The Canadian Journal of Program Evaluation Vol. 9 No. 2 Pages 57–70
ISSN 0834-1516 Copyright © 1994 Canadian Evaluation Society

MEASURING PROGRAM EFFECTS IN THE PRESENCE OF SELECTION BIAS: THE EVOLUTION OF PRACTICE

Frank Eaton
Abt Associates of Canada
Ottawa, Ontario

Abstract: The author applies a variety of methods for estimating the effects of employment and training programs on labor market behavior. This article traces how these methods have evolved over the past 10 years, focusing on the problem of selection bias. It discusses the context of measuring effects related to labor market activity, and then presents the essential elements of selection bias modeling and the specific changes that led to methods currently in use. It also provides an assessment of the benefits and potential weaknesses of these techniques.

Résumé: L'auteur a utilisé plusieurs méthodes afin d'estimer les effets des programmes d'emploi et de formation sur l'expérience des participants dans le marché du travail. Cet article décrit l'évolution de ces méthodes au cours des dix années passées et traite plus particulièrement du problème du biais de sélection. Il aborde le contexte d'estimation des effets au marché du travail, puis il présente les éléments de l'analyse du biais de sélection et l'acheminement vers les méthodes utilisées actuellement. Il présente aussi une évaluation des avantages et des faiblesses potentielles de ces méthodes.

This article focuses on estimating the effect of a government program on the labor market experience of its participants, where the process of selecting the participants could bias the estimates. Some training and job creation programs strive to directly affect participants' wages, employability, and earnings; for others, such effects are secondary but still important. This article examines the author's experience in evaluations of a series of employment programs administered by Employment and Immigration Canada. It focuses specifically on the problem of dealing with selection bias in the context just described.

Debate on the optimal method for measuring effects within the labor market has continued for many years. Robert LaLonde and James

Heckman have figured prominently in the debate (Heckman, 1979; Heckman, Hotz, & Dabos, 1987; Heckman & Robb, 1985; Heckman & Singer, 1985; Lalonde & Maynard, 1987), but many others have contributed as well. A prominent issue in the debate concerns the choice between randomized experiments and nonexperimental or quasi-experimental methods. Contention over this issue is strongest with respect to social experiments or pilot tests of programs, where randomized experiments seem intuitively appropriate. In the context of training under a continuing program such as the Canadian Jobs Strategy, however, random rejection of qualified participants is politically, if not morally, inappropriate. For this reason, where analysts have tried this approach they have experienced a good deal of "leakage," which is to say that individuals who should not have participated in the program did so anyway. They might, for example, have participated after finding out about the program when, by design, they should not have; or they might have participated because some program administrators were unwilling to deny the program to these individuals. The focus of this article, therefore, is on nonexperimental methods.

The methods described here apply mainly to data on individuals. Without exception, the evaluations of programs at Employment and Immigration, as described above, have enjoyed the availability of data observed for individuals. Although some of the methods could work with grouped data, this article does not deal with that form of analysis.

DEFINITION AND STRUCTURE OF THE PROBLEM

Before considering how to measure effects, let us define the information required and examine the data typically available in such a context. Then we can begin looking at ways of using the data to determine the effect that a program might have.

Definition of *Effect*

We generally interpret *program effect* as indicating something that happened to participants in the program that would not have happened in the absence of the program. (Analysts often describe this technique as a comparison with a counterfactual.) Training programs usually strive to improve the earnings of their participants. Therefore, as an example of program effect, let us consider how annual earnings of trainees differ from what they would have been had the trainees not participated in the program.

Pre-Post Comparison

Some early efforts at measuring this kind of effect compared participants' earnings at some point after they stopped training with their earnings before training began. Differences in these values, adjusted for inflation, represented an effect of the training program. But we must consider how well pre-program earnings represent what would have happened to trainees in the absence of the program.

Comparison of Measures at Two Points in Time

Let us start by looking at a relatively simple measure of pre-program earnings, such as earnings in the year before training. This measure reflects an arbitrary and not necessarily representative period in the participant's pre-training history. For example, it is quite plausible that an individual's motive for participating in government training is to overcome a recent reversal of employment fortune, sometimes called the "dip" phenomenon. Earnings measured during such a period may provide a distorted indication of likely behavior in the absence of the program. Also, other factors, such as marital status, number of dependent children, or conditions in the local labor market, might influence the labor force behavior of individuals between the pre-program and post-program periods, independently of the effect of the training program. Therefore, it is difficult to justify simple comparisons of this kind.

