ALTERNATIVE METHODS FOR EVALUATING THE IMPACT OF INTERVENTIONS
An Overview

James J. HECKMAN and Richard ROBB, Jr.*
University of Chicago, Chicago, IL 60637, USA

This paper presents methods for estimating the impact of training on earnings when non-random selection characterizes the enrollment of persons into training. We explore the benefits of cross-section, repeated cross-section and longitudinal data for addressing this problem by considering the assumptions required to use a variety of new and conventional estimators given access to various commonly encountered types of data. We investigate the plausibility of assumptions needed to justify econometric procedures when viewed in the light of prototypical decision rules determining enrollment into training. We examine the robustness of the estimators to choice-based sampling and contamination bias.

1. Introduction

This paper considers the problem of estimating the effect of training on earnings when enrollment into training is the outcome of a non-random selection process. The analysis of training presented here serves as a prototype for the analysis of the closely related problems of deriving selection bias free estimates of the impacts of unionism, migration, job turnover, unemployment and affirmative action programs on earnings.

Our previous study [Heckman and Robb (1985)] investigates the prior restrictions needed to be imposed to secure consistent estimators of the selection bias free impact of training on earnings when the analyst has access to different types of data. We consider the plausibility of these restrictions in the light of economic theory.

Here we report the key findings from our previous paper. Occasionally, the reader is referred to that paper for precise statements of technical conditions or proofs that do not provide essential insights for understanding the main points.

*We thank Ricardo Barros for helpful comments. This research was supported by NSF Grants SES-8107963 and SES-8411242 and NIH Grant R01-HD-19226 to the Quantitative Methods Center at NORC.

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. Assuming random sampling techniques are applied to collect all three types of data, the three sources form a hierarchy: longitudinal data can be used to generate a single cross-section or a set of temporal cross-sections in which the identities of individuals are ignored, and repeated cross-sections can be used as single cross-sections.

Our conclusions are rather startling. Although longitudinal data are widely regarded 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 fixed effect specifications of earnings functions are adopted. Estimators based on repeated cross-section data for unrelated persons identify the same parameter. An analogous statement holds for virtually all longitudinal estimators.

However, we question the plausibility of conventional specifications. They are not motivated by economic theory, and when examined in that light they seem implausible. We propose richer longitudinal specifications of the earnings process and enrollment decision derived from economic theory. In addition, we propose a variety of new estimators. A few of these estimators require longitudinal data, but for most, longitudinal 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.

We also question recent claims that cross-section approaches to estimating the impact of training on earnings are strongly dependent on arbitrary assumptions about distributions of unobservables and about the nature of exclusion restrictions. While some widely-used cross-section estimators suffer from this defect, such commonly invoked assumptions are not an essential feature of the cross-sectional approach. 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 non-parametric 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 enrollment rule. For most cross-section estimators this requires precise specification of the decision rule. Longitudinal and repeated cross-section estimators do not require this.
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) non-randomness of available samples and especially oversampling of trainees (the choice based sample problem)?
   (b) time inhomogeneity in the environment ("non stationarity")?
   (c) the absence of a control group of non-trainees or the contamination of the control group so that the training status of individuals is not known for the control sample?

Notably absent from this list of questions is any mention of the efficiency of estimators for cross-section, repeated cross-section and longitudinal data. 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 in order to estimate a single coefficient when the analyst has access to different types of data. Since different assumptions about the underlying model are invoked in order to justify the validity of alternative estimators, an efficiency comparison is often meaningless. Under the assumptions about an underlying model that justify one estimator, 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, for two reasons conventional efficiency comparisons are often meaningless. 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 economic 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 may have a much smaller sampling variance than 'efficient' longitudinal estimators fit on much smaller samples. In this
sense, our proposed cross-section and repeated cross-section estimators may be *feasibly efficient* given the relative sizes of the samples for the two types of data sources.

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.

This paper is organized as follows. Section 2 describes the notation and an economic model for enrollment of individuals into training. Sections 3–9 each begin by listing a set of findings from our earlier paper. The findings are discussed and illustrated via simple examples. The paper concludes with a brief summary.

