How do we know if a program made a difference? A helicopter tour of the econometrics of non-randomly allocated community-level programs

Gustavo Angeles
Assistant Professor, Department of Maternal and Child Health, University of North Carolina at Chapel Hill.
Address:
Carolina Population Center
University of North Carolina at Chapel Hill
CB-8120 University Square East
Chapel Hill, NC 27516-3997 USA
Phone: 1-919-843-9471
Fax: 1-919-966-2391
Email: gangeles@email.unc.edu

Peter M. Lance
Research Associate, MEASURE Evaluation Project, Carolina Population Center, University of North Carolina at Chapel Hill.
Address: Carolina Population Center
University of North Carolina at Chapel Hill
CB-8120 University Square East
Chapel Hill, NC 27516-3997 USA
Phone: 1-919-843-3753
Fax: 1-919-966-2391
Email: pmlance@email.unc.edu

Extended Abstract

Researchers throughout the social sciences frequently engage in a familiar pursuit: estimation of the impact of a program. Their aim is to learn about how the introduction or intensification of a program alters outcomes from what would otherwise prevail. They often have available to them data on the behavior or human resource outcomes of program participants and non-participants (or those exposed to varying program intensity levels) drawn from the messy laboratory of the human experience. The challenge with data of this nature is that the program was likely not randomly assigned across the population of interest: it becomes difficult to say that those not exposed to it form a legitimate control group for those who were exposed in the sense that true randomization of the program across communities might provide. Recognizing this problem, in recent decades various disciplines have offered a range of perspectives and practical tools or strategies for dealing with it. In this paper we offer a survey of methods from the tradition of econometrics, with a focus on instances where our data is multi-level in nature (with, for instance, variables defined at the individual and community levels) and the failure of experimental variation is driven by community-level characteristics. These techniques are not necessarily exclusive to applied microeconometrics, but it is there that they have generally proven most popular and are the most well known. Certainly, however, the basic conceptual concerns motivating their use in applied economics emerge again and again in other disciplines as well. We pay particular attention to evaluation of programs operating at the community level (so that there is no variation in exposure within the community) though in many instances our tools generalize to other cases.[1] Our goal is to provide an approachable but nuanced survey that will offer non-economists a helpful and practical guide to these valuable tools.

We begin by characterizing the problem. There are many scenarios that readily come to mind. For instance, evaluation of the impact of a community-level family planning program (for instance, placement of a family planning clinic in communities) on fertility by simply comparing fertility in communities exposed to and not exposed to the program might yield misleading results if the program was non-randomly assigned across them.[2] If, for instance, the program was applied to communities that had higher baseline fertility to begin with, the available data might even suggest that the program *caused* higher fertility. (This is the familiar endogenous program placement problem discussed by Rosenzweig and Wolpin (1986), Angeles et al. (1998) and others.) The estimation problem in this case stems from the fact that the program was not assigned randomly, and

---

[1] Consider for instance the case where individuals within communities can elect to participate in the program (and hence there will be variation in the program or initiative within communities) but their decision is nonetheless influenced by community-level factors. Though the estimation techniques that we discuss can often yield valid estimates in these cases, we sidestep explicit discussion of selection into programs by *individuals* (as opposed to communities) because we would then inevitably be compelled to make a long and distracting digression to discuss formally individual-level selection models, which are somewhat tangential to the central thrust of the session.

[2] In the setting of explicit multilevel modeling, we might estimate a logistic regression of a woman's fertility status on various controls for her own characteristics (or those of her household), those of her community, and an indicator variable for the program operating in her community.

hence those communities not receiving it do not provide a valid control for those that did using estimators that rely on straightforward comparison of outcomes. Non-random assignment implies that the program is effectively correlated with community-level characteristics which influence the outcomes of interest. This renders invalid estimates based on simple comparison of outcomes. These characteristics can range from things that operate at the community-level and are fundamentally unobservable (for example cultural norms regarding fertility) to things that operate at the individual level and are conceptually observable but simply weren't for whatever reason included in the available data (for instance, individual income, which can vary on average between program and non-program communities). This basic framework for thinking about the problem can be extended in any number of directions. However, the basic nature of the problem remains more or less the same: we want to make inferences about the effect of variation in various programs across communities as if that variation had been experimental, while in most cases this is not a safe assumption. The key idea is that, if the program is correlated with unobserved community-level characteristics, estimates of the effect of the program on individual behavior using straightforward statistical methods (that essentially amount to comparison of outcomes across communities) might be contaminated by those other community-level factors. In essence, the estimate is picking up the effect of these other unobserved community-level variables rather than the pure program impact that we seek.[3]

We then shift emphasis toward discussion of the three most popular "classes" of estimators applied to the evaluation of community-level initiatives. These are simple cross-sectional estimators, fixed-effects estimators and instrumental variables estimators. The first type identifies program effects by examining the degree to which various outcomes (eg child health) seem to vary with the distribution of a program or initiative *across* communities. These estimators are thus often broadly referred to as "cross-sectional." In some sense they implicitly involve assuming random or experimental variation in programs.[4] We briefly describe instances where researchers have attempted to justify this approach, and focus on the consequences of using these models in the face of non-random program or initiative assignment. While the basic concepts apply to the very general case (in which the "program" is rather broadly defined as some community-level determinant of behavior and need not necessarily even be a tool of policymakers), we also place special emphasis on the problem of endogenous program placement (where the variable of interest is policy tractable and its' assignment across communities is systematically guided by various community characteristics). A program may be consciously targeted across communities or, alternatively, nonrandom assignment can be the artifact of inadvertent targeting due to the practical realities of program

[3] Another way of thinking about this is that the program indicator now serves statistically not only as a source of information about exposure to the program, but also as a proxy for the community characteristics with which it is associated.

