History and Context
Within-Study Comparisons in Economics

Jeffrey Smith
Professor of Economics and Public Policy
University of Michigan
econjeff@umich.edu

OPRE Methodological Advancement Meeting
Washington, DC
September 6-7, 2012

Introduction

Thanks!

Who is that man?

Main point: what can and should be learned from within-study comparisons in economics

Outline / roadmap

Defining terms

LaLonde (1986)

Dehejia and Wahba (1999, 2002)

Defining concepts

Smith and Todd (2005)

Discussion

Defining terms

Experimental: involves random assignment to treated and untreated states

Non-experimental: everything else

Quasi-experimental: non-experimental

LaLonde (1986)

Compare the experimental estimates from the National Supported Work Demonstration evaluation with non-experimental estimates using comparison groups drawn from other data sets

Cook (and others) call this a “within study comparison”

LaLonde (1986) motivation

To quote LaLonde:

"The goal is to assess the likely ability of several *econometric methods* to accurately assess the economic benefits of employment and training programs" (604 – italics mine)

Put differently, the goal is to find an estimator that solves the selection problem

LaLonde (1986) basic setup

Comparison group source one: Panel Study of Income Dynamics female heads (continuously) from 1975-1979

Comparison group source two: Current Population Survey persons in the March 1976 CPS in the labor force in 1976 with individual income < 20K and household income < 30K

Using representative samples for comparison groups was standard practice at the time – it is less so now

LaLonde (1986) NSW experiment

The National Supported Work Demonstration examined the impacts of an expensive treatment on four groups with labor market difficulties: long-term AFDC recipients, high school dropouts, ex-convicts and ex-addicts.

LaLonde looks at two groups: AFDC women and the men from the other three groups

Aside: why would you combine these groups?

Treatment group observations were randomly assigned from January 1976 to April 1977

Random assignment took place in 10 sites around the country, all of them in cities.

LaLonde (1986) estimators

Linear regression

First differences regression

Regression with lagged dependent variable

Bivariate normal selection model with exclusion restrictions, estimated using the Heckman (1979) two-step approach

LaLonde (1986) variables

Covariates: age, Black and Hispanic indicators, years of schooling, an indicator for married, and a high school completion indicator (and that is all!)

Outcome variable: Real earnings from NSW survey (treatment group), SSA earnings records (CPS comparison group), PSID survey (PSID comparison group)

Dependent variable: real earnings in 1978

Lagged dependent variable: real earnings in 1975

Exclusion restriction variables: urban residence (!), employment status in 1976 (!), AFDC status in 1975 (!), number of children (!)

LaLonde (1986) results

The non-experimental impact estimates vary widely across estimators

The non-experimental impact estimates vary widely across comparison groups

Limited specification tests combined with a priori reasoning do not rule out all of the poorly-performing estimators

Bivariate normal model results are wrong for the reasons already described plus problems with choice-based sampling.

LaLonde (1986) conclusions

"… policymakers should be aware that the available non-experimental evaluations of employment and training programs may contain large and unknown biases resulting from specification errors." (617)

This paper was widely interpreted to mean that only experiments could provide credible estimates of the impact of active labor market policies.

It directly resulted in the choice of an experimental design for the National Job Training Partnership Act Study

LaLonde (1986) alternative reading

The data are much (all?) of the problem, not the methods.

Why would you expect the handful of covariates here to solve the selection problem?

No measures of past AFDC participation or number of children for the women

No measures related to crime for the ex-convicts or to substance use for the ex-addicts

Lagged annual earnings not well aligned with the time of treatment and measured and aligned differently for the treated and untreated units.

And we blame the methods?

Dehejia and Wahba (1999, 2002)

Apply matching and weighting methods to LaLonde's data on men (the data on women having been lost due to a misbehaving magnetic field)

Use a sub-sample of the experimental data to allow greater conditioning on "pre" period earnings

Find that matching and weighting yield estimates quite close to those of the experiment in their sample (albeit with quite large standard errors)

Dehejia and Wahba (1999, 2002) conclusions

“The methods we suggest are not relevant in all situations. There may be important unobservable covariates, for which the propensity score method cannot account. However, rather than giving up, or relying on assumptions about the unobserved variables, there is substantial reward in exploring first the information contained in the variables that *are* observed.”

This is all quite correct.

But: unobservable or unobserved?

The literature read this paper to mean that “matching works” even in weak data contexts.

