Drawing Inferences from Self-Selected Samples

Edited by
Howard Wainer

With 17 Figures

Springer-Verlag
New York Berlin Heidelberg
London Paris Tokyo

Howard Wainer
Educational Testing Service
Research Statistics Group
Princeton, New Jersey 08541
U.S.A.

Library of Congress Cataloging in Publication Data
Drawing inferences from self-selected samples.
Papers from a conference sponsored by
Educational Testing Service.
Includes bibliographical references and index.
1. Social sciences—Statistical methods.
2. Sampling (Statistics) 3. Educational statistics.
I. Wainer, Howard. II. Educational Testing Service.
HA31.2.D7 1986 519.5'2 86-15593

Typeset by Science Typographers, Medford, New York.
Printed and bound by R.R. Donnelley and Sons, Harrisonburg, Virginia.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-96379-0 Springer-Verlag New York Berlin Heidelberg
ISBN 3-540-96379-0 Springer-Verlag Berlin Heidelberg New York

Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes

JAMES J. HECKMAN AND RICHARD ROBB

I. Introduction

Social scientists *never* have access to true experimental data of the type sometimes available to laboratory scientists.[1] Our inability to use laboratory methods to independently vary treatments to eliminate or isolate spurious channels of causation places a fundamental limitation on the possibility of objective knowledge in the social sciences. In place of laboratory experimental variation, social scientists use subjective thought experiments. Assumptions replace data. In the jargon of modern econometrics, minimal identifying assumptions are invoked.

Because minimal identifying assumptions cannot be tested with data (all possible minimal assumptions for a model explain the observed data equally well, at least in large samples) and because empirical estimates of causal relationships are sensitive to these assumptions, inevitably there is scope for disagreement in the causal interpretation of social science data. Context, beliefs, and *a priori* theory resolve differences in causal interpretations (see Simon, 1957, for an early statement of this view). By definition, the available data cannot do so, although in principle, experiments can.[2] The solution to the problem of causal inference lies outside of mathematical statistics and depends on contexts which are not universal.

The problem of selection bias in the analysis of social science data is a special case of the general problem of causal inference from social science

[1] That the recently conducted so called social experiments are not laboratory experiments is abundantly clear from the literature on the topic. See, e.g., Fienberg et al. (1985) and the references cited therein.

[2] The second remark in this sentence abstracts from the very real problems of designing or conducting true social experiments. Self selection by agents (including attrition) and the vast multiplicity of possible channels of causal influence make experimentation problematic if not infeasible. Alleged solutions to the problems of self selection and attrition based on arbitrary normality assumptions *inject* into the analysis of experimental data subjective features which the experiments were proposed to avoid.

data. This paper considers alternative assumptions that have been or might be invoked to solve the problem of causal inference created by selection bias. In this paper we consider the following topics.

First, we define selection bias and the "structural" or causal parameters of interest for the following prototypical model. Persons are given (or else select) "treatments," but the assignment of persons to treatments is nonrandom. Differences in measured outcomes among persons are due to the treatment and to factors that would make people different on outcome measures even if there were no causal effect of the treatment. "Treatment," as we use the term, may be a drug trial, a training program, attending school, joining a union, or migrating to a region. For specificity, in this paper we consider a training program but only because it provides a convenient prototypical context against which it is possible to gauge the plausibility of certain identifying assumptions. Except for this contextual content, our analysis is applicable to all of the other treatments mentioned above. We assume that it is not possible to simultaneously observe the same person in the treated and untreated states. If it is possible to observe the same person in both states, the problem of selection bias disappears.[3] By observing the same person in the treated and untreated states, it is possible to isolate the treatment in question without having to invoke any further assumptions.

A careful definition of the causal or structural parameter of interest is an essential aspect of this paper. The literature in social science is unclear on this issue. Conventional "average treatment" definitions often used in the statistics literature (see, e.g., Rosenbaum and Rubin, 1983) often do not define the parameter of interest to behavioral social scientists.

A second topic considered in this paper is the specification of minimal identifying assumptions required to isolate the parameters of interest. We consider the plausibility of these assumptions in the context of well-formulated models of the impact of training on earnings.

We present assumptions required to use three types of widely available data to solve the problem of estimating the impact of training on earnings free of selection bias: (1) a single cross section of post-training earnings, (2) a temporal sequence of cross sections of unrelated people (repeated cross-section data), and (3) longitudinal data in which the same individuals are followed over time. These three types of data are listed in order of their availability and in inverse order of their cost of acquisition. If we assume that random sampling techniques are applied to collect all three types of data, then the three sources form a hierarchy: longitudinal data can be used to generate a single cross section or a set of repeated cross sections in which the identities of individuals are ignored, and repeated cross sections can be used as single cross sections.

---

[3] The Glynn et al. paper (this volume) essentially makes this assumption and thus assumes away selection bias in the mixture modeling section.

Our conclusions are rather startling. Although longitudinal data are widely regarded in the social science and statistical communities as a panacea for selection and simultaneity problems, there is no need to use longitudinal data to identify the impact of training on earnings if conventional specifications of earnings functions are adopted.[4] Estimators based on repeated cross-section data for unrelated persons identify the same parameter. This is true for virtually all longitudinal estimators.

However, in this paper we question the plausibility of conventional specifications. They are not often motivated by any behavioral theory, and when examined in the light of a plausible theory, conventional specifications seem poorly motivated. We propose richer longitudinal specifications of the earnings process and enrollment decision derived from behavioral theory. In addition, we propose a variety of new estimators. A few of these estimators require longitudinal data, but for most, such data are not required. A major conclusion of our paper is that the relative benefits of longitudinal data have been overstated because the potential benefits of cross-section and repeated cross-section data have been understated.

When minimal identifying assumptions are explored for models fit on the three types of data, we find that *different* and not necessarily more plausible assumptions can be invoked in longitudinal analyses than in cross-section and repeated cross-section analyses. The fact that more types of minimal identifying assumptions can be invoked with longitudinal data (since the longitudinal data can be used as a cross section or a repeated cross section) does not make more plausible those assumptions that uniquely exploit longitudinal data.

In analyzing the assumptions required to use various data sources to consistently estimate the impact of training on earnings free of selection bias, we discuss the following topics:

(1) How much prior information about the earnings function must be assumed?
(2) How much prior information about the decision rule governing participation must be assumed?
(3) How robust are the proposed methods to the following commonly encountered features of data on training?
    (a) nonrandomness of available samples and especially oversampling of trainees (the choice-based sample problem);
    (b) time inhomogeneity in the environment ("nonstationarity"); and
    (c) the absence of a control group of nontrainees or the contamination of the control group so that the training status of individuals is not known for the control sample.

[4] Conventional specifications include "fixed effect" or "autoregressive" assumptions for the error terms of earnings equations. These terms are defined below.

We also question recent claims of the sort made in the paper by Glynn et al. (this volume) which state that cross-section approaches to solving the problem of selection bias are strongly dependent on arbitrary assumptions about distributions of unobservables and on certain arbitrary exclusion restrictions (see also Little, 1985, where such claims are also made). While some widely used cross-section estimators suffer from this defect, such commonly invoked assumptions are not an essential feature of the cross-section approach, at least for the type of selection problem considered in this paper. However, we demonstrate that unless explicit distributional assumptions are invoked, all cross-section estimators require the presence of at least one regressor variable in the decision rule determining training. This requirement may seem innocuous, but it rules out a completely nonparametric cross-section approach. Without prior information, it is not possible to cross-classify observations on the basis of values assumed by explanatory variables in the earnings function and do "regressor-free" estimation of the impact of training on earnings that is free of selection bias. A regressor is required in the equation determining enrollment. Without a regressor it is necessary to invoke distributional assumptions. Longitudinal and repeated cross-section estimators do not require a regressor.

A third topic considered in this paper is an assessment of both the "mixture modeling" approach to the selection bias problem advocated by Glynn et al. (this volume) and the propensity score methodology of Rosenbaum and Rubin (1983, 1985) that has been advocated as an alternative to selection bias procedures by Coleman (1985) and Scheuren (1985). We make two points. First, under the Glynn et al. assumptions about the data-generating mechanism, there is no real problem of selection bias. Those authors assume access to data on participants and nonparticipants which when appropriately weighted can produce unbiased estimates of treatment impact. Such data are typically not available. Second, propensity score methods solve the selection bias problem only in special cases that often turn out to be behaviorally uninteresting. The propensity score is not a panacea or genuine alternative methodology for general selection bias problems.

The focus of this paper is on model identification and not on estimation or on the efficiency of alternative estimators. As noted by Tukey in his discussion of this paper, identification is a rather sharp concept that may not be all that helpful a guide to what will "work" in practice. Different estimators may perform quite differently in practice depending on the degree of overidentification or on the nature of the identifying restriction. However, if a parameter is not identified, an estimator of that parameter cannot have any desirable statistical properties. Securing identification is a necessary first step toward construction of a desirable estimator, but certainly is not the last step. This paper concentrates on the necessary first step.

Our focus on identification and on the tradeoffs in assumptions that secure identification should make clear that we are not offering a nostrum for selection bias that "works" in all cases. In our view, the recent literature on this subject has been marred by analysts who claim to offer context-free universal cures for the selection problem. There are almost as many cures as there are contexts, and for that reason no one cure can be said to "work" for all problems.[5]

We focus on identification because the current literature in social science and statistics is unclear on this topic. A major goal of this paper is to demonstrate that previous work on selection bias has often imposed unnecessarily strong assumptions (e.g., normality). Part of the great variability in estimates obtained in some analyses using selection bias procedures may be due to the imposition of different types of extra conditions not required to identify the parameters of interest. Separating out essential from inessential assumptions is a main goal of this paper.

The way we have written this paper may cause some confusion. We establish identifiability by establishing the existence of consistent estimators. Thus, we combine two topics that might fruitfully be decoupled. By establishing the consistency of a variety of estimators, we present a large sample guide to estimation under a variety of assumptions. However, the price of this approach to model identification is that we invoke assumptions not strictly required for identification alone (see Barros, 1986, where this type of separation of assumptions is done). Fewer assumptions are required for identification than are required for consistent estimation. As noted by Tukey in his written comments on this paper, clarity would be served if the reader mentally substituted "c-identified" (for consistency identified) for "identified" everywhere the subject of identification is discussed in this paper.

We have already noted that we use large sample theory in our analysis. Given the size of many social science data sets with hundreds and thousands of independent observations and given the available Monte Carlo evidence, large sample methods are not unreliable. For these reasons, we view our large sample analysis as the natural point of departure for research on selection models.

We do not discuss efficiency or variance questions in this paper. A discussion of efficiency makes sense only within the context of a fully specified model. The focus in this paper is on the tradeoffs in assumptions that must be imposed to estimate a single coefficient when the analyst has access to different types of data. Since different assumptions about the underlying model are invoked to justify the validity of alternative estimators, an efficiency or sampling variance comparison is often meaningless. Under the assumptions about an underlying model that justify one estima-

[5] Except, of course, the cure of genuine experimental data.

tor, properties of another estimator may not be defined. Only by postulating a common assumption set that is unnecessarily large for any single estimator is it possible to make efficiency comparisons. For the topic of this paper—model identification—the efficiency issue is a red herring.

Even if a common set of assumptions about the underlying model is invoked to justify efficiency comparisons for a class of estimators, conventional efficiency comparisons are often meaningless for two reasons. First, the frequently stated claim that longitudinal estimators are more efficient than cross-section estimators is superficial. It ignores the relative sizes of the available cross-section and longitudinal samples. Because of the substantially greater cost of collecting longitudinal data free of attrition bias, the number of persons followed in longitudinal studies rarely exceeds 500 in most social science analyses. In contrast, the available cross-section and repeated cross-section samples have thousands of observations. Given the relative sizes of the available cross-section and longitudinal samples, "inefficient" cross-section and repeated cross-section estimators might have much smaller sampling variances than "efficient" longitudinal estimators that are fit on much smaller samples. In this sense, our proposed cross-section and repeated cross-section estimators might be feasibly efficient given the relative sizes of the samples for the two types of data sources. However, we do not analyze this topic further in our paper.

Second, many of the cross-section and repeated cross-section estimators proposed in this paper require only sample means of variables. They are thus very simple to compute and are also robust to mean zero measurement error in all of the variables. Some more sophisticated longitudinal and cross-section estimators are computationally complex and in practice are often implemented on only a fraction of the available data to save computing cost. Simple methods based on means use all of the data and thus, in practice, might be more efficient.

Barros (1986) presents a very thorough discussion of the efficiency issue for alternative selection estimators for cases where the concept is well defined.

This paper draws heavily on our previous work (Heckman and Robb, 1985). To avoid repetition and to focus on essential points, we refer the reader to our longer companion paper for technical details of certain arguments. However, we use this paper to correct some minor typographical and conceptual errors that appeared in our previous work.

The organization of this paper is as follows. Section II describes the notation and a behavioral model of the enrollment of persons into training. Section III discusses the definition of the appropriate causal or structural parameter of interest. Sections IV–IX present a discussion of alternative estimation methods for different types of data. Section X discusses "mixture modeling" and "propensity score" methods as solutions to the selection bias problems considered in this paper and relates the propensity score method to techniques developed in the econometrics literature. Propensity

score methods are demonstrated to solve only a special and not very interesting case of the general selection bias problem.

