OPTIMAL DESIGN FOR PROGRAM EVALUATIONS[1]

Stuart H. Kerachsky, Mathematica Policy Research
Charles D. Mallar, Mathematica Policy Research and John Hopkins University

A. INTRODUCTION

This paper presents a methodology for selecting a control sample that can be used in program evaluations when the program treatment group is both geographically clustered and preselected.[2] The criteria used to choose a program evaluation design are the classical ones of minimizing bias and maximizing efficiency in the estimation of treatment effects. These criteria are satisfied if (1) there are no systematic differences between the treatment and control groups, and (2) the other variables that explain the behavioral outcomes are observed and included in the statistical analysis, in order to obtain precise estimates of standard errors. Generally, random assignment of the target population to treatment and control groups at the point of entry into the program will minimize bias and maximize efficiency, ceteris paribus. If the assignment deviates from a random one according to known and measured characteristics, unbiased but less efficient estimates can be obtained through multivariate techniques, as long as the treatment and control groups are sufficiently similar so that their behavioral relationships are structurally identical. (See Goldberger, 1972a and b; Cain, 1975; Pitcher, forthcoming; and Conlisk, forthcoming.) The quasi-experimental design developed below for treatment samples that are geographically clustered and preselected approximates random assignments at the point of entry into the program and will generally have the same properties if successfully applied.

B. THE GENERAL PROBLEM AND RESOLUTION

Geographically clustered programs include both those that are size-specific and those that draw heavily from only certain areas of the country. It is assumed that the selection of program sites can be either arbitrary or controlled for in the evaluation design. Evaluations of geographically clustered programs with treatment samples that have been preselected are common in the social sciences. Examples can be found in the evaluations of:

(1) Employment and training programs

(2) Education programs and projects

(3) Variants of state unemployment insurance programs

(4) Different public assistance programs

(5) Local transportation programs.

Many of these evaluations have been of ongoing programs for which random assignments of participants to treatment and control groups at the time of enrollment are not feasible. The potential for political, ethical, budgetary, and operational problems when intervening in the selection process for an ongoing program often precludes random assignment as a viable approach.[3] Consequently, the program treatment group is often preselected.

Previous evaluations have often relied on comparisons between the behavior of the program treatment group and another sample composed of some combination of the following:

(1) People on waiting lists for an oversubscribed program

(2) Early dropouts from the program

(3) Friends or relatives of those in the program

(4) People who have opted not to enroll (including "no shows"), or who have been screened out of the program

(5) Preprogram observations of the treatment group

(6) General population samples, including (at least some) program participants.[4]

The findings from such comparison-group evaluation studies have, in turn, been disputed because the assumptions needed to show unbiasedness and efficiency are not plausible.[5] Two likely sources of bias are unobserved differences in the sample (e.g., in terms of motivation) and overlap between the treatment and comparison groups. Even if there are no unobserved differences and no overlap between the groups so that unbiased measures of the treatment effect can be obtained, observed differences have often reduced the efficiency of estimates of the treatment effects. Furthermore, very disparate samples also strain the credibility of the underlying assumption that the treatment and comparison samples have the same behavioral structures (i.e., that the same equation is applicable to both groups).

Because of the geographic clustering, however, another approach can be developed. A random sample of program participants, combined with a sequentially matched sample from nonprogram sites, can approximate a random assignment strategy and thereby avoid bias and maximize efficiency. This sequential matching involves two distinct steps. First, a random sample of sites similar to those of the program are chosen for the control sites. Second, within these control sites, an appropriate sampling frame is set up, and individuals are randomly selected from the sampling frame for the comparison group. Throughout the remaining discussion it is assumed that the treatment sample for this quasi-experimental design is a random sample of people in the program.[6]

For the control group, a sample of sites must be selected that are similar to, but outside, the areas in which the program is clustered.

Program sites are excluded to minimize biases that result both from self-selection into the program (e.g., unobserved differences in motivation) and from treatments affecting the behavior of persons not in the program (especially for saturation programs). Selection probabilities are then assigned to the remaining sites (the nonprogram sites) in proportion to their similarity to the program site.

Once the control sites are chosen, a selection process similar to the de facto program selection process is then set up within the control sites, to yield a sampling frame of persons with observed and unobserved characteristics similar to program participants. The comparison group is then randomly chosen from the sampling frame with selection probabilities for individuals that are proportional to their similarity to program participants.

The sequential process of obtaining an appropriate comparison sample can be summarized as follows:

(1) Eliminate program sites from which participants are principally recruited.

(2) Assign probabilities of selection to nonprogram sites in proportion to their similarity to program sites.

(3) Randomly select the control sites based on the probabilities as assigned in step (2).

(4) Within control sites, eliminate any program participants.

(5) Assign probability of selection to other persons in proportion to their similarity to program participants.

(6) Randomly select individuals for the comparison group based on the probabilities as assigned in step (5).

