

Evaluating the Econometric Evaluations of Training Programs with Experimental Data

LaLonde (1986)

Gabor Kezdi

CEU

Program Evaluation, Fall 2009

The Program: NSW

National Supported Work Demonstration

- U.S.A. 1970's
 - ▶ relatively small scale program (demonstration purpose)
 - ▶ a non-governmental program run by a not-for-profit institution
- Goal
 - ▶ helping long-term unemployed find long-term regular private jobs
 - ▶ AFDC recipients (women), marginalized people (men)
- Subsidized employment at private-sector employers
 - ▶ all labor costs covered by program for 12 months
 - ▶ starting wages below market level, relatively steep profile conditional on performance
- Complemented by weekly consulting (in groups)
- **Randomized assignment** to treatment and control group

Experimental Results

Average treatment effect on earnings

AFDC sample

Year	Treatments	Controls
1975	\$895 (81)	\$877 (90)
1976	\$1,794 (99)	\$646 (63)
1977	\$6,143 (140)	\$1,518 (112)
1978	\$4,526 (270)	\$2,885 (244)
1979	\$4,670 (226)	\$3,819 (208)
Number of Observations	600	585

$$\hat{ATE} = \$850 \text{ per year}$$

Male sample

Year	Treatments	Controls
1975	\$3,066 (283)	\$3,027 (252)
1976	\$4,035 (215)	\$2,121 (163)
1977	\$6,335 (376)	\$3,403 (228)
1978	\$5,976 (402)	\$5,090 (227)
Number of Observations	297	425

$$\hat{ATE} = \$880 \text{ per year}$$

LaLonde's methodological exercise

- Let's pretend that there is no experimental control group
 - ▶ and compare the treatment group to some other group
- Search for control group using publicly available large surveys
- Estimate the average treatment effect
 - ▶ various methods, various sets of control variables
- Compare each nonexperimental estimate to the experimental average treatment effect

- $2 \times 4 = 8$ control groups for women, 6 for men
 - ▶ 2 nationally representative sample
 - ▶ 4 sample selection rules (3 rules for men)
 - ▶ common among all:
 - ★ typically not from the same local labor market
 - ★ not all program criteria are observable

Results, AFDC sample

- The 2×2 control groups that one would think are best
 - ▶ (1.2) AFDC recipients (survey 1)
 - ▶ (2.2) AFDC recipients (survey 2)
 - ▶ (1.3) AFDC recipients not working (survey 1)
 - ▶ (1.3) AFDC recipients not working (survey 2)
- 7 regression models

1. comparing means

$$Y_{it} = \alpha + \delta D_i + u_i$$

	(1.2)	(2.2)	(1.3)	(2.3)
$\hat{\delta} =$	+1100	-700	+3000	+900

2. cross-sectional regression with many controls

$$Y_{it} = \alpha + \delta D_i + \beta' x_i + u_i$$

	(1.2)	(2.2)	(1.3)	(2.3)
$\hat{\delta} =$	+870	-240	+2900	+760

Results, AFDC sample, cont.

3. simple diff-in-diffs

$$Y_{it} - Y_{it-1} = \alpha + \delta D_i + u_i$$

$$\hat{\delta} = \begin{array}{cccc} & (1.2) & (2.2) & (1.3) & (2.3) \\ \hat{\delta} = & +2600 & +2200 & +3100 & +2100 \end{array}$$

4. diff-in-diffs controlling for age

$$Y_{it} - Y_{it-1} = \alpha + \delta D_i + \beta a_i + u_i$$

$$\hat{\delta} = \begin{array}{cccc} & (1.2) & (2.2) & (1.3) & (2.3) \\ \hat{\delta} = & +2300 & +2100 & +3000 & +1800 \end{array}$$

Results, AFDC sample, cont.