II. Notation and a Model of Program Participation

A. Earnings Functions

To focus on essential aspects of the problem, we assume that individuals experience only one opportunity to participate in training. This opportunity occurs in period k. Training takes a single period for participants to complete. During training, participants earn no labor income.

Denote the latent earnings of individual i in period t by Y_{it}^*. These are the earnings of an individual in the absence of the existence of any training programs. Latent earnings depend on a vector of observed characteristics X_{it}. Let U_{it} represent the error term in the latent earnings equation and assume that

$$E(U_{it}|X_{it}) = 0.$$

Adopting a linear specification, we write latent earnings as

$$Y_{it}^* = X_{it}\beta + U_{it},$$

where β is a vector of parameters. Linearity is adopted only as a convenient starting point and is not an essential aspect of any of the methods presented in this paper. Throughout this paper we assume that the mean of U_{it} given X_{it} is the same for all X_{it}. Sometimes we will require independence between X_{it} and current, future, and lagged values of U_{it}. When X_{it} contains lagged values of Y_{it}^*, we assume that the equation for Y_{it}^* can be solved for a reduced form expression involving only exogenous regressor variables. Under standard conditions, it is possible to estimate the structure from the reduced form so defined.

Under these assumptions, β is the coefficient of X in the conditional expectation of Y^* given X. Observed earnings Y_{it} are related to latent earnings Y_{it}^* in the following way:

$$\begin{aligned} Y_{it} &= Y_{it}^* + d_i\alpha \qquad && t > k \\ Y_{it} &= Y_{it}^* && t \leq k, \end{aligned}$$

where $d_i = 1$ if the person takes training and $d_i = 0$ otherwise and where α is one definition of the causal or structural effect of training on earnings. Observed earnings are the sum of latent earnings and the structural shift term $d_i\alpha$ that is a consequence of training. Y_{it} is thus the sum of two random variables when $t > k$.

The problem of selection bias arises because d_i may be correlated with U_{it}. This is a consequence of selection decisions by agents. Thus, selection bias is present if

$$E(U_{it}d_i) \neq 0.$$

Observed earnings may be written as

$$\begin{aligned} Y_{it} &= X_{it}\beta + d_i\alpha + U_{it} \qquad t > k \\ Y_{it} &= X_{it}\beta + U_{it} \qquad t \leq k, \end{aligned} \tag{1}$$

where β and α are parameters. Because of the covariance between d_i and U_{it},

$$E(Y_{it}|X_{it}, d_i) \neq X_{it}\beta + d_i\alpha.$$

Equation (1) assumes that training has the same effect on everyone. In the next section we consider issues that arise when α varies among individuals, as is assumed in many analyses of experimental and nonexperimental data (see Fisher, 1953; Rosenbaum and Rubin, 1983). Throughout most of this paper we ignore effects of training which grow or decay over time (see our companion paper for a discussion of this topic).

We now develop the stochastic relationship between d_i and U_{it} in equation (1). For this purpose, we present a more detailed notation which describes the enrollment rules that select individuals into training.

B. Enrollment Rules

The decision to participate in training may be determined by a prospective trainee, by a program administrator, or both. Whatever the specific content of the rule, it can be described in terms of an index function framework. Let IN_i be an index of net benefits to the appropriate decision makers from taking training. It includes the loss of income in period k if training is taken. It is a function of observed (Z_i) and unobserved (V_i) variables. Thus,

$$\text{IN}_i = Z_i\gamma + V_i. \tag{2}$$

In terms of this function,

$$\begin{aligned} d_i &= 1 \quad \text{iff} \quad \text{IN}_i > 0 \\ d_i &= 0 \quad \text{otherwise.} \end{aligned}$$

The distribution function of V_i is denoted as $F(v_i) = \Pr(V_i < v_i)$. V_i is assumed to be independently and identically distributed across persons. Let $p = E(d_i) = \Pr(d_i = 1)$ and assume $1 > p > 0$. Assuming that V_i is distributed independently of Z_i (a requirement not needed for most of the estimators considered in this paper), we may write $\Pr(d_i = 1|Z_i) = F(-Z_i\gamma)$, which is sometimes called the "propensity score" in statistics (see, e.g., Rosenbaum and Rubin, 1983). In Section X we demonstrate that a special subclass of econometric selection–correction estimators can be expressed as functions of the propensity score.

The condition for the existence of selection bias

$$E(U_{it}d_i) \neq 0$$

may occur because of stochastic dependence between U_{it} and the unobservable V_i in equation (2) (selection on the unobservables) or because of stochastic dependence between U_{it} and Z_i in equation (2) (selection on the observables).

To interpret various specifications of equation (2), we need a behavioral model. A natural starting point is a model of trainee self-selection based on a comparison of the expected value of earnings with and without training. For simplicity, we assume that training programs accept all applicants.[6]

All prospective trainees are assumed to discount earnings streams by a common discount factor of $1/(1+r)$. Training raises trainee earnings by α per period. While in training, individual i receives a subsidy S_i, which may be negative (so there may be direct costs of program participation). Trainees forego income in training period k. To simplify the expressions, we assume that people live forever.

As of period k, the present value of earnings for a person who does not receive training is

$$PV_i(0) = E_{k-1}\left[\sum_{j=0}^{\infty}\left(\frac{1}{1+r}\right)^j Y^*_{i,k+j}\right].$$

E_{k-1} means that the mathematical expectation is taken with respect to information available to the prospective trainee in period $k-1$. The expected present value of earnings for a trainee is

$$PV_i(1) = E_{k-1}\left[\sum_{j=1}^{\infty}\left(\frac{1}{1+r}\right)^j Y^*_{i,k+j} + \sum_{j=1}^{\infty}\frac{\alpha}{(1+r)^j}\right].$$

The risk neutral wealth-maximizing decision rule is to enroll in the program if $PV_i(1) > PV_i(0)$ or, letting IN_i denote the index function in the decision rule of equation (2),

$$\mathrm{IN}_i = PV_i(1) - PV_i(0) = E_{k-1}(S_i - Y_{ik} + \alpha/r), \tag{3}$$

so the decision to train is characterized by the rule

$$\begin{aligned} d_i &= 1 \quad \text{iff} \quad E_{k-1}(S_i - Y_{ik} + \alpha/r) > 0 \\ d_i &= 0 \quad \text{otherwise.} \end{aligned} \tag{4}$$

Let W_i be the part of the subsidy which the analyst observes (with associated coefficient ϕ) and let τ_i be the part which he or she does not observe:

$$S_i = W_i\phi + \tau_i.$$

[6] Our previous paper considers more general models.

A special case of this model arises when agents possess perfect foresight so that $E_{k-1}(S_i) = S_i$, $E_{k-1}(Y_{ik}) = Y_{ik}$, and $E_{k-1}(\alpha/r) = \alpha/r$. Collecting terms,

$$\begin{aligned} d_i &= 1 \quad \text{iff} \quad S_i - Y_{ik} + \alpha/r = W_i\phi + \alpha/r - X_{ik}\beta + \tau_i - U_{ik} > 0 \\ d_i &= 0 \quad \text{otherwise.} \end{aligned} \tag{5}$$

Then $\tau_i - U_{ik} = V_i$ in equation (2) and (W_i, X_{ik}) corresponds to Z_i in equation (2). If we assume that (W_i, X_{ik}) is distributed independently of V_i, then inequalities (5) define a standard discrete choice model. This assumption is only required for some of the estimators discussed in this paper.

Suppose decision rule (5) determines enrollment. If the costs of program participation are independent of U_{it} for all t (so both W_i and τ_i are independent of U_{it}), then $E(U_{it}d_i) = 0$ only if the mean of the unobservables in period t does not depend on the unobservables in period k, or

$$E(U_{it}|U_{ik}) = 0 \qquad t > k.$$

Whether or not U_{it} and d_i are uncorrelated hinges on the serial dependence properties of U_{it}. If U_{it} is a moving average of order of m so that

$$U_{it} = \sum_{j=1}^{m} a_j \epsilon_{i,t-j},$$

where $\epsilon_{i,t-j}$ are iid, then for $t - k > m$, $E(U_{it}d_i) = 0$. However, if U_{it} follows a first-order autoregressive scheme, then $E(U_{it}|U_{ik}) \neq 0$ for all t and k.

The enrollment decision rules derived in this subsection give context to the selection bias problem. The estimators discussed in this paper differ greatly in their dependence on particular features of these rules. Some estimators do not require that these decision rules be specified at all, while other estimators require a great deal of *a priori* specification of these rules. Given the inevitable controversy that surrounds specification of enrollment rules, there is always likely to be a preference by analysts for estimators that require little prior knowledge about the decision rule.

III. Random Coefficients and the Structural Parameter of Interest

We identify two different definitions associated with the notion of a selection bias-free estimate of the impact of training on earnings. The first notion defines the structural parameter of interest as the impact of training on earnings if people are randomly assigned to training programs. The second notion defines the structural parameter of interest in terms of the difference between the

postprogram earnings of the trained and what the earnings in postprogram years for these same individuals would have been in the absence of training. The two notions come to the same thing only when training has an equal impact on everyone or else assignment to training is random with respect to earnings and attention centers on estimating the mean response to training. The second notion is the most useful one for forecasting future program impacts when the same enrollment rules that have been used in available samples characterize future enrollment.

In seeking to determine the impact of training on earnings in the presence of nonrandom assignment of persons to training, it is useful to distinguish two questions that are frequently confused in the literature.

Question 1: "*What would be the mean impact of training on earnings if people were randomly assigned to training*?"

Question 2: "*How do the postprogram mean earnings of the trained compare to what they would have been in the absence of training*?"

The second question makes a hypothetical contrast between the postprogram earnings of the trained in the presence and in the absence of training programs. This hypothetical contrast eliminates factors that would make the earnings of trainees different from those of nontrainees even in the absence of any training program. The two questions have the same answer if equation (1) generates earnings so that training has the same impact on everyone. The two questions also have the same answer if there is random assignment to training and if attention centers on estimating the *population* mean response to training.

In the presence of nonrandom assignment and variation in the impact of training among persons, the two questions have different answers. Question two is the appropriate one to ask if interest centers on forecasting the change in the mean of the post-training earnings of trainees compared to what *they* would have earned in the absence of training when the same selection rule pertains to past and future trainees. It is important to note that the answer to this question is all that is required to estimate the future program impact if future selection criteria are like past criteria and all that is required is to evaluate the gross return from training (the return exclusive of leisure and direct costs).[7]

To clarify these issues, we consider a random coefficient version of equation (1) in which α varies in the population. In this model, the impact

[7]There is a third question that might be asked: "What would be the effect of training on the earnings of the trained if the future selection rule for trainees differs from the past selection rule?" This question is more ambitious than the two stated in the text and requires that more assumptions be made. Given the general interest in questions one and two, we feel that a discussion of the answers to these two questions should precede a discussion of the answer to question three.

of training may differ across persons and may even be negative for some people. We write in place of equation (1)

$$Y_{it} = X_{it}\beta + d_i\alpha_i + U_{it} \qquad t > k.$$

Define $E(\alpha_i) = \bar{\alpha}$ and $\epsilon_i = \alpha_i - \bar{\alpha}$ so $E(\epsilon_i) = 0$. With this notation, we can rewrite the equation above as

$$Y_{it} = X_{it}\beta + d_i\bar{\alpha} + (U_{it} + d_i\epsilon_i). \tag{6}$$

Note that the expected value of the term in parentheses is nonzero. X_{it} is assumed to be independent of (U_{it}, ϵ_i). An alternative way to derive this equation is to express it as a two-sector switching model following Roy (1951), Goldfeld and Quandt (1976), Heckman and Neumann (1977), and Lee (1978). Let

$$Y_{1it} = X_{it}\beta_1 + U_{1it}$$

be the wage of individual i in sector 1 in period t. Let

$$Y_{0it} = X_{it}\beta_0 + U_{0it}$$

be the wage of individual i in sector 0. X_{it} is independent of (U_{1it}, U_{0it}). Let $d_i = 1$ if a person is in sector 1 and let $d_i = 0$ otherwise. We may write the observed wage as

$$\begin{aligned} Y_{it} &= d_iY_{1it} + (1 - d_i)Y_{0it} \\ &= X_{it}\beta_0 + E(X_{it}|d_i = 1)(\beta_1 - \beta_0)d_i \\ &\quad + \{[X_{it} - E(X_{it}|d_i = 1)](\beta_1 - \beta_0) + U_{1it} - U_{0it}\}d_i + U_{0it}. \end{aligned}$$

Letting $\bar{\alpha} = E(X_{it}|d_i = 1)(\beta_1 - \beta_0)$, $\epsilon_i = [X_{it} - E(X_{it}|d_i = 1)](\beta_1 - \beta_0) + U_{1it} - U_{0it}$, $\beta_0 = \beta$, and $U_{0it} = U_{it}$, produces equation (6).

In this model there is a fundamental nonidentification result when no regressors appear in the decision rule of equation (2). Without a regressor in equation (2) and in the absence of any further distributional (or moment) assumptions, it is not possible to identify $\bar{\alpha}$ unless $E(\epsilon_i|d_i = 1, Z_i) = 0$ or some other known constant.