Longitudinal Multivariate Regression Models

Regression analysis provides one way of overcoming, at least to some extent, the problems associated with simple comparison across time. Multivariate regression models can control for the influence of the factors mentioned above by including them as independent variables in the model. This approach thus eliminates the effect of known and measured influences, helping to isolate the effect of participation in the program. Furthermore, if we also include, as independent variables, measures of pre-program earnings taken over several different periods (hence "longitudinal"), our approach uses a much broader, and likely more stable, base of information to predict hypothetical behavior in the absence of the program.

Appealing as this method seems, it suffers from three significant problems. The first concerns limitations in the available data. We never know the extent to which the factors that we include in the model capture all the relevant influences we seek to eliminate from the analysis. For example, an individual's employment experience may be affected by that of his or her spouse, especially where the spouse's job entails relocating to another community. In some cases, data on factors that we believe have a strong influence on employment might not be available or measured reliably. Weaknesses in the data limit the usefulness of this approach, especially with regard to pre-program histories. Typically, data concerning pre-program earnings come from data files associated with the administration of the training program, and include only simple measures such as earnings in the year before training, as described earlier. Experience suggests it is rare to have more information than this available. Furthermore, capturing such information during a post-program survey would undermine its credibility because of problems of recall over a period that often covers several years.

A second problem is more technical in nature and leads to biased estimates of the program's effect. Economic theory suggests that the unexplained portion of an individual's earnings during a particular period may depend more on the unexplained portion of his or her earnings in the previous period than on measured factors. Econometricians call this phenomenon “serial correlation of errors.” If it occurs, it indirectly affects the resulting estimates of the program's effect on earnings by inflating their standard errors, thereby obscuring the proper inferences with respect to statistical significance.

Finally, although this kind of model can detect the “dip” phenomenon, the third problem associated with it concerns its ability to extrapolate or to project pre-program data into the post-program period. Is the dip the beginning of a new trend that would have continued in the absence of the training program, or would the individuals who experienced it have rebounded to some extent even without the program? This model does not clearly answer this question.

Comparison Group

Another approach to representing the likely behavior of participants in the absence of the program involves using a comparison group. Assuming that an experiment involving random assignment is not an acceptable method for measuring the effect of this kind of program, a comparison group probably offers the next best alternative to a randomly selected control group. The comparison group serves a similar function, in that it should provide an estimate of what would have happened to the participants had the program not been available.

In forming a useful comparison group, then, we seek to include people who resemble participants but who, for some unknown reason, did not participate. Program administrators strive to provide training to all who qualify to receive it. One might expect, therefore, that finding such a comparison group could prove difficult. However, budgetary constraints often impose limits on the numbers who can participate. Also, despite the best efforts of administrators, some candidates for training simply never hear about the program or have a chance to apply for it. The Longitudinal Labour Force Data Base at Employment and Immigration Canada has repeatedly proved to be a valuable source of comparison groups. However, such data are not always available and constructing the comparison group often requires considerable computer processing. It is nonetheless important that this task receive adequate attention and resources.

Balance

The comparison group should be roughly the same size as the group of participants being studied. In deciding how large to make the comparison group, we consider the efficiency of the analysis. In statistical language, this term refers to the ability of the analysis, given that a true program effect exists, to conclude that the effect measured from sampled data is statistically significant. Efficiency depends on the number of observations used in the analysis. A more efficient analytical method will produce statistically significant conclusions with fewer observations. One can prove mathematically that this kind of analysis will achieve maximum efficiency when the comparison group has as many members as there are participants in the analysis.

Matching

Early studies, to emulate the structure of a randomized experiment, created the comparison group as several subgroups, whose members matched program participants exactly with respect to some attributes. Age and gender are the two most basic attributes used for this purpose. For example, a subgroup might consist of males or females in one of five age categories, for a total of ten subgroups altogether. When constructing matched comparison groups, I have often created further subgroups using province and pre-program earnings or employment.

A second approach makes greater use of available information by retaining the original scale of measurement for the variables, rather than collapsing their values into categories. Among the variables just discussed, this comment applies to age and earnings. The approach uses a measure of distance between pairs of participants and candidates for the comparison groups. We use the term *distance* in the context of the variable space defined by the matching characteristics. Researchers often call this approach "nearest-neighbor" matching. Various algorithms exist for selecting the members of the comparison group from the available candidates. All seek to minimize the distance between pairs of selected points using different criteria.

