless

The causal effect on wages, $Y_i^*(1) - Y_i^*(0)$, is only well-defined for the always-employed (EE), since wages require employment.

Questions

Use R or Stata for the empirical questions. Create the following variables:

- `employed`: Binary indicator equal to 1 if `re78` > 0

- `wage`: Equal to `re78` if employed, `NA` otherwise

1. **Baseline effects.** Estimate the effect of treatment on:

   (a) Total earnings (`re78`)

   (b) Employment probability (`employed`)

   (c) Wages conditional on employment (`wage`, using only employed observations)

   Is the treatment effect on employment statistically significant? What does this imply for interpreting the wage effect?

2. **The selection problem.** The NSW program guaranteed a job during the program period, which ended before 1978. Nevertheless, treated individuals may have higher employment in 1978 due to skills acquired or job connections made.

   (a) If treatment increases employment, would you expect the "marginal" workers brought into employment by treatment to have higher or lower wages than the average employed worker? Why?

   (b) Given your answer, would you expect the naive wage comparison (treated employed vs. control employed) to overestimate or underestimate the true wage effect for always-employed workers?

3. **Manski bounds (worst-case).** Without any assumptions about selection, we can still bound the treatment effect. The idea: for individuals with missing outcomes, assume the "worst case".

   Let $[y_L, y_U]$ be the support of wages (e.g., use the observed min and max among employed). The **Manski bounds** are:

   $$\begin{aligned} \text{Lower bound:} \quad & \mathbb{E}[Y|D=1,S=1]\cdot P(S=1|D=1) + y_L \cdot P(S=0|D=1) \\ & - \mathbb{E}[Y|D=0,S=1]\cdot P(S=1|D=0) - y_U \cdot P(S=0|D=0) \end{aligned}$$

   $$\begin{aligned} \text{Upper bound:} \quad & \mathbb{E}[Y|D=1,S=1]\cdot P(S=1|D=1) + y_U \cdot P(S=0|D=1) \\ & - \mathbb{E}[Y|D=0,S=1]\cdot P(S=1|D=0) - y_L \cdot P(S=0|D=0) \end{aligned}$$

   (a) Explain the intuition: why do we use y_L for treated non-employed and y_U for control non-employed when computing the lower bound?

   (b) Compute the Manski bounds for the treatment effect on wages. Are these bounds informative?

4. **Lee bounds: The monotonicity assumption.** Lee (2009) shows that under a **monotonicity assumption**—treatment affects selection in the same direction for everyone—we can obtain tighter bounds.

   Monotonicity: $S_i(1) \geq S_i(0)$ for all i (treatment weakly increases employment for everyone).

   (a) Under this assumption, there are no "defiers" (EN types). Explain intuitively why this helps obtain tighter bounds.

   (b) Do you think monotonicity is plausible in the NSW context? Why / why not?

(c) Under monotonicity, what is the "excess" employment rate induced by treatment? Define:

$$p_0 = \frac{P(S = 1|D = 1) - P(S = 1|D = 0)}{P(S = 1|D = 1)}$$

This is the fraction of treated employed who are "compliers" (employed only because of treatment). Compute p_0 from the data.

5. **Deriving the Lee bounds.** The key insight: to compare "apples to apples," we need to trim the treated distribution to remove compliers. Since we cannot identify individual compliers, we trim from the tails.

   Let $F_{Y|D=1,S=1}$ be the CDF of wages among treated employed. The Lee bounds are:

   $$\begin{aligned} \text{Lower bound:} \quad & \mathbb{E}[Y|D = 1, S = 1, Y \leq y_{1-p_0}] - \mathbb{E}[Y|D = 0, S = 1] \\ \text{Upper bound:} \quad & \mathbb{E}[Y|D = 1, S = 1, Y \geq y_{p_0}] - \mathbb{E}[Y|D = 0, S = 1] \end{aligned}$$

   where y_q is the q-th quantile of wages among treated employed.

   (a) Explain the intuition: why does trimming the top p_0 fraction of treated wages give a lower bound? And trimming the bottom p_0 fraction give an upper bound?

   (b) *Derivation.* Show that under monotonicity:

   $$\mathbb{E}[Y^*(1)|\text{EE}] = \mathbb{E}[Y|D = 1, S = 1, Y \leq y_{1-p_0}] \quad \text{or} \quad \mathbb{E}[Y|D = 1, S = 1, Y \geq y_{p_0}]$$

   depending on whether compliers have the highest or lowest wages among treated employed. Since we don't know which, we use both to form bounds.

   Hint: Start from the fact that $\{D = 1, S = 1\}$ contains EE and NE types. The fraction of NE is p_0. If NE have the highest wages, then trimming the top p_0 gives EE. If NE have the lowest wages, trimming the bottom p_0 gives EE.

6. **Computing Lee bounds.** Using the data:

   (a) Find the quantiles y_{p_0} and y_{1-p_0} of wages among treated employed.

   (b) Compute the Lee bounds for the treatment effect on wages.

   (c) Compare to the Manski bounds. How much tighter are the Lee bounds?

   (d) Do the Lee bounds exclude zero? What can you conclude about the wage effect for always-employed workers?

7. **Visualization.** Create two figures:

   (a) **Histogram with trimming:** Plot the wage distribution among treated employed. Add vertical lines at y_{p_0} and y_{1-p_0} (the trimming thresholds), shade the regions that are trimmed for lower vs. upper bounds, and add a vertical line for the mean wage among control employed.

   (b) **Bounds comparison:** Create a figure comparing the naive estimate, Manski bounds, and Lee bounds side-by-side. Show the point estimate and the width of each interval. This should dramatically illustrate how much the monotonicity assumption tightens the bounds.

8. **Log wages.** Repeat the Lee bounds analysis using log(`wage`) instead of levels. (Note: since we condition on employment, log wages are well-defined.)

   (a) Compute Lee bounds for the effect on log wages.

   (b) How do these compare to Lee (2009) Table 5, which reports bounds of $[-0.083, 0.569]$ for NSW log wages?

   Note: Lee (2009) uses a slightly different NSW sample and includes covariates. Your results may differ but should be in a similar range.