To see this, note that

$$E(Y_{it}|d_i = 1, Z_i, X_{it}) = X_{it}\beta + \bar{\alpha} + E(\epsilon_i|d_i = 1, Z_i) + E(U_{it}|d_i = 1, Z_i)$$

$$E(Y_{it}|d_i = 0, Z_i, X_{it}) = X_{it}\beta + E(U_{it}|d_i = 0, Z_i).$$

Unless $E(\epsilon_i|d_i = 1, Z_i)$ is known, without invoking distributional assumptions, it is impossible to decompose $\bar{\alpha} + E(\epsilon_i|d_i = 1, Z_i)$ into its constituent components unless there is independent variation in $E(\epsilon_i|d_i = 1, Z_i)$ across observations, i.e., unless a regressor appears in equation (2). Without a regressor, $E(\epsilon_i|d_i = 1, Z_i)$ is a constant which cannot be distinguished from $\bar{\alpha}$.

This means that in models without regressors in the decision rule, we might as well work with the redefined model

$$Y_{it} = X_{it}\beta + d_i\alpha^* + \{U_{it} + d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]\}, \tag{7}$$

where

$$\alpha^* = \bar{\alpha} + E(\epsilon_i | d_i = 1),$$

and content ourselves with the estimation of α^*. If everywhere we replace α with α^*, the fixed coefficient analysis of equation (1) applies to equation (7), provided that account is taken of the new error component in the disturbance when computing variances.

The parameter α^* answers question two. It addresses the question of determining the effect of training on the people selected as trainees. This parameter is useful in making forecasts when the same selection rule operates in the future that has operated in the past. In the presence of nonrandom selection into training, it does not answer question one. Indeed, without regressors in the decision rule of equation (2), question one cannot be answered, so far as we can see, unless specific distributional assumptions are invoked.

Random assignment of persons to training does not usually represent a relevant or interesting policy option. For this reason, we will focus attention on question two. Hence, if the training impact varies among individuals, we will seek to estimate α^* in equation (7). Since equation (7) may be reparametrized in the form of equation (1), we work exclusively with the fixed coefficient earnings function. Our earlier paper gives precise statements of conditions under which $\bar{\alpha}$ is identified in a random coefficient model (see Barros, 1986, for a more complete discussion).

Much of the statistical literature assumes that $\bar{\alpha}$ is the parameter of interest (see Fisher, 1953; Lee, 1978; Rosenbaum and Rubin 1983). In the context of estimating the impact of nonrandom treatments that are likely to be nonrandomly assigned in the future, $\bar{\alpha}$ is not an interesting policy or evaluation parameter since it does not recognize selection decisions by agents. Only if random assignment is to be followed in the future is there interest in this parameter. Of course, α^* is interesting for prediction purposes only to the extent that current selection rules will govern future participation. In this paper we do not address the more general problem of estimating future policy impacts when selection rules are changed. To answer this question requires stronger assumptions on the joint distribution of ϵ_i, U_{it}, and V_i than are required to estimate $\bar{\alpha}$ or α^*.

It is also important to note that any definition of the structural treatment coefficient is conditioned on the stability of the environment in which the program is operating. In the context of a training program, a tenfold expansion of training activity may affect the labor market for the trained and raise the cost of the training activity (and hence the content of programs). For either $\bar{\alpha}$ or α^* to be interesting parameters, it must be assumed that such effects are not present in the transition from the sample period to the future. If they are present, it is necessary to estimate how the change in the environment will affect these parameters. In this paper we abstract from these issues, as well as other possible sources of interdepen-

dence among outcomes. The resolution of these additional problems would require stronger assumptions than we have invoked here.[8]

Before concluding this section, it is important to note that there is a certain asymmetry in our analysis which, while natural in the context of models for the evaluation of the impact of training on earnings, may not be as natural in other contexts. In the context of a training program (and in the context of the analysis of schooling decisions), it is natural to reason in terms of a latent earnings function Y_{it}^* which exists in the absence of schooling or training options. "U_{it}" can be interpreted as latent ability or as skill useful in both trained and untrained occupations. Because of the natural temporal ordering of events, pretraining earnings is a natural concept and α_i is the markup (in dollar units) of skills due to participation in training. Note that nothing in this formulation restricts agents to have one or just two skills. Training can uncover or produce a new skill or enhance a single common skill. Parameter α^* is the gross return to training of the trained before the direct costs of training are subtracted.

In other contexts there is no natural temporal ordering of choices. In such cases the concept of α^* must be refined since there is no natural reference state. Corresponding to a definition of the gross gain using one state as a benchmark, there is a definition of gross gain using the other state as a benchmark. In the context of the Roy model [discussed following equation (6)], it is appropriate for an analysis of economic returns to outcomes to compute a gross gain for those who select sector 1 which compares *their* average earnings in sector 1 with what they would have earned on average in sector 0 and to compute a gross gain for those who select sector 0 which compares their average earnings in sector 0 with what they would have earned on average in sector 1.

To state this point more clearly, assume that X_{it} in the expression following equation (6) is a constant ($=1$), and drop the time subscripts to reach the following simplified Roy model:

$$Y_{1i} = \mu_1 + U_{1i}$$
$$Y_{0i} = \mu_0 + U_{0i}.$$

In this notation,

$$\bar{\alpha} = \mu_1 - \mu_0$$
$$\epsilon_i = U_{1i} - U_{0i}.$$

The average gross gain for those who enter sector 1 from sector 0 is

$$\alpha_1^* = E(Y_{1i} - Y_{0i} | d_i = 1) = \bar{\alpha} + E(\epsilon_i | d_i = 1).$$

[8] This issue renders invalid use of estimates from the pilot negative income tax programs as estimates of the impact of a national negative income tax program. In the context of data from large-scale training programs, this issue is less cogent.

The average gross gain for those who enter sector 0 from sector 1 is

$$\alpha_0^* = E(Y_{0i} - Y_{1i} | d_i = 0) = -\bar{\alpha} - E(\epsilon_i | d_i = 0).$$

Both coefficients compare the average earnings in the outcome state and the average earnings in the alternative state for those who are in the outcome state. In a more general analysis, both α_1^* and α_0^* might be of interest. Provided that $\bar{\alpha}$ can be separated from $E(\epsilon_i | d_i = 1)$, α_0^* can be estimated exploiting the facts that $E(\epsilon_i) = 0$ and $E(d_i) = p$ are assumed to be known or estimable. No further identification conditions are required. For the sake of brevity and to focus on essential points, we do not develop this more general analysis here. The main point of this section—that $\bar{\alpha}$, the parameter of interest in statistical studies of selection bias, is not the parameter of behavioral interest—remains intact.

IV. Cross-Section Procedures

Standard cross-section procedures invoke unnecessarily strong assumptions. All that is required to identify α in a cross section is access to a regressor in equation (2). In the absence of a regressor, assumptions about the marginal distribution of U_{it} can be exploited to produce consistent estimators of the training impact.

A. Without Distributional Assumptions a Regressor Is Needed

Let $\overline{Y}_t^{(1)}$ denote the sample mean of trainee earnings and let $\overline{Y}_t^{(0)}$ denote the sample mean of nontrainee earnings:

$$\overline{Y}_t^{(1)} = \frac{\sum d_i Y_{it}}{\sum d_i}$$

$$\overline{Y}_t^{(0)} = \frac{\sum (1 - d_i) Y_{it}}{\sum (1 - d_i)},$$

assuming $0 < \sum d_i < I_t$, where I_t is the number of observations in period t. We retain the assumption that the data are generated by a random sampling scheme. If no regressors appear in equation (1), then $X_{it}\beta = \beta_t$ and

$$\operatorname{plim} \overline{Y}_t^{(1)} = \beta_t + \alpha + E(U_{it} | d_i = 1)$$

$$\operatorname{plim} \overline{Y}_t^{(0)} = \beta_t + E(U_{it} | d_i = 0).$$

Thus,

$$\operatorname{plim}\left(\overline{Y}_t^{(1)} - \overline{Y}_t^{(0)}\right) = \alpha + \frac{E(U_{it} | d_i = 1)}{1 - p}$$

since $pE(U_{it}|d_i=1)+(1-p)E(U_{it}|d_i=0)=0$. Even if p were known, α cannot be separated from $E(U_{it}|d_i=1)$ using cross-section data on sample means. Sample variances do not aid in securing identification unless $E(U_{it}^2|d_i=0)$ or $E(U_{it}^2|d_i=1)$ are known *a priori*. Similar remarks apply to the information available from higher moments unless they are restricted in some *a priori* fashion.

B. Overview of Cross-Section Procedures Which Use Regressors

If, however, $E(U_{it}|d_i=1, Z_i)$ is a nonconstant function of Z_i, it is possible (with additional assumptions) to solve this identification problem. Securing identification in this fashion explicitly precludes a fully nonparametric strategy in which both the earnings function of equation (1) and decision rule of equation (2) are estimated in each (X_{it}, Z_i) stratum. For within each stratum, $E(U_{it}|d_i=1, Z_i)$ is a constant function of Z_i and α is not identified from cross-section data. Restrictions across strata are required.

If $E(U_{it}|d_i=1, Z_i)$ is a nonconstant function of Z_i, it is possible to exploit this information in a variety of ways depending on what else is assumed about the model. Here we simply sketch alternative strategies. In our earlier paper, we presented a systematic discussion of each approach.

(a) Suppose Z_i or a subset of Z_i is exogenous with respect to U_{it}. Under conditions specified more fully below, the exogenous subset may be used to construct an instrumental variable for d_i in equation (1), and α can be consistently estimated by instrumental variables methods. No explicit distributional assumptions about U_{it} or V_i are required (Heckman, 1978). The enrollment rule of equation (2) need not be fully specified.

(b) Suppose that Z_i is distributed independently of V_i and the functional form of the distribution of V_i is known. Under standard conditions, γ in equation (2) can be consistently estimated by conventional methods in discrete choice analysis (Amemiya, 1981). If Z_i is distributed independently of U_{it}, $F(-Z_i\hat{\gamma})$ can be used as an instrument for d_i in equation (1) (Heckman, 1978).

(c) Under the same conditions as specified in (b),

$$E(Y_{it}|X_{it}, Z_i) = X_{it}\beta + \alpha[1-F(-Z_i\gamma)].$$

γ and α can be consistently estimated using $F(-Z_i\hat{\gamma})$ in place of $F(-Z_i\gamma)$ in the preceding equation (Heckman 1976, 1978) or else the preceding equation can be estimated by nonlinear least squares, estimating β, α and γ jointly (given the function form of F) (Barnow et al., 1980).

(d) If the functional forms of $E(U_{it}|d_i=1, Z_i)$ and $E(U_{it}|d_i=0, Z_i)$ as functions of Z_i are known up to a finite set of parameters, it is sometimes possible to consistently estimate β, α, and the parameters of the conditional

means from the (nonlinear) regression function

$$E(Y_{it}|d_i, X_{it}, Z_i) = X_{it}\beta + d_i\alpha + d_iE(U_{it}|d_i = 1, Z_i) + (1 - d_i)E(U_{it}|d_i = 0, Z_i). \tag{8}$$

One way to acquire information about the functional form of $E(U_{it}|d_i = 1, Z_i)$ is to assume knowledge of the functional form of the joint distribution of (U_{it}, V_i) (e.g., that it is bivariate normal), but this is not required. Note further that this procedure does not require that Z_i be distributed independently of V_i in equation (2) (Heckman, 1980).

(e) Instead of (d), it is possible to do a two-stage estimation procedure if the joint density of (U_{it}, V_i) is assumed known up to a finite set of parameters and Z_i is distributed independently of V_i and U_{it}. In stage one, $E(U_{it}|d_i = 1, Z_i)$ and $E(U_{it}|d_i = 0, Z_i)$ are determined up to some unknown parameters by conventional discrete choice analysis. Then the regression of equation (8) is run using estimated E values in place of population E values on the right hand side of the equation (Heckman, 1976). In Section X we establish the relationship between this procedure and the propensity score method of Rosenbaum and Rubin (1983).

(f) Under the assumptions of (e), use maximum likelihood to consistently estimate α (Heckman, 1978).

Note that a separate value of α may be estimated for each cross section, so that depending on the number of cross sections it is possible to estimate growth and decay effects in training (i.e., α_t can be estimated for each cross section).

Conventional selection bias approaches (d), (e), and (f) as well as (b) and (c) rely on strong distributional assumptions or else assumptions about nonlinearities in the model, but in fact these are not required. Distributional assumptions are usually not motivated by behavioral theory. Given that a regressor appears in the decision rule of equation (2), if it is uncorrelated with U_{it}, the regressor is an instrumental variable for d_i. It is not necessary to invoke strong distributional assumptions, but if they are invoked, Z_i need not be uncorrelated with U_{it}. In many papers, however, Z_i and U_{it} are usually assumed to be independent. The imposition of overidentifying "information," if false, may lead to considerable bias and instability in the estimates. However, the overidentifying assumptions are testable, and so such false restrictions need not be imposed. Conventional practice imposes these overidentifying restrictions without testing them. We next discuss the instrumental variables procedure in greater detail.[9]

---

[9]Notice that in a transition from a fixed coefficient model to a random coefficient model the analysis of this section focuses on estimation of α^* for the latter. Clearly, U_{it} in equation (7), is redefined to include $d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]$. With this modification, all of our analysis in the text remains intact.