### 2. Notation and a model of program participation

#### 2.1. 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 earnings of individual $i$ in period $t$ by $Y_{it}$. Earnings depend on a vector of observed characteristics, $X_{it}$. Post-program earnings ($t > k$) also depend on a dummy variable, $d_{it}$, which equals one if the $i$th individual participates and is zero if he does not. Let $U_{it}$ represent the error term in the earnings equation and assume that $E[U_{it}] = 0$. Adopting a linear specification,

$$
Y_{it} = X_{it}\beta + d_{it}\alpha + U_{it}, \quad t > k,
$$

$$
= X_{it}\beta + U_{it}, \quad t \leq k,
$$

where $\beta$ and $\alpha$ are parameters.

Throughout this paper we assume that $X_{it}$ is uncorrelated with $U_{it}$. When $X_{it}$ contains lagged values of $Y_{it}$, we assume that (1) can be solved for a reduced form expression of exogenous variables, and we use that expression in place of (1). In some cases, independence between $X_{it}$ and lagged, current or future values of $U_{it}$ will be required as well.

Eq. (1) assumes that training has the same effect on everyone. In the next section we consider issues that arise when $\alpha$ varies among individuals. Throughout most of this paper we ignore effects of training which grow or decay over time. [See Heckman and Robb (1985) for a discussion of this topic.]
We now turn to the stochastic relationship between $d_i$ and $U_{it}$ in (1). For this purpose, we develop a more detailed notation which describes the enrollment rules that select individuals into training.

2.2. 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 benefits to the appropriate decision-makers from taking training. It is a function of observed ($Z_i$) and unobserved ($V_i$) variables. Thus

$$IN_i = Z_i \gamma + V_i. \tag{2}$$

In terms of this function,

$$d_i = 1 \quad \text{iff} \quad IN_i > 0$$

$$= 0 \quad \text{otherwise.}$$

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 $1 > p > 0$.

The central problem considered in this paper arises when the decision to take training is not random with respect to the disturbance in the earnings function. More precisely, the problem of selection bias arises when

$$\E[U_{it},d_i] \neq 0.$$

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

To interpret various specifications of eq. (2), we need to specify an economic 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. The earnings function is assumed to be (1). For simplicity, we assume that training programs accept all applicants. Our previous paper considers more general models.

All prospective trainees are assumed to discount earnings streams by a common discount factor $1/(1 + r)$. From (1) training raises trainee earnings by $\alpha$ 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 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

$$PV_i(1) = E_{k-1} \left( S_i + \sum_{j=1}^{\infty} \left( \frac{1}{1+r} \right)^j Y_{i,k+j} + \sum_{j=1}^{\infty} \frac{\alpha}{(1+r)^j} \right).$$

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 decision rule (2),

$$IN_i = PV_i(1) - PV_i(0) = E_{k-1}[S_i - Y_{ik} + \alpha/r]. \quad (3)$$

so the decision to train is characterized by the rule

$$d_i = 1 \quad \text{iff} \quad E_{k-1}[S_i - Y_{ik} + \alpha/r] > 0, \quad (4)$$

$$= 0 \quad \text{otherwise}.$$

Let $W_i$ be the part of the subsidy which the econometrician observes (with associated coefficient $\phi$) and let $\tau_i$ be the part which he does not observe:

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

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,

$$d_i = 1 \quad \text{iff} \quad S_i - Y_{ik} + \alpha/r = W_i\phi + \alpha/r - X_{ik}\beta + \tau_i - U_{ik} > 0, \quad (5)$$

$$= 0 \quad \text{otherwise}.$$

Then $(\tau_i - U_{ik}) = \gamma_i$ in (2) and $(W_i, X_{ik})$ corresponds to $Z_i$ in (2). Assuming that $(W_i, X_{ik})$ is distributed independently of $V_i$, makes (5) a standard discrete choice model.

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 $\tau_i$ are independent of $U_{it}$), then $E[U_{it}d_i] = 0$ only if the unobservables in period $t$ are (mean)
independent of the unobservables in period $k$ or

$$E[U_{it}|U_{ik}] = 0 \quad \text{for} \quad t > k.$$  

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 $m$ so

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

where the $\epsilon_{i,t-j}$ are iid, then for $t - k > m$, $E[U_{it}d_i] = 0$. On the other hand, if $U_{it}$ follows a first-order autoregressive scheme, then $E[U_{it}|U_{ik}] \neq 0$ for all finite $t$ and $k$.

