Economics Letters 151 (2017) 82–90

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

A cautionary tale on using panel data estimators to measure program impacts

Casey J. Wichman [a,*], Paul J. Ferraro [b]

[a] Resources for the Future, 1616 P St. NW, Washington, DC 20036, United States

[b] Bloomberg School of Public Health, Carey Business School, Whiting School of Engineering, Johns Hopkins University, 100 International St., Baltimore, MD 21202, United States

HIGHLIGHTS

- We compare experimental and observational estimates of environmental program impact.
- We expand the sample of comparison units to improve covariate balance.
- Despite similarity of covariates and baseline trends, bias of the estimator worsens.
- Fixed-effects panel estimators and indirect tests of their validity are no panacea.

ARTICLE INFO

Article history:
Received 31 May 2016
Received in revised form
19 November 2016
Accepted 22 November 2016
Available online 13 December 2016

JEL classification:
C52
C93
D12
H42
Q25

Keywords:
Design replication
Program evaluation
Matching
Panel data
Water conservation

ABSTRACT

We compare experimental and nonexperimental estimates from a social and informational messaging experiment. Our results show that applying a fixed effects estimator in conjunction with matching to pre-process nonexperimental comparison groups cannot replicate an experimental benchmark, despite parallel pre-intervention trends and good covariate balance. The results are a stark reminder about the role of untestable assumptions – in our case, conditional bias stability – in drawing causal inferences from observational data, and the dangers of relying on single studies to justify program scaling-up or canceling.

1. Introduction

Researchers using observational data often confront the question: what is the ideal experiment to identify my causal relationship? Less common is the question: how accurate is the estimate of my observational design relative to an experimental benchmark? To consider this question, researchers use "design replications", or "within-study designs", in which causal estimates from randomized experiments are compared to estimates from nonexperimental replications (Cook et al., 2008). In theory, nonexperimental designs can perform as well as experimental designs. These design replications allow researchers to examine the validity of the assumptions used to identify causal effects in specific nonexperimental contexts. How best to interpret the results of design replications has, however, been contentious (Lalonde, 1986; Heckman et al., 1997; Smith and Todd, 2005; Dehejia, 2005).

One source of contention is the failure of design replication studies to consider the sensitivity of their results to the choice of sample (Smith and Todd, 2005).[1] In a design replication study

* Corresponding author.
E-mail addresses: wichman@rff.org (C.J. Wichman), pferraro@jhu.edu (P.J. Ferraro).

[1] Or, as a referee pointed out, one might interpret Smith and Todd's (2005) analysis as changing the *population*, rather than the sample. We explore this

http://dx.doi.org/10.1016/j.econlet.2016.11.029
0165-1765/

using a fixed effects panel data (FEPD) estimator in conjunction with matching to pre-process the comparison group data, Ferraro and Miranda (forthcoming) show that an observational design using comparison households from a neighboring county can replicate results from an experimental design.[2] Through a bootstrapping exercise, they further demonstrate that the treatment effect estimates are not sensitive to the choice of sample within the two counties.

An alternative way to assess sensitivity to sample choice is to expand the pool of untreated units. Conventional wisdom suggests that increasing the number of comparison units from which to select a comparison group should (weakly) improve nonexperimental designs (Heckman et al., 1997). We assess this wisdom by extending the design of Ferraro and Miranda with the addition of a second group of untreated households, which are observationally more similar to the treated households. Including additional comparison households greatly improves covariate balance and yields parallel pre-treatment trends in outcomes. Despite these improvements, however, we find that the FEPD estimator, with or without pre-processing the data, performs worse: it no longer replicates the experimental benchmark.

2. An experimental benchmark and nonexperimental comparison groups

Our experimental benchmark comes from a randomized controlled trial (RCT) with over 100,000 households in Cobb County, Georgia (Ferraro and Price, 2013). In the RCT, a water utility sent messages to households to induce voluntary reductions in water use. Each treatment group comprised approximately 11,700 households and the control group, 71,600 households. Treatment assignment was randomized at the household level within nearly 400 meter route strata (i.e., small neighborhoods).[3]

We examine two of Ferraro and Price's treatments: (i) a **technical information treatment**, which instructed households on strategies to reduce water use; and (ii) a **social comparison treatment**, which augmented the technical information with social norm-based encouragement and a social comparison in which own consumption was compared to median county consumption. In the original experiment, the social comparison treatment induced a large (approximately 5%) statistically significant reduction in water consumption while the technical information treatment displayed a small (approximately 0.5%) statistically insignificant effect.

