# We found causality in a hopeless place. Challenges of causality in demographic observational studies

**Bruno Arpino**
University of Florence (Italy)
bruno.arpino@unifi.it

**Abstract**
Demographers are often confronted with the goal of establishing a causal link between demographic events (e.g., fertility, union formation and dissolution) and socio-economic, health and other types of measures. Since experiments are commonly not a feasible strategy, demographic studies usually rely on observational data. Not being able to manipulate the treatment assignment, demographers have to deal with several issues, such as omitted variable bias and reverse causality. The aims of this paper are to review the methods commonly used by demographers to estimate causal effects in observational studies and to discuss strengths and limitations of these methods and of their implementation in demographic studies.

## Long abstract

### How do demographers deal with causality in observational studies?

First, I will review the methods commonly used by demographers to estimate causal effects. I will consider quantitative observational studies in which the authors explicitly aimed at assessing causal effects and published in the demographic journals with the highest impact.

Results from a preliminary search on the journals *Demography* and *European Journal of Population* indicate that when demographers explicitly aim at estimating causal effects, the methods most commonly used are: propensity score matching (and similar approaches, like propensity score weighting or alternative matching methods), regression models with (individual, family or twin) fixed effects and instrumental variable models. Simultaneous equations and structural equations models are also often employed.

In the rest of the paper I will focus on comparing strengths and limitations of three of the most commonly used methods: propensity score matching, instrumental variables models and panel data models with individual fixed effects. For the first two I will consider both the case of cross-sectional and longitudinal data. To illustrate the discussion, I will use real data from the British Household Panel Study (BHPS). Then I will simulate data that mimic the BHPS data structure to better highlight what are the consequences of violations of assumptions underlying each method.

### Methods

In observational studies, direct comparison of outcomes across treatment groups can give rise to biased estimates because groups being compared may be different due to lack of randomization. Subjects with certain characteristics may have higher probabilities than others to be exposed to the treatment. If these characteristics are also related to the outcome under investigation, an unadjusted comparison of the groups is likely to produce wrong conclusions about the treatment effect.

Propensity scores, defined as the probability to receive the treatment conditional on the set of observed variables, were introduced by Rosenbaum and Rubin (1983) as a one-dimensional summary of the multidimensional set of covariates, such that when the propensity scores are balanced across the treatment and control groups, the distribution of all covariates are balanced across the two groups. In this way, the problem of adjusting for a multivariate set of observed characteristics reduces to adjusting for the one-dimensional propensity score.

Propensity scores can be used in several different ways (such as matching, weighting, stratification or regression). Propensity score matching (PSM) is the most commonly used. PSM consists in matching treated and control individuals with similar values of the propensity score. Rubin (2001) argues that an advantage of propensity score methods is that they allow observational studies to be designed similar to randomized experiments: the design of the study is separated from the analysis of the effect of the treatment on the outcome. Crucially important for the successful design of observational studies based on estimated propensity scores is the assessment of the balance achieved in the distribution of covariates between treated and control individuals after matching. Such diagnostics enable applied researchers to determine whether conditioning on the estimated propensity score has removed observed systematic differences between treated and control groups.

Importantly, propensity score methods can only ensure balance of background variables used in its estimation, and consequently, causal inferences based on these

methods carry an assumption that no unobserved confounders exist. In the presence of unobserved confounders PSM gives biased estimates.

Panel data models with individual fixed effects (FE) is a common technique used to estimate causal effects when unobserved confounders may exist. FE models exploit the panel structure of the data to remove unobserved time invariant variables (Wooldridge, 2010). FE models cannot adjust for time variant confounders. Moreover, they give biased estimates in the presence of reverse causality, i.e. when the it is the outcome to have a causal effect on the treatment variable and not the other way around (or both directions of the causality are possible).

Compared to FE, PSM has pros and cons. PSM is in general more robust than parametric regression models, especially when the distributions of covariates in the two groups being compared are very different because the regression estimator depends heavily on extrapolation using the specific functional form of the model (Imbens 2014). In addition, PSM can highlight initial differences in observed covariates among treated and control units and offers the opportunity to check whether it was successful in reducing these imbalances. Moreover, implementing PSM on longitudinal data offers several advantages (Arpino and Aassve 2013). First, we can match on covariates measured before the treatment is measured. In this way, we avoid controlling for covariates that could be on the causal pathway between the treatment and the outcome (mediators), generating biased estimates (e.g., Imbens 2004; Rosenbaum 1984). Second, the lagged value of the outcome variable can be included in the set of matching covariates. Matching on the lagged outcome, similar to panel models with individual fixed-effects, allows controling for time-invariant unobserved confounders (Imai and Kim 2019; Arpino and Aassve 2013; Athey and Imbens 2006). Both PSM and FE cannot deal with time-variant confounders.