C. The Instrumental Variable Estimator

This estimator is the least demanding in the *a priori* conditions that must be satisfied for its use. It requires the following assumptions:

(a) There is at least one variable in $Z_{it}(Z_i^e)$ with a nonzero γ coefficient in equation (2), such that for some known transformation of Z_i^e, $[g(Z_i^e)]$, $E[U_{it}g(Z_i^e)] = 0$. (9a)

(b) Array X_{it} and d_i into a vector $J_{1it} = (X_{it}, d_i)$. Array X_{it} and $g(Z_i^e)$ into a vector $J_{2it} = [X_{it}, g(Z_i^e)]$. In this notation, it is assumed that

$$E\left(\sum_{i=1}^{I_t} \frac{J_{2it}' J_{1it}}{I_t}\right) \tag{9b}$$

has full column rank uniformly in I_t for I_t sufficiently large where I_t denotes the number of individuals in period t.

With these assumptions, the instrumental variable (IV) estimator

$$\begin{bmatrix}\hat{\beta}\\ \hat{\alpha}\end{bmatrix}_{\text{IV}} = \left(\sum_{i=1}^{I_t} \frac{J_{2it}' J_{1it}}{I_t}\right)^{-1} \sum_{i=1}^{I_t} \frac{J_{2it}' Y_{it}}{I_t}$$

is consistent for (β, α) regardless of any covariance between U_{it} and d_i.

It is important to notice how weak these conditions are. The functional form of the distribution of V_i need not be known. Z_i need not be distributed independently of V_i. Moreover, $g(Z_i^e)$ may be a nonlinear function of variables appearing in X_{it} as long as conditions 9(a) and 9(b) are satisfied.

The instrumental variable $g(Z_i^e)$ may also be a lagged value of time-varying variables appearing in X_{it}, provided the analyst has access to longitudinal data. The rank condition (9b) will generally be satisfied in this case as long as X_{it} exhibits serial dependence. Thus, longitudinal data (on exogenous characteristics) may provide a source of instrumental variables.

D. Identification Through Distributional Assumptions About the Marginal Distribution of U_{it}

If no regressor appears in the decision rule of equation (2), the estimators presented so far in this section cannot be used to consistently estimate α. Heckman (1978) demonstrates that if (U_{it}, V_i) are normally distributed, α is identified even if there is no regressor in equation (2). His conditions are overly strong.

If U_{it} has zero third and fifth central moments, α is identified even if no regressor appears in the enrollment rule. This assumption about U_{it} is implied by normality or symmetry of the density of U_{it}, but it is weaker than either provided that the required moments are finite. The fact that α can be identified by invoking distributional assumptions about U_{it} illustrates the more general point that there is a tradeoff between assumptions about regressors and assumptions about the distribution of U_{it} that must be invoked to identify α.

We have established that under the following assumptions, α in equation (1) is identified:

$$\begin{aligned} &\text{(a)}\quad E(U_{it}^3) = 0 \\ &\text{(b)}\quad E(U_{it}^5) = 0 \\ &\text{(c)}\quad (U_{it}, V_i) \qquad \text{is iid.} \end{aligned} \tag{10}$$

A consistent method-of-moments estimator can be devised that exploits these assumptions (see Heckman and Robb, 1985). Find $\hat{\alpha}$ that sets a weighted average of the sample analogs of $E(U_{it}^3)$ and $E(U_{it}^5)$ as close to zero as possible.

To simplify the exposition, suppose that there are no regressors in the earnings function of equation (1), so $X_{it}\beta = \beta_t$. The proposed estimator finds the value of $\hat{\alpha}$ that sets

$$\frac{1}{I_t}\sum_{i=1}^{I_t}\left[(Y_{it} - \bar{Y}_t) - \bar{\alpha}(d_i - \bar{d})\right]^3 \tag{11a}$$

and

$$\frac{1}{I_t}\sum_{i=1}^{I_t}\left[(Y_{it} - \bar{Y}_t) - \hat{\alpha}(d_i - \bar{d})\right]^5 \tag{11b}$$

as close to zero as possible in a suitably chosen metric where as before, an overbar denotes sample mean. A pair of moments is required in order to pick the unique consistent root. In our companion paper, we establish the existence of a unique root that sets expressions (11a) and (11b) to zero in large samples. Obviously other moment restrictions could be used to identify α.[10]

---

[10] The remark in footnote 9 applies with full force in this section. Different assumptions are being made in the case of estimating α^* than are invoked in the case of estimating the fixed coefficient model, i.e., third and fifth moment assumptions are being invoked about $U_{it} + d_i[\epsilon_i - E(\epsilon_i | d_i = 1)]$ in the former case. The main point—that if *some* moment assumptions are being invoked it is possible to estimate α or α^*—remains intact.

E. Selection on Observables

In the special case in which

$$E(U_{it}|d_i, Z_i) = E(U_{it}|Z_i),$$

selection is said to occur on the observables. Such a case can arise if U_{it} is distributed independently of V_i in equation (2), but U_{it} and Z_i are stochastically dependent (i.e., some of the observables in the enrollment equation are correlated with the unobservables in the earnings equation). In this case U_{it} and d_i can be shown to be conditionally independent given Z_i. If it is further assumed that U_{it} and V_i conditional on Z_i are independent, then U_{it} and d_i can be shown to be conditionally independent given Z_i. In the notation of Dawid (1979) as used by Rosenbaum and Rubin (1983),

$$U_{it} \perp\!\!\!\perp d_i | Z_i,$$

i.e., given Z_i, d_i is strongly ignorable. In a random coefficient model the required condition is

$$(U_{it} + \epsilon_i d_i) \perp\!\!\!\perp d_i | Z_i.$$

The strategy for consistent estimation presented in Section B must be modified; in particular, methods (a)–(c) are inappropriate. However, method (d) still applies and simplifies because

$$E(U_{it}|d_i = 1, Z_i) = E(U_{it}|d_i = 0, Z_i) = E(U_{it}|Z_i),$$

so that we obtain in place of equation (8)

$$E(Y_{it}|d_i, Y_{it}, Z_i) = X_{it}\beta + d_i\alpha + E(U_{it}|Z_i). \tag{8'}$$

Specifying the joint distribution of (U_{it}, Z_i) or just the conditional mean of U_{it} given Z_i, produces a formula for $E(U_{it}|Z_i)$ up to a set of parameters. The model can be estimated by nonlinear regression. Conditions for the existence of a consistent estimator of α are presented in our companion paper (see also Barnow et al., 1980).

Method (e) of Section B no longer directly applies. Except in unusual circumstances (e.g., a single element in Z_i), there is no relationship between any of the parameters of $E(U_{it}|Z_i)$ and the propensity score $\Pr(d_i = 1|Z_i)$, so that conventional two-stage estimators generated from discrete choice theory do not produce useful information. Method (f) produces a consistent estimator provided that an explicit probabilistic relationship between U_{it} and Z_i is postulated.[11]

[11] The remarks made in footnotes 9 and 10 apply in this section as well.

F. Summary

Conventional cross-section practice invokes numerous extraneous assumptions to secure identification of α. These overidentifying restrictions are rarely tested, although they are testable. Strong distributional assumptions are not required to estimate α. Assumptions about the distributions of unobservables are rarely justified by an appeal to behavioral theory. Assumptions about the presence of regressors in enrollment equations and assumptions about stochastic dependence relationships among U_{it}, Z_i, and d_i are sometimes justified by behavioral theory.

V. Repeated Cross-Section Methods for the Case When Training Identity of Individuals Is Unknown

In a time homogeneous environment, estimates of the population mean earnings formed in two or more cross sections of unrelated persons can be used to obtain selection-bias-free estimates of the training effect even if the training status of each person is unknown (but the population proportion of trainees is known or can be consistently estimated). With more data, the time homogeneity assumption can be partially relaxed.

Assuming a time homogeneous environment and access to repeated cross-section data and random sampling, it is possible to identify α (a) without any regressor in the decision rule, (b) without need to specify the joint distribution of U_{it} and V_i, and (c) without any need to know which individuals in the sample enrolled in training (but the proportion of trainees must be known or consistently estimable).

To see why this claim is true, suppose that no regressors appear in the earnings function.[12] In the notation of equation (1), $X_{it}\beta = \beta_t$. Then, assuming that a random sampling scheme generates the data,

$$\operatorname{plim} \overline{Y}_t = \operatorname{plim} \frac{\sum Y_{it}}{I_t} = E(\beta_t + \alpha d_i + U_{it}) = \beta_t + \alpha p \qquad t > k$$

$$\operatorname{plim} \overline{Y}_{t'} = \operatorname{plim} \frac{\sum Y_{it'}}{I_{t'}} = E(\beta_{t'} + U_{it'}) = \beta_{t'} \qquad t' < k.$$

[12] If regressors appear in the earnings functions, the following procedure can be used. Rewrite equation (1) as $Y_{it} = \beta_t + X_{it}\pi + d_i\alpha + U_{it}$. It is possible to estimate π from preprogram data. (This assumes there are no time invariant variables in X_{it}. If there are such variables, they may be deleted from the regressor vector and π appropriately redefined without affecting the analysis.) Replace Y_{it} by $Y_{it} - X_{it}\hat{\pi}$ and the analysis in the text goes through. Note that we are assuming that no X_{it} variables become nonconstant after period k.

In a time homogeneous environment, $\beta_t = \beta_{t'}$ and

$$\operatorname{plim} \frac{\overline{Y}_t - \overline{Y}_{t'}}{\hat{p}} = \alpha,$$

where $\hat{p}$ is a consistent estimator of $p = E(d_i)$.

With more than two years of repeated cross-section data, one can apply the same principles to identify α while relaxing the time homogeneity assumption. For instance, suppose that population mean earnings lie on a polynomial of order $L-2$:

$$\beta_t = \pi_0 + \pi_1 t + \cdots + \pi_{L-2} t^{L-2}.$$

From L temporally distinct cross sections, it is possible to estimate consistently the $L-1$ π-parameters and α provided that the number of observations in each cross section becomes large and there is at least one preprogram and one postprogram cross section.

If the effect of training differs across periods, it is still possible to identify α_t provided that the environment changes in a "sufficiently regular" way. For example, suppose

$$\beta_t = \pi_0 + \pi_1 t$$

$$\alpha_t = \phi_0 (\phi_1)^{t-k} \qquad t > k.$$

In this case, π_0, π_1, ϕ_0, and ϕ_1 are identified from the means of four cross sections, as long as at least two of these means come from a preprogram period and two come from successive postprogram periods.

In our companion paper we state more rigorously the conditions required to consistently estimate α using repeated cross-section data which do not record the training identity of individuals. Section IX examines the sensitivity of this class of estimators to violations of the random sampling assumption.

VI. Longitudinal Procedures

Most longitudinal procedures require knowledge of certain moments of the joint distribution of unobservables in the earnings and enrollment equations. We present several illustrations of this claim, as well as a counterexample. The counterexample identifies α by assuming only that the error term in the earnings equation is covariance stationary.

We now consider four examples of estimators which use longitudinal data.

A. The Fixed Effects Method

This method was developed by Mundlak (1961, 1978) and refined by Chamberlain (1982). It is widely used in recent social science data analyses.

It is based on the following assumption:

$$E(U_{it} - U_{it'} | d_i, X_{it} - X_{it'}) = 0 \quad \text{for all } t, t', t > k > t'. \tag{12}$$

As a consequence of this assumption, we may write a difference regression as

$$E(Y_{it} - Y_{it'} | X_{it} - X_{it'}, d_i) = (X_{it} - X_{it'})\beta + d_i\alpha \qquad t > k > t'.$$

Suppose that condition (12) holds and the analyst has access to one year of preprogram and one year of postprogram earnings. Regressing the difference between postprogram earnings in any year and earnings in any preprogram year on the change in regressors between those years and a dummy variable for training status produces a consistent estimator of α.

Some decision rules and error processes for earnings produce condition (12). For example, consider a certainty environment in which the earnings residual has a permanent-transitory structure:

$$U_{it} = \phi_i + \epsilon_{it} \tag{13}$$

where ϵ_{it} is a mean zero random variable independent of all other values of $\epsilon_{it'}$ for $t \neq t'$ and is distributed independently of ϕ_i, a mean zero person-specific time-invariant random variable. Assuming that S_i in the decision rule of equation (5) is distributed independently of all ϵ_{it} except possibly for ϵ_{ik}, then condition (12) will be satisfied. With two periods of data (in t and t', $t > k > t'$) α is just identified. With more periods of panel data, the model is overidentified and hence condition (12) is subject to test (Chamberlain, 1982).

Condition (12) may also be satisfied in an environment of uncertainty. Suppose equation (13) governs the error structure in equation (1) and

$$E_{k-1}(\epsilon_{ik}) = 0,$$

but

$$E_{k-1}(\phi_i) = \phi_i$$

so that agents cannot forecast innovations in their earnings but they know their own permanent component. Provided that S_i is distributed independently of all ϵ_{it} except possibly for ϵ_{ik}, this model also produces condition (12).