Disadvantages of nearest-neighbor matching are that it might fail to select the number of candidates needed and that it often requires intensive computer processing. If the criteria for selection are too limiting, this approach will not yield as many matches as needed. Relaxing the criteria will increase the numbers selected but reduce the overall closeness of match. If the comparison group candidates do not match the participants well to begin with, this method offers no advantage over the first one. Also, depending on the specific algorithm used in constructing the comparison group, the processes of calculating all the pairwise distances involved and of selecting candidates to minimize a given distance-based criterion can consume a significantly large amount of computing resources.

Having looked at approaches to matching, let us consider to what extent the comparison group *should* match the participants, a subject of some debate in the past decade. A comparison group that matches the participants with respect to pre-program levels of a variable we will use in estimating the effect of the program helps to enhance the credibility of the analysis. At the same time, if the variables used for matching will also appear in the analysis, in a regression model, for example, the matching process will be redundant. Note that regression analysis offers much more efficiency than comparing mean earnings between pairs of several subgroups. On the other hand, if the comparison group differs too much from the participants, it can hardly represent the conditions we might expect the participants to have experienced had the training program not been available. As a rule, one should select members of the comparison group who appear to have been eligible to participate but who did not.

Overlap

Let us consider the above reasons for matching in more practical terms. We seek data from the comparison group to model the likely behavior of participants in the absence of the training program. Therefore, certainly in the context of regression models, we should ensure that the ranges covered by the comparison group's variables coincide as much as possible with the corresponding ranges for the participants. Any model that we use to estimate an outcome based on a certain set of conditions will provide much better estimates if the data used to determine the structure of the model reflect the same conditions as much as possible. Therefore, although it might seem redundant to match the comparison groups exactly, with respect to variables used in regression models, the distributions of the matching variables for the comparison group should roughly resemble those for the participants.

Form of the Outcome Measure

One outcome measure is the level of earnings following participation in the program. This measure could represent a figure that a participant or a member of the comparison group reports as his or her current-level earnings at the time of a follow-up survey. If more detailed information was available, one could calculate, for example, average earnings since leaving the training program.

The simplest analysis would entail comparing average values of such a measure for participants and for the comparison group. As I indicated earlier, regarding a similar simple comparison of averages, the analysis would be much more efficient if it controlled for other influences on earnings, thus isolating, to the extent possible, the effect of the program on earnings. Also, for this approach to appear credible, the comparison group must resemble the participants very closely with respect to average pre-program earnings. If construction of the comparison group used the "subgroup" approach we described above, as opposed to the nearest-neighbor method, then this average should match closely for each pair of subgroups.

To some extent, we can overcome the effects of a poor match of this kind by differencing the outcome variable. In other words, rather than comparing post-program levels of earnings, we construct the change in earnings from before training to the post-program period. The nature of this measure will depend on the quality of data available for both periods. Its definition could be subject to the arbitrariness described earlier. However, these deficiencies pose less of a problem in this context, if one is willing to assume they affect the comparison group in roughly the same way as they do the participants.

Measurable Influential Factors

A further enhancement to the analysis, using either form of the dependent variable, employs regression models to control for measurable factors that affect earnings. We can thus isolate and measure more precisely the effect of the training program, by including it as a binary or dummy variable in the regression model. It serves a useful purpose in distinguishing between such factors that remain constant over time and those that vary.

Constant Factors

Factors that do not change over time typically include gender, membership in a visible minority, location where the training occurred, and level of education at the time training began. This group could also include factors for which measurements would normally change over time but that are available for only one point in time (such as at the time of a postprogram follow-up survey) or at two points of time so close together that few observations exhibit any change (such as marital status). It could also include age, if we define age at a specific point in time (such as when training began or when a follow-up survey took place).

Changing Factors

Factors that change over time include marital status, number of dependent children, number of contributors to household income, and indicators of local labor market activity, such as the unemployment rate or the ratio of employed to total population. The variation in these factors over time raises the question of how we ought to measure them and use them in the analysis. If the period of time covered by the data is relatively short, in practice many of these variables will change very little, if at all. Models I have developed in several studies, owing to limitations on the data available, incorporate such variables measured at a single point in time, usually at the time of a survey.

One time-varying factor that we have often included in our analysis is the duration of the posttraining period. The value of including such a variable is that it permits some inference about whether the estimated program effect changes over time. In some cases, the main interest in such studies lies in the extent to which the program continues to have an effect or in its pattern of growth or attenuation following training. Including in the analysis one or more variables that represent the duration of the posttraining period provides some insight into this pattern. The analytical model may accommodate more elaborate patterns through the use of higher powers of this variable, such as its square or cube.