3. 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 post-program earnings of the trained and what the earnings in post-program 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 and attention centers on estimating the mean response to training. The second notion is frequently 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 non-random assignment of persons to training, it is useful to distinguish two questions that are frequently confused in the literature:

Q1: ‘What would be the mean impact of training on earnings if people were randomly assigned to training?’

Q2: ‘How do the post-program 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 post-program 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 non-trainees even in the absence of any training program. The two questions have the same answer if eq. (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 attention centers on estimating the population mean response to training.

In the presence of non-random assignment and variation in the impact of training among persons, the two questions have different answers. Question 2 is the appropriate one to ask if interest centers on forecasting the change in the mean of the post-training earnings of trainees 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.

To clarify these issues, we consider a random coefficient version of (1) in which \( \alpha \) varies in the population. In this model, the impact of training may differ across persons and may even be negative for some people. We write in place of (1)

\[
Y_{it} = X_{it}'\beta + d_i\alpha_i + U_{it}, \quad t > k.
\]

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

\[
Y_{it} = X_{it}'\beta + d_i\bar{\alpha} + \{U_{it} + d_i\epsilon_i\}. \tag{6}
\]

An alternative way to derive this equation is to express it as a two-sector switching model following Roy (1951), Heckman and Neumann (1977) and Lee (1978). Let

\[
y_{1it} = X_{1it}'\beta_1 + U_{1it}
\]

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

\[
y_{0it} = X_{0it}'\beta_0 + U_{0it}
\]

be the wage of individual \( i \) in sector 0. Letting \( d_i = 1 \) if a person is in sector 1 and letting \( d_i = 0 \) otherwise, we may write the observed wage as

\[
Y_{it} = d_iY_{1it} + (1 - d_i)Y_{0it}
\]

\[
= X_{it}'\beta_0 + E[X_{it}|d_i = 1](\beta_1 - \beta_0) d_i
\]

\[
+ [(X_{it} - E[X_{it}|d_i = 1])(\beta_1 - \beta_0) + U_{1it} - U_{0it}] d_i + U_{0it}.
\]

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 \( U_{0it} = U_{1it} \), produces eq. (6).

In this model there is a fundamental non-identification result when no regressors appear in the decision rule (2). Without a regressor in (2) and in the
absence of any further distributional 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

\[
E[Y_{it} | d_i = 1, Z_i, X_{it}] = X_{it} \beta + \bar{\alpha} + E[\epsilon_i | d_i = 1, Z_i, X_{it}]
\]

\[+ E[U_{it} | d_i = 1, Z_i, X_{it}],
\]

\[
E[Y_{it} | d_i = 0, Z_i, X_{it}] = X_{it} \beta + E[U_{it} | d_i = 0, Z_i, X_{it}].
\]

Unless \( E[\epsilon_i | d_i = 1, Z_i, X_{it}] \) is known, without invoking distributional assumptions it is impossible to decompose \( \bar{\alpha} + E[\epsilon_i | d_i = 1, Z_i, X_{it}] \) into its constituent components unless there is independent variation in \( E[\epsilon_i | d_i = 1, Z_i, X_{it}] \) across observations [i.e., a regressor appears in (2)]. Without a regressor, \( E[\epsilon_i | d_i = 1, Z_i, X_{it}] \) 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

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

where

\[
\alpha^* = \bar{\alpha} + E[\epsilon_i | d_i = 1],
\]

and content ourselves with the estimation of \( \alpha^* \). If everywhere we replace \( \alpha \) with \( \alpha^* \), the fixed coefficient analysis of eq. (1) applies to (7).

The parameter \( \alpha^* \) answers Q2. 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 as has operated in the past. In the presence of non-random selection into training it does not answer Q1. Indeed, without regressors in decision rule (2), this question cannot be answered unless specific distributional assumptions are invoked.

Random assignment of persons to training does not usually represent a relevant 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 \( \alpha^* \) in (7). Since eq. (7) may be reparametrized in the form of eq. (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.