We investigate the plausibility of condition (12) with respect to more general decision rules and error processes in Section VIII.[13]

B. U_{it} Follows a First-Order Autoregressive Process

Suppose next that U_{it} follows a first-order autoregression:

$$U_{it} = \rho U_{i,t-1} + \nu_{it} \tag{14}$$

---

[13] We repeat the point made in footnotes 9, 10, and 11 that if α^* is the coefficient of interest, U_{it} is redefined to be $U_{it} + d_i[\epsilon_i - E(\epsilon_i | d_i = 1)]$.

where $E(\nu_{it}) = 0$ and the ν_{it} are mutually independently (not necessarily identically) distributed random variables with $\rho \neq 1$. Substitution using equations (1) and (14) to solve for U_{it} yields

$$Y_{it} = \left[X_{it} - \left(X_{it'}\rho^{t-t'}\right)\right]\beta + \left(1 - \rho^{t-t'}\right)d_i\alpha + \rho^{t-t'}Y_{it'} + \left\{\sum_{j=0}^{t-(t'+1)} \rho^j \nu_{i,t-j}\right\} \qquad t > t' > k. \tag{15}$$

Assume further that the perfect foresight rule of equation (5) determines enrollment and that the ν_{ij} are distributed independently of S_i and X_{ik} in equation (5). Heckman and Wolpin (1976) invoke similar assumptions in their analysis of affirmative action programs. If X_{ij} is independent of $\nu_{ij'}$ for all j, j' (an overly strong condition) then (linear or nonlinear) least squares applied to equation (15) consistently estimates α as the number of observations becomes large. (The appropriate nonlinear regression increases efficiency by imposing the implied cross coefficient restrictions.) As is the case with the fixed effect estimator, increasing the length of the panel and keeping the same assumptions, the model becomes overidentified (and hence testable) for panels with more than two observations.[14]

C. U_{it} Is Covariance Stationary

The next procedure invokes an assumption implicitly used in many papers on training (e.g., Ashenfelter, 1978; Bassi, 1983), but exploits the assumption in a novel way. We assume

(a) U_{it} is covariance stationary so $E(U_{it}U_{i,t-j}) = \sigma_j$ for $j \geq 0$;
(b) access to at least two observations on preprogram earnings in t' and $t' - j$ as well as one observation on postprogram earnings in t where $t - t' = j$; and (16)
(c) $pE(U_{it'}|d_i = 1) \neq 0$.

Unlike the two previous examples, we make no assumptions here about the

[14] In the context of estimating α^* in the random coefficient model, it is not natural to specify equation (14) in the text for the redefined U_{it}. In general if U_{it} has an autoregressive representation, $U_{it} + d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]$ will not. A more natural specification models error component $d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]$ as a permanent postprogram component in the error term. In place of the error term in braces in Equation (15), write

$$\sum_{j=0}^{t-(t'+1)} \rho^j \nu_{i,t-j} + \phi_i\left(1 - \rho^{t-t'}\right) \qquad t, t' > k$$

where $\phi_i = d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]$. Orthogonality conditions will *not* be satisfied between ϕ_i and $Y_{it'}$, and an instrument for lagged $Y_{it'}$ will be required to consistently estimate α^* or else the time series methods of MaCurdy (1982) will have to be invoked to obtain consistent estimators.

appropriate enrollment rule or about the stochastic relationship between U_{it} and the cost of enrollment S_i.

By the argument of footnote 12, we lose no generality by suppressing the effect of regressors in equation (1). Thus let

$$Y_{it} = \beta_t + d_i\alpha + U_{it} \qquad t > k$$
$$Y_{it'} = \beta_{t'} + U_{it'} \qquad t' < k,$$

where β_t and $\beta_{t'}$ are period-specific shifters.

From a random sample of preprogram earnings from periods t' and $t'-j$, σ_j can be consistently estimated from the sample covariances between $Y_{it'}$ and $Y_{i,t'-j}$:

$$m_1 = \frac{\Sigma\left[(Y_{it'} - \overline{Y}_{t'})(Y_{i,t'-j} - \overline{Y}_{t'-j})\right]}{I_t}$$

$$\operatorname{plim} m_1 = \sigma_j.$$

If $t > k$ and $t - t' = j$ so that the postprogram earnings data are as far removed in time from t' as t' is removed from $t'-j$, form the sample covariance between Y_{it} and $Y_{it'}$:

$$m_2 = \frac{\Sigma\left[(Y_{it} - \overline{Y}_t)(Y_{it'} - \overline{Y}_{t'})\right]}{I_t},$$

which has the probability limit

$$\operatorname{plim} m_2 = \sigma_j + \alpha p E(U_{it'} | d_i = 1) \qquad t > k > t'.$$

From the sample covariance between d_i and $Y_{it'}$,

$$m_3 = \frac{\Sigma\left[(Y_{it'} - \overline{Y}_{t'})d_i\right]}{I_t}$$

with probability limit

$$\operatorname{plim} m_3 = pE(U_{it'} | d_i = 1) \qquad t' < k.$$

Combining this information and assuming $pE(U_{it'} | d_i = 1) \neq 0$ for $t' < k$,

$$\operatorname{plim} \hat{\alpha} = \operatorname{plim} \frac{m_2 - m_1}{m_3} = \alpha.$$

For panels of sufficient length (e.g., more than two preprogram observations or more than two postprogram observations), the stationarity assumption can be tested. Thus as before, increasing the length of the panel converts a just identified model to an overidentified one.[15]

---

[15]As in footnotes 9, 10, and 11, we emphasize that different assumptions are being made in the random coefficient version of the model than are made in the fixed coefficient version. Note, however, that in this section we do not require that variances be equal in preprogram and postprogram periods so that the estimator $\hat{\alpha}$ is still appropriate as an estimator for α^* if, e.g., U_{it} is uncorrelated with $d_i[\epsilon_i - E(\epsilon_i | d_i = 1)]$ for all t.

D. An Unrestricted Process for U_{it} When Agents Do Not Know Future Innovations in Their Earnings

The estimator proposed in this subsection assumes that agents cannot perfectly predict future earnings. More specifically, for an agent whose relevant earnings history begins N periods before period k, we assume that

(a) $$E_{k-1}(U_{ik}) = E(U_{ik}|U_{i,k-1},\ldots,U_{i,k-N}),$$

i.e., that predictions of future U_{it} are made solely on the basis of previous values of U_{it}. Past values of the exogenous variables are assumed to have no predictive value for U_{ik}.

In addition, we assume that

(b) the relevant earnings history goes back N periods before period k;
(c) the enrollment decision is characterized by equation (4);
(d) S_i and X_{ik} are known as of period $k-1$ when the enrollment decision is being made;
(e) X_{it} is distributed independently of U_{ij} for all t and j; and
(f) S_i is distributed independently of U_{ij} for all j.

Defining

$$\psi_i = (Y_{i,k-1} - X_{i,k-1}\beta,\ldots,Y_{i,k-N} - X_{i,k-N}\beta)$$

and

$$G(\psi_i) = E(d_i|\psi_i),$$

under the conditions given above, α can be consistently estimated. We define

$$p = E(d_i),$$

and

$$c = \frac{E[U_{it}(G(\psi_i) - p)]}{E(G(\psi_i) - p)^2}.$$

We rewrite equation (1) in the following way:

$$Y_{it} = X_{it}\beta + d_i\alpha + c(G(\psi_i) - p) + [U_{it} - c(G(\psi_i) - p)]. \qquad (17)$$

This defines an estimating equation for the parameters of the model. In the transformed equation

$$E\{X_{it}'[U_{it} - c(G(\psi_i) - p)]\} = 0$$

by assumption (e) above. The transformed residual is uncorrelated with $c(G(\psi_i) - p)$ from the definition of c.

Thus, it remains to show that

$$E\{d_i[U_{it} - c(G(\psi_i) - p)]\} = 0.$$

Before proving this it is helpful to notice that as a consequence of assumptions (a), (d), and (e),

$$\begin{aligned} &E(d_i|U_{it}, U_{i,t-1}, \ldots, U_{i,k-1}, \ldots, U_{i,k-N}) \\ &\quad = E(d_i|U_{i,k-1}, \ldots, U_{i,k-N}) \qquad t > k. \end{aligned} \tag{18}$$

This relationship is proved in our companion paper. Since only preprogram innovations determine participation and because U_{it} is distributed independently of X_{ik} and S_i in the decision rule of equation (4), the conditional mean of d_i does not depend on postprogram values of U_{it} given all preprogram values.

Intuitively, the term $U_{it} - c(G(\psi_i) - p)$ is orthogonal to $G(\psi_i)$, the best predictor of d_i based on ψ_i; if $U_{it} - c(G(\psi_i) - p)$ were correlated with d_i, it would mean that U_{it} helped to predict d_i, contradicting condition (18).

The proof of the proposition uses the fact from condition (18) that $E(d_i|\psi_i, U_{it}) = G(\psi_i)$ in computing the expectation

$$\begin{aligned} &E\{d_i[U_{it} - c(G(\psi_i) - p)]\} \\ &\quad = E[E\{d_i[U_{it} - c(G(\psi_i) - p)]\}|\psi_i, U_{it}] \\ &\quad = E\{[U_{it} - c(G(\psi_i) - p)]E(d_i|\psi_i, U_{it})\} \\ &\quad = E\{[U_{it} - c(G(\psi_i) - p)]G(\psi_i)\} \\ &\quad = 0 \end{aligned}$$

as a consequence of the definition of c.

The elements of ψ_i can be consistently estimated by fitting a preprogram earnings equation and forming the residuals from preprogram earnings data to estimate $U_{i,k-1}, \ldots, U_{k,k-N}$. One can assume a functional form for G and estimate the parameters of G using standard methods in discrete choice applied to enrollment data.[16]

VII. Repeated Cross-Section Analogs of Longitudinal Procedures

Most longitudinal procedures can be fit on repeated cross-section data. Repeated cross-section data are cheaper to collect, and they do not suffer from problems of nonrandom attrition which plague panel data.

The previous section presented longitudinal estimators of α. In all cases but one, however, α can actually be identified with repeated cross-section data. Here we establish this claim. Our earlier paper gives additional

[16] In the context of estimating α^* the estimator of this section requires that predictions of future $U_{it} + d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]$ are based solely on preprogram values of U_{it} $(t < k)$.

examples of longitudinal estimators which can be implemented on repeated cross-section data. We have been unable to produce a repeated cross-section estimator of the method given in Section VID.

A. The Fixed Effect Model

As in Section VIA, assume that condition (12) holds so

$$E(U_{it}|d_i=1)=E(U_{it'}|d_i=1)$$
$$E(U_{it}|d_i=0)=E(U_{it'}|d_i=0) \qquad t>k>t'$$

for all t, t'. Let $X_{it}\beta=\beta_t$ and define $\hat{\alpha}$ in terms of the notation of Section IVA

$$\hat{\alpha}=\left(\overline{Y}_t^{(1)}-\overline{Y}_t^{(0)}\right)-\left(\overline{Y}_{t'}^{(1)}-\overline{Y}_{t'}^{(0)}\right).$$

Assuming random sampling, consistency of $\hat{\alpha}$ follows immediately from condition (12):

$$\begin{aligned}\operatorname{plim}\alpha &= \left[\alpha+\beta_t-\beta_{t'}+E(U_{it}|d_i=1)-E(U_{it}|d_i=0)\right]\\ &\quad-\left[\beta_t-\beta_{t'}+E(U_{it'}|d_i=1)-E(U_{it'}|d_i=0)\right]\\ &=\alpha.\end{aligned}$$

As in the case of the longitudinal version of this estimator, with more than two cross sections, the hypothesis of condition (12) is subject to test (i.e., the model is overidentified).

B. U_{it} Follows a First-Order Autoregressive Process

In one respect the preceding example is contrived. It assumes that in preprogram cross sections we know the identity of future trainees. Such data might exist (e.g., individuals in the training period k might be asked about their preperiod k earnings to see if they qualify for admission), but this seems unlikely. One advantage of longitudinal data for estimating α in the fixed effect model is that if the survey extends before period k, the identity of future trainees is known.

The need for preprogram earnings to identify α is, however, only an artifact of the fixed-effect assumption of equation (13). Suppose instead that U_{it} follows a first-order autoregressive process given by equation (14) and that

$$E(v_{it}|d_i)=0 \qquad t>k \tag{19}$$

as in Section VIB. With three successive postprogram cross sections in which the identity of trainees is known, it is possible to identify α.

