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)

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


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?

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.

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?

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)


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


Further reading of potential interest