To construct nonexperimental comparison groups, we use households from neighboring Fulton County (used by Ferraro and Miranda, forthcoming), and nearby Gwinnett County. Cobb, Fulton, and Gwinnett counties had similar water pricing policies and the same water sources, weather patterns, state and metro regulatory environments, and other regional confounding factors during the experiment. To our knowledge, there were no contemporaneous policy changes in the comparison counties. We believe these comparison groups thus meet the Heckman et al. and Cook et al. criteria for effective observational designs.

3. Empirical strategy

Our identification strategy uses repeated observations on households to control for unobserved and unchanging characteristics that are related to water consumption and exposure to the treatment (Angrist and Krueger, 1999). Our design relies on the common linear, additive FEPD estimator,

$$w_{it} = \alpha + \boldsymbol{A}_i'\gamma + \boldsymbol{X}_{it}'\beta + \delta Treat_{it} + \lambda_{\boldsymbol{t}} + \varepsilon_{it}, \tag{1}$$

where w_{it} is monthly water use for household i at time t; $\boldsymbol{A}_i$ is a vector of fixed (time-invariant) household characteristics; $\boldsymbol{X}_{it}$ is a vector of time-varying household characteristics; $Treat_{it}$ is a treatment indicator; and $\lambda_{\boldsymbol{t}}$ are time fixed effects. Under an assumption of conditional bias stability, Eq. (1) provides an unbiased estimator of the Average Treatment Effect, δ, which was also the estimand estimated by the RCT. Conditional bias stability asserts that conditional on $\boldsymbol{X}_{it}$, pre-program differences in outcomes between treatment and comparison groups are stable across post-program periods. Ferraro and Miranda make the case for the plausibility of this assumption in the study context.

3.1. Data and samples

We use household water consumption data from the Cobb County Water System, the Fulton County Water Service Division, and the Gwinnett Department of Water Resources. We have thirteen months of pre-treatment data (May 2006–May 2007) and four months of post-treatment data (June–September 2007). The county tax assessor databases provide home and property characteristics, and the 2000 US Census provides data on neighborhood characteristics at the block-group level.

Table 1 shows average water consumption in thousands of gallons during key watering seasons for Cobb households in the experiment, and for Fulton and Gwinnett households. We also consider covariates that are observable to policymakers and that theory or empirical studies suggest could be important confounders in a study on water conservation (e.g., Ferraro and Miranda, 2014; Wichman et al., 2016). Overall, Gwinnett households appear to be more similar to treatment households along water use and socioeconomic characteristics than do Fulton households.

4. Observational measuring sticks

Drawing causal inferences in any nonexperimental design requires making untestable assumptions (e.g., model dependence, unconfoundedness, and so on).[4] To overcome model dependence, researchers are increasingly using matching techniques to reweight the sample so that treatment and comparison groups are similar and, thus, rely less heavily on parametric assumptions (Ho et al., 2007). Furthermore, observing parallel trends in outcomes prior to treatment is commonly used to support the conditional bias stability assumption. As in Ferraro and Miranda, we focus on these two empirical heuristics in our analysis.

4.1. Does trimming and matching improve covariate balance?

Following Ferraro and Miranda (forthcoming), we first use the **full sample** of treated and comparison households. Second, we construct a **trimmed sample** using the optimal trimming rule of Crump et al. (2009) to remove observations with extreme propensity scores.[5] Third, we construct two **matched samples**. We use nearest-neighbor (1:1) Mahalanobis covariate matching

point empirically by examining two comparison groups separately (i.e., as distinct populations) as well as jointly (i.e., as different draws from the same population).

[2] Pre-processing in our context refers to matching or trimming to reweight the sample prior to applying a parametric estimator.

[3] For more details on the experiment and randomization, see Ferraro and Price (2013).

[4] Causal inference in experimental designs also relies on untestable assumptions (Heckman and Smith, 1995), but fewer than are required in nonexperimental designs.

[5] Based on a logit model, our optimal trimming rule discards observations with estimated propensity scores outside the interval [0.03, 0.97].