Unmeasurable Influential Factors

The advantage of the above kind of regression model applies only to the extent that available data reflect all factors that have a significant influence on earnings. Measuring and recording all possible attributes that might affect an individual's earnings is practically impossible. Despite the best efforts of people who maintain administrative databases or design survey instruments, one can always find some characteristic that, for whatever reasons, the data do not address. We refer to such characteristics as "unmeasurable." Researchers in this field often associate them with individual traits such as motivation or other latent aspects of a person's background or personality that affect the progress of his or her career.

Some researchers in this field have attempted to develop scales to measure motivational factors explicitly. However, I treat them here as unmeasurable factors because experience suggests that artificial measures of motivation do not adequately explain differences in labor market behavior between program participants and the comparison group. As will be seen below when we discuss using such information to deal with selection bias, it suffers a major weakness because it usually pertains to the period following participation in the training program, whereas the critical interest lies in the measurement of motivation *before* training, as a potential determinant of participation in the program.

Let us consider unmeasurable characteristics that would most plausibly affect earnings. As with the measured factors, we categorize the unmeasurable influences according to whether they change over time. We will see in a moment the influence that these kinds of variables have on our efforts to estimate program effects.

Constant Factors

Most individuals who might take government-sponsored training will likely have already developed most of the personality traits that might influence their careers. Therefore, we tend to think of these as latent traits, constant over the period of time under study, especially if constraints dictate that this period is relatively short. Some changes in attitudes and behaviors with respect to work and career may occur. To the extent that they occur as a result of participation in the program, we consider them part of the program's effect. Any other changes of this kind we can reasonably assume to have equal influence for participants and comparison group members alike.

Changing Factors

Unmeasurable factors that change over time consist mainly of conditions in the functioning of the labor market. To some extent, our analysis takes gross labor market trends into account by incorporating measures such as the local rate of unemployment, which is measurable. The very specific labor market conditions facing each individual should be approximately the same for participants and comparison group members. However, because of the latent motivational traits we have assumed above to be constant, some individuals may gain access to information about the labor market and about career opportunities that others will not.

DEALING WITH SELECTION BIAS

Before presenting methods to deal with the problem of selection bias, let us define that bias precisely in terms of the concepts and variables discussed so far. Simply put, the problem arises when unmeasured factors that affect a person's earnings also affect whether that person participates in the training program. For example, a person who is intrinsically more highly motivated toward career progress is more likely to participate in training, but would also be more likely to achieve higher earnings without the training. As will be seen below, the unmeasured factors of greatest concern are those that change over time, especially those involving information that could contribute to a person's decision to take training. The bias arising in this situation manifests itself in the estimates of the program's effects on earnings, and may be upward or downward. In other words, we cannot say beforehand whether the selection bias will cause estimates of the program's effect to be higher or lower than the actual effects we seek to estimate.

Although we cannot always directly determine the influence of selection bias on earnings, several methods permit us to adjust the estimates of program effect for selection bias, given some acceptable assumptions.

Inverse of Mill's Ratio

This method, initially developed by James Heckman (1979), has a somewhat complex mathematical derivation. It consists of two regression models applied in sequence. The first model (probit) uses data known for both the participants and the comparison group to estimate the probability of participation for each individual in the analysis. The inclusion of a certain function (called the inverse of Mill's ratio) of these estimated probabilities in the earnings model corrects the estimate of program effect for selection bias, as Heckman showed mathematically. However, this approach requires a specialized method of fitting the regression model of earnings that is not available in most statistical software packages. (William Green has implemented it, including the correct method of asymptotic covariance estimation given by Lee, Maddala, and Trost [1980], in a software product called Limdep.) Also, it has a reputation for being relatively inefficient, in the sense discussed above.

This approach offers the advantage of a simple test for the presence of selection bias. The test enables us to determine the effect of the correction, and whether a correction is even necessary.

Instrumental Variables

A more widely known method of econometric analysis offers an alternative approach to dealing with selection bias. The method of instrumental variables provides a solution to an even broader class of problems of which selection bias is but one example. Without going into technical details, available to the interested reader in any standard econometrics textbook, the method requires a variable that is correlated with participation in the program but not with the error term (what remains of earnings after adjusting for all other known factors) in the regression model.

Compared to Heckman's original method, this approach is much more widely available in software and usually achieves more efficient estimation of program effects. On two occasions when Professor Heckman has worked with us on evaluations of training studies, he has stated a preference for using the method of instrumental variables. He suggests using, as an instrument for the participation dummy variable, the probability of participating in the program as estimated by a probit model for each participant and member of the comparison group. Where this estimated probability fails to function properly as an instrument, Professor Heckman has suggested using a higher power of it, such as its square or cube.