To establish this result, let the three postprogram periods be t, $t+1$ and $t+2$. Assuming, as before, that no regressor appears in equation (1),

$$\operatorname{plim} \overline{Y}_j^{(1)} = \beta_j + \alpha + E(U_{ij}|d_i = 1)$$

$$\operatorname{plim} \overline{Y}_j^{(0)} = \beta_j + E(U_{ij}|d_i = 0).$$

From condition (19),

$$E(U_{i,t+1}|d_i = 1) = \rho E(U_{it}|d_i = 1)$$

$$E(U_{i,t+1}|d_i = 0) = \rho E(U_{it}|d_i = 0)$$

$$E(U_{i,t+2}|d_i = 1) = \rho^2 E(U_{it}|d_i = 1)$$

$$E(U_{i,t+2}|d_i = 0) = \rho^2 E(U_{it}|d_i = 0).$$

Using these formulae, it is straightforward to verify that $\hat{\rho}$ defined by

$$\hat{\rho} = \frac{\left(\overline{Y}_{t+2}^{(1)} - \overline{Y}_{t+2}^{(0)}\right) - \left(\overline{Y}_{t+1}^{(1)} - \overline{Y}_{t+1}^{(0)}\right)}{\left(\overline{Y}_{t+1}^{(1)} - \overline{Y}_{t+1}^{(0)}\right) - \left(\overline{Y}_{t}^{(1)} - \overline{Y}_{t}^{(0)}\right)}$$

is consistent for ρ and that $\hat{\alpha}$ defined by

$$\hat{\alpha} = \frac{\left(\overline{Y}_{t+2}^{(1)} - \overline{Y}_{t+2}^{(0)}\right) - \hat{\rho}\left(\overline{Y}_{t+1}^{(1)} - \overline{Y}_{t+1}^{(0)}\right)}{1 - \hat{\rho}}$$

is consistent for α.[17]

For this model, the advantage of longitudinal data is clear. Only two time periods of longitudinal data are required to identify α, but three periods of repeated cross-section data are required to estimate the same parameter. However, if Y_{it} is subject to measurement error, the apparent advantages of longitudinal data become less clear. Repeated cross-section estimators are robust to mean zero measurement error in the variables. The longitudinal regression estimator discussed in Section VIB does not identify α unless the analyst observes earnings without error. Given three years of longitudinal data and assuming that measurement error is serially uncorrelated, one could instrument Y_{it}, in equation (15), using earnings in the earliest year as an instrument. This requires one more year of data. Thus one advantage of the longitudinal estimator disappears in the presence of measurement error.[18] With four or more repeated cross sections, the model is obviously overidentified and hence subject to test.

---

[17] This estimator is obviously consistent for either the fixed coefficient (α) or random coefficient (α^*) model since $E\{d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]|d_i = 1\} = 0$.

[18] Recall from our discussion in footnote 14 that in the random coefficient model developed there an instrument for $Y_{it'}$ is required even in the absence of measurement error.

C. Covariance Stationarity

For simplicity we assume that there are no regressors in the earnings equation and let $X_{it}\beta = \beta_i$ (see Heckman and Robb, 1985, for the case in which regressors are present). Assume that conditions (16) are satisfied. Before presenting the repeated cross section estimator, it is helpful to record the following facts:

$$\text{Var}(Y_{it}) = \alpha^2(1-p)p + 2\alpha E(U_{it}|d_i = 1)p + \sigma_u^2 \qquad t > k \quad (20a)$$

$$\text{Var}(Y_{it'}) = \sigma_u^2 \qquad t' < k \quad (20b)$$

$$\text{Cov}(Y_{it}, d_i) = \alpha p(1-p) + pE(U_{it}|d_i = 1). \quad (20c)$$

Note that $E(U_{it}^2) = E(U_{it'}^2)$ by virtue of assumption (16a). Then

$$\hat{\alpha} = [p(1-p)]^{-1}\left\{\left[\frac{\Sigma(Y_{it} - \bar{Y}_t)d_i}{I_t}\right.\right.$$
$$-\left\{\left[\frac{\Sigma(Y_{it} - \bar{Y}_t)d_i}{I_t}\right]^2\right. \quad (21)$$
$$\left.\left.\left. -p(1-p)\left[\frac{\Sigma(Y_{it} - \bar{Y}_t)^2}{I_t} - \frac{\Sigma(Y_{it'} - \bar{Y}_{t'})^2}{I_{t'}}\right]\right\}^{1/2}\right]\right\}$$

is consistent for α.

This expression arises by subtracting equation (20b) from (20a). Then use equation (20c) to get an expression for $E(U_{it}|d_i = 1)$ which can be substituted into the expression for the difference between equation (20a) and (20b). Replacing population moments by sample counterparts produces a quadratic equation in $\hat{\alpha}$, with the negative root given by equation (21). The positive root is inconsistent for α.[19]

Notice that the estimators of Sections VIC and VIIC exploit different features of the covariance stationarity assumptions. The longitudinal procedure only requires that $E(U_{it}U_{i,t-j}) = E(U_{it'}U_{i,t'-j})$ for $j > 0$; variances need not be equal across periods. The repeated cross section analog above only requires that $E(U_{it}U_{i,t-j}) = E(U_{it'}U_{i,t'-j})$ for $j = 0$; covariances may

[19] This estimator requires that the variance of U_{it} $(t > k)$ be the same as the variance of $U_{it'}$ $(t' < k)$. Thus, in the random coefficient model, if U_{it} has a constant variance, $U_{it} + d_i[\epsilon_i - E(\epsilon_i|d_i = 1)]$ will not have the same variance as $U_{it'}$. It is possible, but artificial, to invoke equality of the variances for the two disturbance terms. Thus, in this sense our proposed covariance stationary estimator is *not* robust when applied to estimate α^* in repeated cross-section data.

differ among equispaced pairs of the U_{it}. With more than two cross sections, the covariance stationarity assumption is overidentifying and hence subject to test.

VIII. First Difference or Fixed Effect Methods

Plausible economic models do not justify first difference methods. Lessons drawn from these methods are misleading.

A. Models which Justify Condition (12)

Whenever condition (12) holds, α can be estimated consistently from the difference regression method described in Section VIA. This section presents a model which satisfies condition (12): the earnings residual has a permanent-transitory structure, the decision rules of equations (4) or (5) determine enrollment, and S_i is distributed independently of the transitory component of U_{it}.

However, this model is rather special. It is very easy to produce plausible models that do not satisfy condition (12). For example, even if equation (13) characterizes U_{it}, if S_i in equation (5) does not have the same joint (bivariate) distribution with respect to all ϵ_{it}, except for ϵ_{ik}, condition (12) may be violated.

Even if S_i in equation (5) is distributed independently of U_{it} for all t, it is still not the case that condition (12) is satisfied in a general model. For example, suppose X_{it} is distributed independently of all U_{it} and let

$$U_{it} = \rho U_{i,t-1} + \nu_{it},$$

where ν_{it} is a mean zero, iid random variable and $|\rho| < 1$. If $\rho \neq 0$ and the perfect foresight decision rule characterizes enrollment, condition (12) is not satisfied for $t > k > t'$ because

$$\begin{aligned} E(U_{it}|d_i = 1) &= E(U_{it}|U_{ik} + X_{ik}\beta - \alpha/r < S_i) \\ &= \rho^{t-k}E(U_{ik}|d_i = 1) \\ &\neq E(U_{it'}|d_i = 1) = E(U_{it'}|U_{ik} + X_{ik}\beta - \alpha/r < S_i) \end{aligned}$$

unless the conditional expectations are linear (in U_{ik}) for all t and $k - t' = t - k$. In that case

$$E(U_{it'}|d_i = 1) = \rho^{k-t'}E(U_{ik}|d_i = 1),$$

so $E(U_{it} - U_{it'}|d_i = 1) = 0$ only for t, t' such that $k - t' = t - k$. Thus, condition (12) is not satisfied for all $t > k > t'$.

For more general specifications of U_{it} and stochastic dependence between S_i and U_{it}, condition (12) will not be satisfied.

B. More General First Difference Estimators

Instead of condition (12), assume that

$$\begin{aligned} E[(U_{it}-U_{it'})(X_{it}-X_{it'})] &= 0 \quad \text{for some} \quad t, t', t>k>t' \\ E[(U_{it}-U_{it'})d_i] &= 0 \quad \text{for some} \quad t>k>t'. \end{aligned} \tag{22}$$

Two new ideas are embodied in this assumption. In place of the assumption that $U_{it}-U_{it'}$ be conditionally independent of $X_{it}-X_{it'}$ and d_i, we only require uncorrelatedness. Also, rather than assume that $E(U_{it}-U_{it'}|d_i, X_{it}-X_{it'})=0$ for all $t>k>t'$, the correlation needs to be zero only for some $t>k>t'$. For the appropriate values of t and t', least squares applied to the differenced data consistently estimates α.

Our companion paper presents three examples of models that satisfy condition (22) but not (12). Here we discuss one of them.

Suppose that

(a) U_{it} is covariance stationary;
(b) U_{it} has a linear regression on U_{ik} for all t (i.e., $E(U_{it}|U_{ik})=\beta_{tk}U_{ik}$);
(c) the U_{it} are mutually independent of (X_{ik}, S_i) for all t;
(d) α is common to all individuals (so the model is of the fixed coefficient form); and
(e) the environment is one of perfect foresight where the decision rule of equation (5) determines participation.

(23)

Under these assumptions, condition (22) is satisfied.

To see this, note that (a) and (b) above imply that there exists a δ such that

$$\begin{aligned} U_{it} = U_{i,k+j} &= \delta U_{ik} + \omega_{it} \qquad j>0,\ t=k+j \\ U_{it'} = U_{i,k-j} &= \delta U_{ik} + \omega_{it'} \qquad j>0,\ t'=k-j \end{aligned}$$

for *some* $j>0$, and

$$E(\omega_{it}|U_{ik}) = E(\omega_{it'}|U_{ik}) = 0.$$

Now observe that

$$E(U_{it}|d_i=1) = \delta E(U_{ik}|d_i=1) + E(\omega_{it}|d_i=1).$$

But, as a consequence of assumption (c) above,

$$E(\omega_{it}|d_i=1)=0$$

since $E(\omega_{it})=0$ and because (c) guarantees that the mean of ω_{it} does not depend on X_{ik} and S_i. Similarly,

$$E(\omega_{it'}|d_i=1)=0$$

and thus condition (22) holds.

Linearity of the regression does not imply that the U_{it} are normally distributed (although if the U_{it} are joint normal, the regression is linear). The multivariate t density is just one example of many examples of densities with linear regressions.[20]

C. Anomalous Features of First Difference Estimators

Nearly all of the estimators require a control group (i.e., a sample of nontrainees). The only exception is the fixed effect estimator in a time homogeneous environment. In this case, if conditions (12) or (22) hold, if we let $X_{it}\beta = \beta_t$ to simplify the exposition, and if the environment is time homogeneous so $\beta_t = \beta_{t'}$, then

$$\hat{\alpha} = \overline{Y}_t^{(1)} - \overline{Y}_{t'}^{(0)}$$

consistently estimates α. The frequently stated claim that "if the environment is stationary, you don't need a control group" (see, e.g., Bassi, 1983) is false except for the special conditions which justify use of the fixed effect estimator.

Most of the procedures considered here can be implemented using only postprogram data. The covariance stationarity estimators of Sections VIC and VIIC, certain repeated cross-section estimators, and first difference methods constitute an exception to this rule. In this sense, those estimators are anomalous.

Fixed effect estimators are also robust to departures from the random sampling assumption. For instance, suppose conditions (12) or (22) are satisfied, but that the available data oversample or undersample trainees [i.e., the proportion of trainees in the sample does not converge to $p = E(d_i)$]. Suppose further that the analyst does not know the true value of p. Nevertheless, a first difference regression continues to identify α. Most other procedures do not share this property.

IX. Nonrandom Sampling Plans

Virtually all methods can be readily adjusted to account for choice-based sampling or measurement error in training status. Some methods require no modification at all.

The data available for analyzing the impact of training on earnings are often nonrandom samples. Frequently they consist of pooled data from two

[20] For reasons already discussed in footnote 14, the estimator proposed in this section is less attractive (and requires redefinition) in the context of estimating α^* in a random coefficient model.

sources: (a) a sample of trainees selected from program records and (b) a sample of nontrainees selected from some national sample. Typically, such samples overrepresent trainees relative to their proportion in the population. This creates the problem of choice-based sampling analyzed by Manski and Lerman (1977) and Manski and McFadden (1981).

A second problem, contamination bias, arises when the training status of certain individuals is recorded with error. Many control samples such as the Current Population Survey or Social Security Work History do not reveal whether or not persons have received training.

Both of these sampling situations combine the following types of data:

(A) earnings, earnings characteristics, and enrollment characteristics for a sample of trainees ($d_i = 1$);
(B) earnings, earnings characteristics, and enrollment characteristics for a sample of nontrainees ($d_i = 0$); and
(C) earnings, earnings characteristics, and enrollment characteristics for a national "control" sample of the population (e.g., CPS or Social Security records) where the training status of persons is not known.

If type (A) and (B) data are combined and the sample proportion of trainees does not converge to the population proportion of trainees, the combined sample is a choice-based sample. If type (A) and (C) data are combined with or without type (B) data, there is contamination bias because the training status of some persons is not known.

Most procedures developed in the context of random sampling can be modified to consistently estimate α using choice-based samples or contaminated control groups (i.e., groups in which training status is not known for individuals). In some cases, a consistent estimator of the population proportion of trainees is required. We illustrate these claims by showing how to modify the instrumental variables estimator to address both sampling schemes. Our companion paper gives explicit case by case treatment of these issues for each estimator developed there.

A. The Instrumental Variable (IV) Estimator: Choice-Based Sampling

If condition (9a) is strengthened to read

$$\begin{aligned} E(X_{it}' U_{it} | d_i) &= 0 \\ E(g(Z_i^c) U_{it} | d_i) &= 0 \end{aligned} \tag{24}$$

and a modified condition (9b) is also met, the IV estimator is consistent for α in choice-based samples.