[4] More precisely, in a setting where we observe some but not all community characteristics influencing the outcome of interest, these estimators assume that any variation in program placement beyond what can be accounted for with community-level factors that we do observe (and hence can control for in a multi-level regression) is random within the context of the regression model (in other words, this estimator allows for the possibility that remaining variation in programs across communities is systematic, as long as the variables associated that remaining variation are not among the unobserved community characteristics that also influence the outcome of interest).

implementation. We then consider within- or fixed-effects-estimators. This is a broad class of estimators, but the common thread running through them is the identification of program effects by considering variation *within* communities over time. Program effects are identified from the relationship between changes in outcomes of interest (within the community over time) and intertemporal variation in program placement or intensity. One of the members of this family that we will discuss is the difference-in-difference estimator. Finally, we consider instrumental variables approaches to addressing community level unobserved heterogeneity. We discuss the basic intuitive logic of the estimator as well as the classical conditions that an instrument must meet. We also emphasize the care that must be taken in interpreting instrumental variables estimates and outline briefly the sources of the skepticism of some researchers toward these models. We do not advocate one approach over the others, but instead focus on the idea that each involves tradeoffs and hence the appropriate estimator will depend on the particular circumstances confronting the researcher.

For each class of estimator, we provide coverage for several points of emphasis:

- The type of data needed: Can the model be estimated with data at a point in time? Do we need data from several points in time? If so, do we need to have observations drawn from the same communities or, ultimately, the same units of analysis (eg individuals) at different points in time?
- The assumptions behind the estimator: Does it require certain assumptions regarding the nature of the community-level characteristics? (For instance, are they assumed to be fixed or varying over time under the model in question?) What additional assumptions regarding the data are necessary to justify the estimation strategy? Our emphasis is not on rendering judgment regarding the legitimacy of a given assumption, but conveying the notion that these assumptions are really about behavior (in other words, making an assumption for the purposes of motivating one estimator or another is really equivalent to making an assumption about behavior), and as such their plausibility will vary from setting to setting,
- Sources of identification: All of the models we consider "identify" the effect of the program by considering the relationships between various variables. For instance, in the discussion above simple comparison of fertility rates implicitly identifies program effects by relating variation in program implementation across communities to variation in fertility across the same communities. Understanding the sources of identification sheds light on the type of variation one must observe in their data to use a given estimator.
- Interpretation of results: Some of the popular methods introduce subtle but important (and often not well understood) complications to the interpretation of the results. For instance, some program evaluation estimators actually yield, under certain circumstances, what are referred to as local average treatment effects (Imbens and Angrist (1994)). These provide the effect of the program only for some unidentified subpopulation. We will provide a very intuitive introduction to these issues, emphasizing interpretation relevant for evaluating program impact.

This basic approach allows us to provide a common frame of reference for the various types of estimator.

Our entire discussion is framed in terms of a single canonical example (estimation of the effect of a local health initiative on child health captured by height for age). It is our hope that, by anchoring the conceptual discussion with a single unified example, our audience will be able to understand better the nature of the different estimation strategies and recognize where their intrinsic differences lie. We conclude with a brief review of two estimation tools that have received particular attention in recent years, propensity score estimation and regression discontinuity design estimation. Though both have been applied to a wide variety of empirical circumstances, we also discuss them in the narrower context of community-level variables that complicate estimation of the impact of various initiatives. Our talk will be largely conceptual in nature, while the paper delves into the issues in more detail, both in the conceptual and analytical sense. The paper is intended to serve as a compact and accessible manual on the econometrics of program evaluation for researchers interested mainly in applying these tools in their own work evaluating the impact of various programs (as opposed to understanding the deeper theoretical intricacies of these models for the purpose of advancing the methodological literature). The paper also provides examples of the various estimators using the STATA statistical package.

Sources

Angeles, G., D. Guilkey, T. Mroz. (1998). ``Purposive Program Placement and the Estimation of Family Planning Program Effects in Tanzania." *Journal of the American Statistical Association* 93(443): 884-899.

Imbens, G. and J. Angrist. (1994). "Identification and Estimation of Local Average Treatment Effects" *Econometrica* 62: 467-476.

Rosenzweig, M., K. Wolpin. (1986). ``Evaluating the Effects of Optimally Distributed Programs." *American Economic Review* 76(3): 470-482.