Although theory suggests that this method might be generally preferable, it can generate problems when the instrument is correlated with other explanatory variables in the model. We have tried using transformations of the probability as instruments, including the inverse of Mill's ratio, to improve on matters, with mixed results. The problem becomes particularly acute when the probit model does a poor job of estimating probability of participation, that is, when it does not accurately predict membership in the two groups.

Controlling for Preprogram Levels

A third method of overcoming the problem of selection bias involves incorporating preprogram earnings in the analysis, in a manner parallel to that used earlier. This method offers a very practical advantage. People who have difficulty grasping the technical aspects of selection bias find this approach convincing because it intuitively makes the comparison group appear similar to a randomly selected control group. I outline two ways of using this information in the context of the methods discussed above.

Difference Model

First, let us consider the simple difference in earnings: posttraining earnings less pretraining earnings. One likely effect of selection bias is that pretraining earnings for participants will be quite different from those for the comparison group. If we apply the above analysis to the difference in earnings, however, we may remove this effect. Under certain specifications of the model, one can show that selection bias will arise from unmeasured characteristics only if they are of the kind that varies over time.

Lagged-Dependent Model

This model presents a slightly more general form of the difference model just discussed. Instead of working with a dependent variable formed as a difference, we add pretraining earnings as an independent variable on the right side of the regression model. In this formulation, the model accommodates a broader range of relationships between posttraining earnings and pretraining earnings.

Serial Correlation

Both of the above methods present a risk of further bias in the estimates of program impact, arising from serial correlation. This is a technical, statistical problem that arises when there is a strong relationship between a measure of earnings at one time and a comparable measure taken a specific interval before. My experience suggests, however, that such problems have a slight enough impact to justify using this kind of analysis.

Separate Models

Another extension of the above method involves specifying separate regression models for the participants and for the comparison group. With this arrangement, we estimate the effect of the program on earnings as the average difference in earnings that the models predict if we apply the estimated regression coefficients from each model to the data for the participants only. This method offers two advantages. First, the model for the participants can include variables related specifically to the training, such as duration of training or whether the participant completed the training as planned. Second, this method allows for differences, between participants and nonparticipants, in the relationships between earnings and the explanatory variables. The extent to which the two groups will require different models for this reason could depend on how closely they match each other.

EVALUATION OF THE METHODS

Summary of Discussion

The foregoing discussion, and the experience that generated it, suggests regression analysis of data from both participants and a comparison group as the preferred method for measuring the effects of training in the face of selection bias. The regression model should include information pertaining to earnings in the pretraining period, and as many measured variables as are available. If one is willing to assume that no selection bias arises from unmeasured factors that change over time, then an ordinary least squares regression model that controls for pretraining earnings will remove all selection bias from the analysis. Otherwise, one must use either Heckman’s two-step model or instrumental variables. Using separate models for the participant and comparison groups will provide a more general approach, but may not be necessary.

Recommendation

I recommend performing three parallel regression analyses of the kind just described. Although theory and experience may suggest that certain specific regression techniques should work better than others, I have unhappily found somewhat inconsistent behavior among these models. Using ordinary least squares, Heckman two-step, *and* instrumental variables models provides a choice of models from which to select estimates of program effect. You can make the choice on the basis of both the consistency of the estimates and the standard diagnostics of the regression models. Heckman's two-step method, though it may suffer from relative inefficiency, provides a convenient test for the presence of selection bias. If it shows insignificant selection bias, one may feel justified in using the estimate of program effect from the ordinary least squares regression model, which should generally be the most efficient.

REFERENCES

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 156–161.

Heckman, J.J., Hotz, V.J., & Dabos, M. (1987). Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review, 11*(4), 395–427.

Heckman, J.J., & Robb, R., Jr. (1985). Alternative methods for evaluating the impact of interventions. In J.J. Heckman and B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 156–245). Cambridge: Cambridge University Press.

Heckman, J.J., & Singer, B. (1985). Social science duration analysis. In J.J. Heckman and B. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 39–110). Cambridge: Cambridge University Press.

LaLonde, R., & Maynard, R. (1987). How precise are evaluations of employment and training programs: Evidence from a field experiment. *Evaluation Review, 11*(4), 428–451.

Lee, L.-F., Maddala, G.S., & Trost, R.P. (1980). Asymptotic covariance matrices of two-stage probit and two-stage tobit methods for simultaneous equations models with selectivity. *Econometrica, 48*(2), 491–503.