To see why this is so, write the normal equations for the IV estimator in the following form:

$$\begin{bmatrix} \frac{\Sigma X_{it}' X_{it}}{I_t} & \frac{\Sigma X_{it}' d_i}{I_t} \\ \frac{\Sigma g(Z_i^e) X_{it}}{I_t} & \frac{\Sigma g(Z_i^e) d_i}{I_t} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\Sigma X_{it}' Y_{it}}{I_t} \\ \frac{\Sigma g(Z_i^e) Y_{it}}{I_t} \end{bmatrix} \tag{25}$$

$$= \begin{bmatrix} \frac{\Sigma X_{it}' X_{it}}{I_t} & \frac{\Sigma X_{it}' d_i}{I_t} \\ \frac{\Sigma g(Z_i^e) X_{it}}{I_t} & \frac{\Sigma g(Z_i^e) d_i}{I_t} \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \begin{bmatrix} \frac{\Sigma X_{it}' U_{it}}{I_t} \\ \frac{\Sigma g(Z_i^e) U_{it}}{I_t} \end{bmatrix}.$$

Since condition (24) guarantees that

$$\operatorname*{plim}_{I_t \to \infty} \frac{\Sigma X_{it}' U_{it}}{I_t} = 0 \qquad \text{and}$$

$$\operatorname*{plim}_{I_t \to \infty} \frac{\Sigma g(Z_i^e) U_{it}}{I_t} = 0 \tag{26}$$

and modified rank condition (9b) holds, the IV estimator is consistent.

In a choice-based sample, let the probability that an individual has enrolled in training be p^*. Even if conditions (9a) and (9b) are satisfied, there is no guarantee that condition (26) will be met without invoking condition (24). This is so because

$$\operatorname*{plim}_{I_t \to \infty} \frac{\Sigma X_{it}' U_{it}}{I_t} = E(X_{it}' U_{it} | d_i = 1) p^* + E(X_{it}' U_{it} | d_i = 0)(1 - p^*)$$

$$\operatorname*{plim}_{I_t \to \infty} \frac{\Sigma g(Z_i^e) U_{it}}{I_t} = E(g(Z_i^e) U_{it} | d_i = 1) p^* + E(g(Z_i^e) U_{it} | d_i = 0)(1 - p^*).$$

These expressions are not generally zero, so the IV estimator is generally inconsistent.

In the case of random sampling, $p^* = \Pr[d_i = 1] = p$ and the above expressions are identically zero. They are also zero if condition (24) is satisfied. However, it is not necessary to invoke condition (24). Provided p is known, it is possible to reweight the data to secure consistent estimators under the assumptions of Section IV. Multiplying equation (1) by the

weight

$$\omega_i = d_i \frac{p}{p^*} + (1 - d_i)\left(\frac{1-p}{1-p^*}\right)$$

and applying IV to the transformed equation produces an estimator that satisfies condition (24). It is straightforward to check that weighting the sample at hand back to random sample proportions causes the IV method to consistently estimate α and β (see Heckman and Robb, 1985).

B. The IV Estimator: Contamination Bias

For data of type (C) (see beginning of Section IX), d_i is not observed. Applying the IV estimator to pooled samples (A) and (C) assuming that observations in (C) have $d_i = 0$ produces an inconsistent estimator. However, it is possible to construct a consistent estimator for this case.

In terms of the IV equations (25), from sample (C) it is possible to generate the cross products

$$\frac{\Sigma X_{it}' X_{it}}{I_c}, \quad \frac{\Sigma g(Z_i^e) X_{it}}{I_c}, \quad \frac{\Sigma X_{it}' Y_{it}}{I_c}, \quad \frac{\Sigma g(Z_i^e) Y_{it}}{I_c},$$

which converge to the desired population counterparts where I_c denotes the number of observations in sample (C). Missing is information on the cross products

$$\frac{\Sigma X_{it}' d_i}{I_c}, \quad \frac{\Sigma g(Z_i^e) d_i}{I_c}.$$

Notice that if d_i were measured accurately in sample (C),

$$\operatorname*{plim}_{I_c \to \infty} \frac{\Sigma X_{it}' d_i}{I_c} = pE[X_{it}' | d_i = 1]$$

$$\operatorname*{plim}_{I_c \to \infty} \frac{\Sigma g(Z_i^e) d_i}{I_c} = pE[g(Z_i^e) | d_i = 1].$$

But the means of X_{it} and $g(Z_i^e)$ in sample (A) converge to

$$E(X_{it}' | d_i = 1) \quad \text{and} \quad E(g(Z_i^e) | d_i = 1),$$

respectively. Hence, inserting the sample (A) means of X_{it} and $g(Z_i^e)$ multiplied by p in the second column of the matrix IV equations (25) produce a consistent IV estimator provided that in the limit the size of samples (A) and (C) both approach infinity.

C. Repeated Cross-Section Methods with Unknown Training Status and Choice-Based Sampling

The repeated cross-section estimators discussed in Section V are inconsistent when applied to choice-based samples unless additional conditions are assumed. For example, when the environment is time homogeneous and condition (12) also holds, $(\bar{Y}_t - \bar{Y}_{t'})/p$ remains a consistent estimator of α in choice-based samples as long as the same proportion of trainees are sampled in periods t' and t. If a condition such as condition (12) is not met, it is necessary to know the identity of trainees to weight the sample back to the proportion of trainees that would be produced by a random sample to obtain consistent estimators. Hence, the class of estimators that does not require knowledge of individual training status is not robust to choice-based sampling.

D. Control Function Estimators

A subset of cross-sectional and longitudinal procedures is robust to choice-based sampling. Those procedures construct a control function K_{it} with the following properties:

$$\begin{aligned} &K_{it} \text{ depends on variables } \ldots, Y_{i,t+1}, Y_{it}, Y_{i,t-1}, \ldots, X_{i,t+1}, X_{it}, \\ &X_{i,t-1}, \ldots, d_i \text{ and parameters } \psi \text{ and} \\ &\text{(a) } E(U_{it} - K_{it} | d_i, X_{it}, K_{it}, \psi) = 0 \\ &\text{(b) } \psi \text{ is identified.} \end{aligned} \tag{27}$$

When inserted into the earnings function of equation (1), K_{it} purges the equation of dependence between U_{it} and d_i. Rewriting equation (1) to incorporate K_{it},

$$Y_{it} = X_{it}\beta + d_i\alpha + K_{it} + (U_{it} - K_{it}). \tag{28}$$

The purged disturbance $(U_{it} - K_{it})$ is orthogonal to the right-hand side variables in the new equation. This (possibly nonlinear) regression applied to equation (28) consistently estimates the parameters (α, β, ψ). Moreover, condition (27) implies that $(U_{it} - K_{it})$ is orthogonal to the right-hand side variables conditional on d_i, X_{it} and K_{it}:

$$E(Y_{it} | X_{it}, d_i, K_{it}) = X_{it}\beta + d_i\alpha + K_{it}.$$

Thus, if type (A) and (B) data are combined in any proportion, least squares estimators of equation (28) consistently estimate (α, β, ψ) provided the number of trainees and nontrainees in the sample both approach infinity. The class of control function estimators which satisfies condition (27) can be implemented without modification in choice-based samples.

The sample selection bias methods (d)–(e) described in Section IVB and the propensity score methods formulated in Section X exploit the control function principle. Our companion paper gives further examples of control function estimators.

E. Summary and Conclusions on Robustness Properties

Repeated cross-section estimators which do not exploit knowledge of the training identity of persons are not robust to choice-based sampling nor can they be weighted to produce consistent estimators of α. (Repeated cross-section estimators with training identity known can obviously be reweighted to produce consistent estimators.) However, these estimators are robust to contamination bias provided that the population proportion of trainees is known or can be consistently estimated.

A major conclusion of our analysis is that with the exception of repeated cross-section estimators with the training status of persons unknown, using robustness to contamination bias or choice-based sampling as a criterion for selecting estimators does not suggest a clear ordering of cross-section, repeated cross-section, or longitudinal estimators. Control function estimators are robust to choice-based sampling, but such estimators can be formed on all three types of data sets.

X. The Propensity Score and Mixture Models as Solutions to the General Problem of Selection Bias

Under special conditions, the control function of condition (27) *can be expressed as a function of the propensity score* $\Pr(d_i = 1 | Z_i)$ *and no other variables. If selection occurs only on observables, the control function of condition* (27) *can be expressed solely as a function of the propensity score. Propensity score methods offer no solution to the general problem of selection bias. The mixture modeling approach of Glynn et al.* (1986) *assumes access to data not usually available in analyzing problems of selection bias.*

A. Propensity Score Methods

In a series of papers, Rosenbaum and Rubin (1983, 1985) have advocated the use of the propensity score [$\Pr(d_i = 1 | Z_i)$] in a matching estimation method for controlling or reducing bias in observational studies. Some authors (e.g., Scheuren, 1985; Coleman, 1985) have proposed use of the

propensity score as an alternative to more conventional selection bias methods. The propensity score methodology suggests use of $\Pr(d_i = 1|Z_i)$ as a control function in a matching procedure.

Two distinct topics should be distinguished in evaluating this proposed "cure" for selection bias: (a) a statement of conditions under which there exists a control function that depends solely on the propensity score (and some parameters) and (b) the validity of matching methods. We do not discuss (b) in this paper.[21]

Assuming the existence of selection bias $[E(U_{it}d_i) \neq 0]$ and that X_{it} is distributed independently of U_{it} given Z_i, the propensity score methodology assumes in the fixed coefficient model that d_i and U_{it} are conditionally independent given Z_i and X_{it}

$$d_i \perp\!\!\!\perp U_{it}|(Z_i, X_{it}). \tag{29}$$

In a random coefficient model it assumes that

$$d_i \perp\!\!\!\perp (U_{it} + \epsilon_i d_i)|(Z_i, X_{it}).$$

We confine our discussion to the fixed coefficient case. The modifications required in our analysis for the random coefficient case are obvious.

For there to be a nontrivial selection problem $[E(U_{it}d_i) \neq 0]$ and for equation (29) to be satisfied, the only source of selection bias must be selection on the observables (as defined in Section IVE). Selection on unobservables is ruled out. Using the law of iterated expectations in the manner of Rosenbaum and Rubin (1983, Theorem 3), if equation (29) is true, then

$$\begin{aligned} &E\left[Y_{it}|d_i, X_{it}, \Pr(d_i = 1|Z_i, X_{it})\right] \\ &\quad = X_{it}\beta + d_i\alpha + E\left[U_{it}|\Pr(d_i = 1|Z_i, X_{it})\right]. \end{aligned} \tag{30}$$

The term $E[U_{it}|\Pr(d_i = 1|Z_i, X_{it})]$ may be used as a control function in the sense of the definition of condition (27). Instead of conditioning on Z_i, X_{it}, it is sufficient to condition on $\Pr(d_i = 1|Z_i, X_{it})$ in constructing a control function. If X_{it} is distributed independently of V_i given Z_i, then $\Pr(d_i = 1|Z_i, X_{it}) = \Pr(d_i = 1|Z_i)$ and the analysis may be conducted in terms of the propensity score defined in Section IIB. Thus, it is possible to reduce the scale of the matching problem if matching methods are used (assuming

[21] We note that the published literature on matching offers no formal proofs of any desirable property of matching estimators in the case in which regressor variables are continuously distributed (it is trivial to establish optimality properties for matching in the case in which all regressors are categorical with finite categories). Recent claims about the robustness of matching methods in the case in which the functional form of a regression model is unknown are not yet supported by published systematic theoretical arguments or by compelling Monte Carlo or empirical evidence. Assertions about the generality of matching methods remain to be substantiated. (See Barros, 1986, for some general theorems on the consistency and asymptotic normality of matching methods with continuous regressors.)

that X_{it} and Z_i are categorical variables with a finite number of categories) and consistent estimators of α are produced by matching. Otherwise, postulating the functional form of $E(U_{it}|\Pr(d_i = 1|Z_i, X_{it})]$, it is possible to use regression to consistently estimate α under conditions postulated in our companion paper.

A key assumption underlying the method is the existence of at least one regressor in the decision rule of equation (2). As noted in Section III, this rules out a completely nonparametric estimation strategy. However, Z_i need not be distributed independently of V_i, although in practice (e.g., Rosenbaum and Rubin, 1985) this is assumed to be the case with V_i logistically distributed. The repeated cross-section and longitudinal estimators, as well as the cross-section estimators that assume knowledge of the functional form of the distribution of U_{it} [or at least that $E(U_{it}^3) = E(U_{it}^5) = 0$] do not require any regressor in the decision rule.

Since Z_i is not independent of U_{it}, elements of Z_i are not, in general, valid instrumental variables for d_i, [see method (a) in Section IVB]. However, elements of Z_i may be uncorrelated with U_{it} and may be valid instruments. For the same reason, methods (b) and (c) of Section IVB are generally inappropriate. However, method (d) is still appropriate because as a consequence of equation (29),

$$E(U_{it}|d_i = 1, Z_i, X_{it}) = E(U_{it}|d_i = 0, Z_i, X_{it}) = E(U_{it}|Z_i, X_{it}).$$

By virtue of equation (30), one can condition on $\Pr(d_i = 1|Z_i, X_{it})$ rather than on Z_i, X_{it} to arrive at the special control function implicit in Rosenbaum and Rubin (1983).

Selection solely on the observables is a very special case of the general problem of selection bias. In the context of evaluating training programs, Z_i may be correlated with U_{it}, but U_{it} and V_i must be independent given Z_i. There can be no unobserved motivational or ability variables common to the equation governing the decision to enroll a person into training and the equation determining his or her potential earnings. In the context of Coleman's work on the choice between public and private schools, the propensity score methodology is valid if selection occurs solely on the observables so that there is no correlation between unobserved person-specific and family-specific motivational and ability factors that affect test outcomes and the decision to place a child in a private school. The propensity score methodology solves a very special problem (already considered by Barnow, et al., 1980) that is of limited interest to social science data analysts.