5. quasi diff-in-diffs with no controls

$$Y_{it} - \rho Y_{it-1} = \alpha + \delta D_i + u_i$$

$$\hat{\delta} = \begin{matrix} & (1.2) & (2.2) & (1.3) & (2.3) \\ +1700 & -450 & +1200 & +1200 \end{matrix}$$

6. quasi diff-in-diffs with few controls

$$Y_{it} - \rho Y_{it-1} = \alpha + \delta D_i + \beta' x_i + u_i$$

$$\hat{\delta} = \begin{matrix} & (1.2) & (2.2) & (1.3) & (2.3) \\ +1500 & +620 & +2900 & +950 \end{matrix}$$

7. quasi diff-in-diffs with many controls

$$Y_{it} - \rho Y_{it-1} = \alpha + \delta D_i + \beta' x_i + \gamma' z_i + u_i$$

$$\hat{\delta} = \begin{matrix} & (1.2) & (2.2) & (1.3) & (2.3) \\ +1800 & +670 & +2900 & -100 \end{matrix}$$

Results, male sample

- The 2×2 control groups that one would think are best
 - ▶ (1.2) men younger 55 not working in 1976 (survey 1)
 - ▶ (2.2) men younger 55 not working in 1976 (survey 2)
 - ▶ (1.3) men younger 55 not working in 1975 and 1976 (survey 1)
 - ▶ (2.3) men younger 55 not working in 1975 and 1976 (survey 2)
- Same 7 regression models

1. comparing means

$$Y_{it} = \alpha + \delta D_i + u_i$$

	(1.2)	(2.2)	(1.3)	(2.3)
$\hat{\delta} =$	-4020	-4100	+700	-1300

2. cross-sectional regression with many controls

$$Y_{it} = \alpha + \delta D_i + \beta' x_i + u_i$$

	(1.2)	(2.2)	(1.3)	(2.3)
$\hat{\delta} =$	-3480	-1670	-510	+220

Results, male sample, cont.

3. simple diff-in-diffs

$$Y_{it} - Y_{it-1} = \alpha + \delta D_i + u_i$$

$$\hat{\delta} = \begin{array}{cccc} & (1.2) & (2.2) & (1.3) & (2.3) \\ \hat{\delta} = & +480 & +2200 & +230 & -1640 \end{array}$$

4. diff-in-diffs controlling for age

$$Y_{it} - Y_{it-1} = \alpha + \delta D_i + \beta a_i + u_i$$

$$\hat{\delta} = \begin{array}{cccc} & (1.2) & (2.2) & (1.3) & (2.3) \\ \hat{\delta} = & -650 & -490 & -1320 & -1390 \end{array}$$

Results, AFDC sample, cont.

5. quasi diff-in-diffs with no controls

$$Y_{it} - \rho Y_{it-1} = \alpha + \delta D_i + u_i$$

$$\hat{\delta} = \begin{array}{cccc} (1.2) & (2.2) & (1.3) & (2.3) \\ -1360 & -1850 & +630 & -1400 \end{array}$$

6. quasi diff-in-diffs with few controls

$$Y_{it} - \rho Y_{it-1} = \alpha + \delta D_i + \beta' x_i + u_i$$

$$\hat{\delta} = \begin{array}{cccc} (1.2) & (2.2) & (1.3) & (2.3) \\ -1690 & -780 & -550 & +20 \end{array}$$

7. quasi diff-in-diffs with many controls

$$Y_{it} - \rho Y_{it-1} = \alpha + \delta D_i + \beta' x_i + \gamma' z_i + u_i$$

$$\hat{\delta} = \begin{array}{cccc} (1.2) & (2.2) & (1.3) & (2.3) \\ -1230 & -320 & +400 & +1470 \end{array}$$

Summary of results

- Non-experimental estimates of ATE vary a lot
 - ▶ by survey
 - ▶ by estimation model
 - ▶ the ex ante better models do not provide better estimates (especially true for diff-in-diffs specifications)
- Results for men are even worse
 - ▶ there virtually no pattern in which results are closer to target
 - ▶ estimates are a lot smaller on average (and more often negative)
- At the same time, using experimental control groups always result in the same ATE estimates regardless of estimation method
 - ▶ not surprisingly if treatment is truly exogenous

Conclusions

- Conclusions drawn originally
 - ▶ these non-experimental methods are not very useful
 - ▶ they are subject to manipulation to a large extent
- Qualification of those conclusions
 - ▶ true given the data at hand
 - ★ control groups from different labor markets
 - ★ program criteria not all observed
 - ★ in particular, pre-program labor market history is not observed to necessary extent
 - ▶ true given selection of control groups
 - ★ may be little overlap of some key right-hand side variables (see later at matching models)