The standard solution to deal with selection on unobservables is to use an instrumental variable (IV) method (Wooldridge 2010), which relies on the availability of a variable (instrument) that satisfies two key conditions: it should be associated with the treatment (relevance) and is should not have direct impact on the outcome (validity). In several empirical applications in demography is very difficult to think about possible variables that can theoretically satisfy these two conditions. Moreover, even if such a variable is available the IV estimator can be unsatisfactory. The reason is that, unless we are willing to impose very strong assumptions, IV estimates refer only to the unobserved sub-sample of the population that reacts to the chosen instrument, i.e. the so called *compliers* (Imbens and Angrist 1994; Angrist et al. 1996). The corresponding parameter estimate is, consequently, the local average treatment effect (LATE) which, in the presence of heterogeneous treatment effects, may be different from average treatment effect (ATE) and the average treatment effect for the treated (ATT) that usually are the parameters of interest.

An advantage of IV over FE is that it can be used also with cross-sectional data and can deal not only with time-variant confounders but also with the problems of reverse causality and measurement error. IV can also be combined with FE when panel data are available.

***The implementation of PSM, FE and IV in demographic studies***

The goal of this section is also to highlight common misuses of PSM, FE and IV or their erroneous interpretation in demographic studies. For example, it will be assessed whether reviewed studies report the balance of covariates that is achieved when using PSM, that is one critical aspect of the method. As stated by Austin (2008): "just as no RCT should be published that does not compare baseline characteristics between the

arms of the trial, so every study using propensity-score matching should compare measured baseline characteristics between treated and untreated subjects in the matched sample." As another example, analyses based on instrumental variables should carefully discuss their theoretical foundations and should empirically assess the plausibility of their assumptions.

**Data: The British Household Panel Study**

Real data that I will use come from the British Household Panel Study (BHPS). The BHPS, is an annual panel survey with a nationally representative sample of about 5,500 households recruited in 1991, containing approximately 10,000 interviewed individuals. Participants are re-interviewed each successive year for 18 years; participants who split from original households to form new ones are followed, and all adult members of these households are also interviewed. Similarly, new members joining sample households become eligible for interview, and children are interviewed beginning at age 16. The BHPS data set provides information on several socioeconomic characteristics, family orientations, fertility and partnership histories, health and subjective wellbeing among others variables that have been widely used in demographic analyses.

**Simulation experiments**

Following Arpino and Cannas (2016) I will design simulation experiments to mimic the observed data in several respects. First, we will keep the same data structure observed in BHPS, that is, the same number of individuals and waves. In this way, in our simulations, we will consider a realistic case of unbalanced panel data. Second, instead of generating values of covariates as realizations of random variables as typically carried out in simulation studies, we will use the same distribution of covariates as observed in the dataset. Finally, the coefficients of covariates in the true models generating the treatment and the outcome will be set to values similar to observed coefficients estimated on the real data. To gain further understanding on the performance of the different methods, we will modify the baseline simulation set-up in 4 ways:

1. introducing one or more unobserved time-invariant or time-variant confounders;
2. varying the amount of within and between individual variation;
3. allowing for reverse causality from the outcome to the treatment;
4. generating instrumental variables with different characteristics in terms of their validity and relevance.

**Conclusion**

I will conclude with a summary of guidelines for good practices to follow in empirical demographic studies and a discussion of alternative methods that received little attention from demographers. I will discuss general limitations of causal arguments based on quantitative empirical observational studies. I will also stress that sound causal inference cannot depend only on the specific method used but it is crucially determined by the quality of theoretical arguments and data available.

## References


Arpino, B., and Aassve, A. (2013). Estimating the causal effect of fertility on economic wellbeing: Data requirements, identifying assumptions and estimation methods. *Empirical Economics*, 44, 355–385.

Athey, S., and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74, 431–497.

Arpino B., and Cannas, M. (2016) Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine*, 35(12), 2074–2091.

Austin PC. (2008) A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27(12): 2037–2049.

Balbo N., and Arpino B. (2016) The role of family orientations in shaping the effect of fertility on subjective well-being: a propensity score matching analysis. *Demography*, 53(4), 955–978.

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data?. American Journal of Political Science, 63(2), 467-490.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W. (2014). Matching methods in practice: Three examples (NBER Working Paper No. 19959). Cambridge, MA: National Bureau of Economic Research.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A*, 147, 656–666.

Rosenbaum PR, Rubin DB. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*,70: 41–55.

Rubin DB. (2001) Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, 2:169–188.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.