The propensity score can also be used (in a different way than that advocated by Rosenbaum and Rubin, 1983) in a setting in which there is selection on the unobservables but there is no selection on the observables (Z_i is independent of U_{it}) and Z_i and X_{it} are distributed independently of V_i, an assumption not required when there is selection solely on the observables. In this case, discussed in Heckman (1980) and our companion

paper (p. 188),

$$E(U_{it}|d_i=1, Z_i) = \frac{\int_{-\infty}^{\infty} u \int_{-Z_i\gamma}^{\infty} f(u,v)\,dv\,du}{\int_{-Z_i\gamma}^{\infty} f(v)\,dv}$$

$$E(U_{it}|d_i=0, Z_i) = \frac{\int_{-\infty}^{\infty} u \int_{-\infty}^{-Z_i\gamma} f(v)\,dv}{\int_{-\infty}^{-Z_i\gamma} f(v)\,dv}$$

Using the facts that

$$\Pr(d_i=0|Z_i) = 1 - F(-Z_i\gamma)$$
$$\Pr(d_i=0|Z_i) = F(-Z_i\gamma)$$

and the strict monotonicity of F (a new assumption), we may write

$$E(U_{it}|d_i=1, Z_i) = \frac{\int_{-\infty}^{\infty} u \int_{F^{-1}[1-\Pr(d_i=1|Z_i)]}^{\infty} f(u,v)\,dv\,du}{\Pr(d_i=1|Z_i)}$$
$$= K_1[\Pr(d_i=1|Z_i)]$$

$$E(U_{it}|d_i=0, Z_i) = \frac{\int_{-\infty}^{\infty} u \int_{-\infty}^{F^{-1}[1-\Pr(d_i=1|Z_i)]} f(u,v)\,dv\,du}{\Pr(d_i=0|Z_i)}$$
$$= K_0[\Pr(d_i=0|Z_i)]$$

where

$$K_1\Pr(d_i=1|Z_i) + K_0\Pr(d_i=0|Z_i) = 0,$$

and where

$$K = K_1 d_i + K_0(1-d_i)$$

is a control function in the sense of condition (27) and is a function solely of the propensity score.[22] Use of the propensity score in this fashion involves no new idea and is just an instance of estimator (d) given in Section IVB.

Note, however, that very different assumptions are required to justify this control function than are required to justify the control function for selection on observables implicit in Rosenbaum and Rubin (1983). Under the assumption that Z_i is distributed independently of U_{it} (or at least that one element of Z_i is uncorrelated with U_{it}), the appropriate elements of Z_i may be used as instruments for d_i, whereas they are invalid instruments under the assumptions of Rosenbaum and Rubin (1983) which produce a

[22] Under the null hypothesis of no selection bias, polynomials in $\Pr(d_i=1|Z_i)$ should not appear in the regression of Y_{it} on X_{it} and d_i (Heckman, 1980). The same test with obvious modification in the conditioning set can be applied to the model of Rosenbaum and Rubin (1983) or Barnow et al. (1980). These are exact tests under the null hypothesis.

nontrivial selection bias problem.[23] Note further that is is possible to test between these two specifications provided a control function

$$E(U_{it}|X_{it}, Z_i, d_i = j) \qquad j = 1, 0$$

exists (see Heckman and Robb, 1985, p. 191). This control function can be used to produce consistent estimators of α for either model, whereas each of the other two control function estimators are invalid under the conditions assumed to justify the other.

B. Mixture Modeling

The "mixture modeling" approach advocated by Glynn et al. (this volume) at this conference assumes access to data not typically available in the analysis of selection bias models. Nonrespondents are randomly sampled in a follow-up sample and give information that suffices to determine the parameters of the outcome distribution for nonrespondents without bias. In our context, they assume that trainees can be randomly placed in nontrainee status for one period. By appropriately weighting estimates for respondents and nonrespondents it is possible to estimate population parameters without bias. A simple consistent weighted mean estimator exists for their model, although the authors do not explicitly present it. [It is implicitly given by choosing a $N(0, \infty)$ prior.]

The Glynn et al. (this volume) paper is disappointing to us because of the caricature it presents of econometric methods for solving selection bias problems. It reiterates a false statement frequently made in the statistics community that econometric selection bias procedures depend on normality assumptions or other strong distributional assumptions. It should be apparent to the reader of this paper and our companion paper that this caricature of econometric work is false.

The Glynn et al. paper considers a model in which respondents and nonrespondents are assumed to come from different distributions. In many choice-theoretic behavioral models it is much more natural to postulate that respondents and nonrespondents (or trainees and nontrainees) have outcomes drawn from a common distribution and that they use a common decision rule in making decisions albeit with different consequences. *Ex post*, not *ex ante*, distributions are different. Missing in their analysis is any explicit causal or behavioral mechanism generating the data.

Equations like our (1) and (2) can be used to characterize the population as a whole, define the parameters of behavioral interest, and provide the context within which it is possible to judge the plausibility of various

[23] If some elements of Z_i are not correlated with U_{it}, they may be valid instruments in the Rosenbaum and Rubin model provided that a rank condition is satisfied.

identifying restrictions. This formulation of the selection problem enables the analyst to clarify the explicit rules generating the selected samples.

XI. Summary

This paper presents alternative methods for estimating the impact of treatments on outcomes when nonrandom selection characterizes the enrollment of persons into treatment categories. In the absence of genuine experimental data, some assumptions must be invoked to solve the problem of selection bias. The choice of an appropriate assumption requires appeal to context, *a priori* beliefs, and prior knowledge. There is no context-free solution to the problem of selection bias despite apparent claims to the contrary in the recent literature in statistics which solves selection problems by imposing ad hoc mathematical structures onto the data.

We have defined the parameters of behavioral interest for a prototypical problem of estimating the impact of training on earnings. We have explored the benefits of having access to cross-section, repeated cross-section, and longitudinal data by considering the assumptions required to use a variety of new and conventional estimators to identify the behavioral parameters of interest. We state just-identifying assumptions which cannot be tested with data and compare those with overidentifying assumptions which in principle are testable. We examine the plausibility of these assumptions when viewed in the light of prototypical decision rules determining the enrollment of persons into training. Since specification of decision rules is an inherently controversial issue, we consider the robustness of various estimators to ignorance about the decision process. Because many samples are choice-based samples and because the problem of measurement error is pervasive, we examine the robustness of estimators to choice based sampling and measurement error.

We find that cross-section selection bias estimators do not require the elaborate distributional assumptions frequently invoked in practice. Such conventional overidentifying assumptions are in principle testable.

A key conclusion of our analysis is that the benefits of longitudinal data have been overstated in the recent literature because a false comparison has been made. Repeated cross-section data can often be used to identify the same parameters as can be identified in longitudinal data. Uniquely longitudinal estimators require assumptions that are different from the assumptions required to justify cross-section or repeated cross-section estimators.

We also consider propensity score methods and mixture modeling approaches recently advocated as solutions to the selection problem in the statistics literature. We find that the propensity score method solves the problem of selection bias for the case in which selection occurs solely on observable characteristics. Mixture modeling, as presented at this conference, assumes access to data typically not available. When it does not (as

in Rubin, 1977), it solves the problem of selection bias by invoking normality assumptions.

Any just-identified solution to the problem of selection bias in nonexperimental data requires an appeal to principles or assumptions that cannot be tested with data. Behavioral social scientists often appeal to context, beliefs, and *a priori* theory. Statisticians tend to substitute ad hoc mathematical assumptions in place of contextual assumptions. Such mathematical assumptions contain implicit behavioral premises, but these are rarely stated. Until these implicit behavioral assumptions are made explicit and the limitations of mathematical statistics are clearly recognized, there will be no convergence in views on the validity, and limits, of competing approaches to the selection bias problem. The "solution" to the selection bias problem lies outside of formal statistics.

Acknowledgments. . This research was supported by NSF SES-8107963 and NIH-1-R01-HD16846-01 to the Quantitative Economics Group at NORC. Heckman is affiliated with that group and the Department of Economics at the University of Chicago. Robb is affiliated with NORC and the Chicago Corporation. We have benefited from helpful comments made by Stephen Stigler and members of the Statistics Workshop at the University of Chicago. We also thank John Tukey for his comments at the ETS Conference and for very helpful correspondence. Don Rubin is also thanked for his comments. Ricardo Barros made especially helpful comments on several drafts of this paper.

Bibliography

Amemiya, T. (1981). "Qualitative response models: A survey." *J. Econ. Lit.*, 19, 1483–1536.

Ashenfelter, O. (1978). "Estimating the effect of training programs on earnings." *Rev. Econ. Statist.*, 60, 47–57.

Barnow, B., Cain, G., and Goldberger, A. (1980). "Issues in the analysis of selectivity bias." In E. Stromsdorfer and G. Farkas (eds.), *Evaluation Studies*, vol. 5. San Francisco: Sage.

Barros, R. (1986). *Three Essays on Selection and Identification Problems in Economics*. Ph.D. thesis, University of Chicago, Chicago, Illinois.

Bassi, L. (1983). *Estimating the Effect of Training Programs with Nonrandom Selection*. Ph.D. thesis, Princeton University, Princeton, New Jersey.

Chamberlain, G. (1982). "Multivariate regression models for panel data." *J. Econometrics*, 18, 1–46.

Coleman, J.C. (1985). "Schools, families and children." Ryerson Lecture, University of Chicago, April 1985.

Cox, D.R. *The Planning of Experiments*. New York: John Wiley, (1958).

Dawid, A.P. (1979). "Conditional independence in statistical theory" (with discussion). *J. Roy. Statist. Soc. Ser. B*, 41, 1–31.

Fienberg, S., Singer, B., and Tanur, J. (1985). "Large-scale social experimentation in

the United States." In A.C. Atkinson and S. Fienberg (eds.), *A Celebration of Statistics*. Berlin/New York: Springer-Verlag.

Fisher, R.A. (1953). *The Design of Experiments*. London: Hafner.

Goldfeld, S. and Quandt, R. (1976). "Techniques for estimating switching regressions." In S. Goldfeld and R. Quandt (eds.), *Studies in Nonlinear Estimation*. Cambridge, Massachusetts: Ballinger.

Heckman, J. (1976). "Simultaneous equations models with continuous and discrete endogenous variables and structural shifts." In S. Goldfeld and R. Quandt (eds.), *Studies in Nonlinear Estimation*. Cambridge, Massachusetts: Ballinger.

Heckman, J. "Dummy endogenous variables in a simultaneous equations system." *Econometrica*, 46, 931–961.

Heckman, J. (1979). "Sample selection bias as a specification error." *Econometrica*, 47, 153–161.

Heckman, J. (1980). "Addendum to sample selection bias as a specification error." In E. Stromsdorfer and G. Farkas (eds.), *Evaluation Studies*, vol. 5. San Francisco: Sage.

Heckman, J. and Neumann, G. (1977). "Union wage differentials and the decision to join unions." Unpublished manuscript, University of Chicago, Chicago, Illinois.

Heckman, J. and Robb, R. (1985). "Alternative methods for evaluating the impact of interventions." In J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data.*, New York: Cambridge University Press, pp. 156–245.

Heckman, J. and Wolpin, K. (1976). "Does the contract compliance program work?: An analysis of Chicago data." *Indust. Labor Relations Rev.*, 19, 415–433.

Lee, L.F. (1978). "Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables." *Intl. Econ. Rev.*, 19, 415–433.

Little, R.J. (1985). "A note about models for selectivity bias." *Econometrica*, 53(6), 1469–1474.

MaCurdy, T. (1982). "The use of time series processes to model the error structure of earnings in a longitudinal data analysis." *J. Econometrics*, 18(1), 83–114.

Manski, C. and Lerman, S. (1977). "The estimation of choice probabilities from choice-based samples." *Econometrica*, 45, 1977–1988.

Manski, C. and McFadden, D. (1981). "Alternative estimators and sample designs for discrete choice analysis." In C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Massachusetts: MIT Press, pp. 117–136.

Mundlak, Y. (1961). "Empirical production functions free of management bias." *J. Farm Econometrics*, 43, 45–56.

Mundlak, Y. (1978). "On the pooling of time series and cross section data." *Econometrica*, 46, 69–85.

Rosenbaum, P. and Rubin, D. (1983). "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70, 41–55.

Rosenbaum, P. and Rubin, D. (1985). "Constructing a control group using multivariate sampling methods that incorporate the propensity score." *Amer. Statist.*, 39(1), 33–38.

Roy, A. (1951). "Some thoughts on the distribution of earnings." *Oxford Econ. Pap.*, 3, 135–146.

Rubin, D. (1977). "Formalizing subjective notions about the effects of nonrespondents in sample surveys." *J. Amer. Statist. Assoc.*, 72(359), 538–543.

Scheuren, F. (1985). "Evaluating manpower training: Some notes on data handling issues." Report to JTLS Panel, U.S. Department of Labor, Washington, D.C.

Simon, H. (1957). "Spurious correlation: A causal interpretation." In H. Simon (ed.), *Models of Man*. New York: John Wiley, 37–49.