Concerns about DW (1999, 2002) from HIST (1998)

Heckman, Ichimura, Smith and Todd (1998) draw some conclusions based on comparing the experimental estimates from the U.S. National Job Training Partnership Act Study to matching estimates obtained using “ideal comparison group data” from four of the experimental sites

Their key conclusions:

Conditioning variables matter a lot, particularly “pre” period outcomes

Putting comparison group members in the same local labor markets matters a lot

Measuring the dependent variable in the same way for the treatment group and comparison group members matters a lot

The LaLonde (1986) data fails to meet all these criteria!

Why then did Dehejia and Wahba (1999, 2002) get such good results?

Defining concepts: what are Dehejia and Wahba (1999, 2002) doing?

Study type 1: examining identification strategies

Example: examining whether unconfoundedness holds for a particular set of conditioning variables in a particular context using experimental estimates as benchmarks

Study type 2: examining the performance of alternative estimators that rely on the same identification strategy

Example: comparing three sets of estimates that all assume unconfoundedness generated by a parametric linear regression, nearest neighbor matching and inverse propensity weighting

Dehejia and Wahba (1999, 2002) combine these without being clear about it

Smith and Todd (2005)

Replicate Dehejia and Wahba (1999, 2002)

If you do what they did, you get what they got – not a minor feat!

If you use an equally (or more) plausible subset of the experimental data, the low bias results disappear, as they do with the full LaLonde (1986) sample

The estimates are quite sensitive to details of the specification

The estimates are quite sensitive to details of the estimation (even how ties are handled!)

Conclusions (narrow)

LaLonde (1986) does not show that non-experimental estimators do not work.

Dehejia and Wahba (1999, 2002) do not show that propensity score matching "works"

Replication is useful and can be carried out peacefully

Conclusions (broad)

There is no magic bullet – no estimator that always solves the selection problem

The question is not “which estimator works”

Instead, we want to know the mapping from data, parameter of interest and institutional context to estimator choice

Sometimes there is no non-experimental estimator that solves the selection problem for a given data set and institutional context – a bummer indeed!

Clever econometrics will usually lose out to bad data

References mentioned in the talk

Dehejia, Rajeev and Sadek Wahba. 1999. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association*. 94(448): 1053-1062.

Dehejia, Rajeev and Sadek Wahba. 2002. “Propensity Score Matching Methods for Non-Experimental Causal Studies.” *Review of Economics and Statistics*. 84(1): 151-161.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd. 1998. “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66(5): 1017-1098.

LaLonde, Robert. 1986. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *American Economic Review* 76(4): 604-620.

Smith, Jeffrey and Petra Todd. 2005a. “Does Matching Overcome LaLonde’s Critique of Nonexperimental Methods?” *Journal of Econometrics* 125(1-2): 305-353.

Smith, Jeffrey and Petra Todd. 2005b. “Rejoinder.” *Journal of Econometrics* 125(1-2): 365-375.

Further reading of potential interest

Agodini, Roberto and Mark Dynarski. 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *Review of Economics and Statistics* 86(1): 180-194.

Bell, Stephen, Larry Orr, John Blomquist, and Glen Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Cook, Thomas, William Shadish, and Vivian Wong. 2008. "Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27(4): 724-750.

Diaz, Juan Jose and Sudhanshu Handa. 2006. "An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program." *Journal of Human Resources* 41(2): 319-345.

Djebbari, Habiba and Jeffrey Smith. 2008. "Heterogeneous Program Impacts: Experimental Evidence from the PROGRESA Program." *Journal of Econometrics* 145(1-2): 64-80.

Fraker, Thomas and Rebecca Maynard. 1987. “The Adequacy of Comparison Group Designs for Evaluation of Employment-Related Programs.” *Journal of Human Resources*. 22(2): 194-227.

Friedlander, Daniel and Philip Robins. 1995. “Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods.” *American Economic Review*. 85(4): 923-937.

Heckman, James and V. Joseph Hotz. 1989. “Choosing Among Alternative Methods of Estimating the Impact of Social Programs: The Case of Manpower Training.” *Journal of the American Statistical Association*. 84(408): 862-874.

Lise, Jeremy, Shannon Seitz and Jeffrey Smith. 2004. “Equilibrium Policy Experiments and the Evaluation of Social Programs.” NBER Working Paper No. 10283.

Michalopoulos, Charles, Howard Bloom, and Carolyn Hill. 2004. “Can Propensity Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?” Review of Economics and Statistics 86(1): 156-179.

Todd, Petra and Kenneth Wolpin. 2006. “Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility.” *American Economic Review* 96(5): 1384-1417.

Shadish, William, M. H. Clark, and Peter Steiner. 2008. “Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments.” Journal of the American Statistical Association 103(484): 1334-1356.

Wilde, Elizabeth and Robinson Hollister. 2007. “How Close is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment.” *Journal of Policy Analysis and Management* 26: 455-477.