This quasi-experimental design will yield treatment and control groups for which the assumptions needed to obtain unbiased and efficient estimates of treatment effects are usually plausible. The two groups are unlikely to differ systematically in either observed or unobserved characteristics, and there is no overlap in the samples. Finally, any observed differences that remain between the treatment and comparison groups can be controlled for in a multivariate estimation framework.

In some instances, the quasi-experimental design developed here will be preferable to random assignments at the time of enrollment. For example, randomization across sites is desirable when the fraction of the population being served is so large that the behavior of a within-site control group could be affected. This would of course be true for saturation programs in which a large portion of the eligible population is enrolled in the program.

C. AN APPLICATION TO AN EVALUATION OF THE JOB CORPS

The methodology developed above had recently been applied in a design for an evaluation of the economic impact of the Job Corps program on its participants (see Kerachsky and Mallar, 1977, for more details). The Job Corps program provides education, training, and support services in a residential setting to youths who come from severely disadvantaged families (youths age 16 to 23). Random assignments of potential enrollees to a control group were not feasible because of operational and other considerations. Therefore, the sequential matching process outlined above was instituted to obtain an appropriate comparison group.

First, program sites--both zip-code regions saturated by Job Corps participation (i.e., high proportions of eligible youths in the program) and zip-code regions proximate to Job Corps centers--were eliminated. Then the remaining regions were assigned selection probabilities in proportion to their similarities to the home regions of Job Corps members, based primarily on the poverty and racial compositions of the regions. Once the control sites were chosen, youths living in the relevant areas were assigned selection probabilities in proportion to their similarity to Job Corps participants, based primarily on their poverty, age, race, and educational status.[7] A sample of youths was then chosen for interviewing. Finally, the baseline questionnaire was designed to measure any observed differences that remained and which are now important for explaining the economic outcomes that are being studied.

This quasi-experimental design seems appropriate for the Job Corps evaluation and should lead to precise estimates of the economic impacts of the program. The assumptions needed for unbiased and efficient estimates of the program treatment effects seem plausible. There is no overlap, and with a large number of observations, the program treatment group should differ from a comparison sample only in terms of access both to information about Job Corps and to Job Corps centers. Therefore, a feasible program evaluation has been designed even within the constraints of an ongoing program.

D. CONCLUSIONS

A widely applicable technique for evaluating ongoing programs has been developed. The strategy for obtaining the comparison group sample is feasible and should lead to unbiased and efficient estimates of program treatment effects. The assumptions needed for minimizing bias and maximizing efficiency are plausible. There should be no overlap between the treatment and comparison samples, unobserved differences between the samples should be minimized, and observed differences should be small enough to be controlled for with a multivariate estimation technique, with only a small loss in efficiency.

FOOTNOTES

1. This paper summarizes a quasi-experimental design that was first developed for and applied to an evaluation of the economic impact of the Job Corps program on its participants (see Kerachsky and Mallar, 1977).

2. See the next section for precise definitions of "geographically clustered" and "preselected."

3. These problems are, of course, less important for experimental and demonstration programs.

4. The closer the match between these general population samples and the program sample, the greater the overlap between the samples--hence, the greater the biases.

5. See Goldstein (1973) for summaries and criticisms of several of the studies of employment and training programs.

6. Random selection as discussed here can be with or without stratifications.

7. Females were oversampled in the comparison group relative to Job Corps participants to increase the efficiency of separate estimates for females.

REFERENCES

Cain, Glen G. "Regression and Selection Models to Improve Nonexperimental Comparisons." In Evaluation and Experiment, edited by Carl A. Bennett and Arthur A. Lumsdaine. New York: Academic Press, 1975.

Conlisk, John. "Choice of Sample Size in Evaluating Manpower Programs." In Research in Labor Economics, Supplement I: Evaluating Manpower Training Programs, edited by Farrell Bloch. Greenwich, Connecticut: JAI Press, forthcoming.

Goldberger, Arthur S. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Institute for Research on Poverty Discussion Paper No. 123-72, Madison: University of Wisconsin, 1972a.

Goldberger, Arthur S. "Selection Bias in Evaluating Treatment Effects: The Case of Interactions." Institute for Research on Poverty Discussion Paper No. 129-72, Madison: University of Wisconsin, 1972b.

Goldstein, Jon H. "The Effectiveness of Manpower Training Programs: A Review of Research on the Impact of the Poor." Paper No. 3 of Studies in Public Welfare, Subcommittee on Fiscal Policy, Joint Economic Committee of the Congress of the United States. Washington, D.C.: Government Printing Office, 1973.

Kerachsky, Stuart H. and Charles D. Mallar. "Design of an Evaluation of the Job Corps." MPR Working Paper No. C-12, Princeton, New Jersey: Mathematica Policy Research, Inc., 1977.

Pitcher, Hugh. "A Sensitivity Analysis to Determine Sample Size for Performing Impact Evaluations." In Research in Labor Economics, Supplement I: Evaluating Manpower Training Programs, edited by Farrell Bloch. Greenwich, Connecticut: JAI Press, forthcoming.