4. Cross-sectional procedures

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

4.1. Without distributional assumptions a regressor is needed

Let $\overline{y}_i^{(1)}$ denote the sample mean of trainee earnings and let $\overline{y}_i^{(0)}$ denote the sample mean of non-trainee earnings:

$$\overline{y}_i^{(1)} = \frac{\sum d_i y_{it}}{\sum d_i}, \quad \overline{y}_i^{(0)} = \frac{\sum (1-d_i) y_{it}}{\sum (1-d_i)},$$

for $0 < \sum d_i < I$, where $I$ is the number of observations. We retain the assumption that the data are generated by a random sampling scheme. If no regressors appear in (1) then $X_i/\beta = \beta_i$ and

$$\text{plim } y_t^{(1)} = \beta_i + \alpha + E[U_{it}|d_i = 1],$$

$$\text{plim } y_t^{(0)} = \beta_i + E[U_{it}|d_i = 0].$$

Thus

$$\text{plim}(\overline{y}_i^{(1)} - \overline{y}_i^{(0)}) = \alpha + E[U_{it}|d_i = 1]/(1-p),$$

since $pE[U_{it}|d_i = 1] + (1-p)E[U_{it}|d_i = 0] = 0$. Even if $p$ were known, $\alpha$ cannot be separated from $E[U_{it}|d_i = 1]$ using cross-sectional data on sample means. Sample variances do not aid in securing identification unless $E[U_{it}^2|d_i = 0]$ or $E[U_{it}^2|d_i = 1]$ is known a priori. Similar remarks apply to the information from higher moments.

4.2. Overview of cross-sectional procedures which use regressors

If, however, $E[U_{it}|d_i = 0, Z_i]$ is a non-constant function of $Z_i$, it is possible (with additional assumptions) to solve this identification problem. Securing identification in this fashion explicitly precludes a fully non-parametric strategy in which both the earnings function (1) and decision rule (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 $\alpha$ is not identified from cross-section data. Restrictions across strata are required.

If $E[U_{it}|d_i = 1, Z_i]$ is a non-constant 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 present 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 eq. (1), and $\alpha$ can be consistently
estimated by instrumental variables methods. No distributional assumptions about $U_i$ or $V_i$ are required [Heckman (1978)].

(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, $\gamma$ in (2) can be consistently estimated by conventional methods in discrete choice analysis [Amemiya (1981)]. If $Z_i$ is distributed independently of $U_i$, $F(-Z_i\gamma)$ can be used as an instrument for $d_i$ in eq. (1) [Heckman (1978)].

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

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

$\gamma$ and $\alpha$ can be consistently estimated using $F(-Z_i\gamma)$ in place of $F(-Z_i\gamma)$ in the preceding equation [Heckman (1976,1978)] or else the preceding equation can be estimated by non-linear least squares, estimating $\beta$, $\alpha$ and $\gamma$ jointly (given the functional form of $F$) [Barnow, Cain and Goldberger (1980)].

(d) If the functional forms of $E[U_{i}|d_i = 1, Z_i]$ and $E[U_{i}|d_i = 0, Z_i]$ as functions of $Z_i$ are known up to a finite set of parameters, it is sometimes possible to consistently estimate $\beta$, $\alpha$ and the parameters of the conditional means from the (non-linear) regression function

$$E[Y_{it}|d_i, Z_i] = X_{it} \beta + d_i \alpha + d_i E[U_{it}|d_i = 1, Z_i]$$

$$+ (1 - d_i) E[U_{it}|d_i = 0, Z_i]. \tag{8}$$

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 (2) [Barnow, Cain and Goldberger (1980)].

(e) Instead of (d), it is possible to use a two-stage estimation procedure if the joint density of $(U_{it}, V_i)$ is assumed known up to a finite set of parameters. 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 regression (8) is run using estimated $E$ values in place of population $E$ values on the right-hand side of the equation.

(f) Under the assumptions of (e), use maximum likelihood to consistently estimate $\alpha$ [Heckman (1978)]. Note that a separate value of $\alpha$ 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 (e.g., $\alpha$, can be estimated for each cross-section).

Conventional selection bias approaches (d)–(f) as well as (b)–(c) rely on strong distributional assumptions but in fact these are not required. Given that a regressor appears in decision rule (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 practice, however, $Z_i$ and $U_{it}$ are usually assumed to be independent. We next discuss the instrumental variables procedure in greater detail.

4.3. 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:

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

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

$$
\sum_{i=1}^{I_t} \left( J_{2it}' J_{1it} / I_t \right)
$$

has full column rank uniformly in $I_t$ for $I_t$ sufficiently large, where $I_t$ denotes the number of individuals in period $t$. (9b)

With these assumptions, the I.V. estimator,

$$
\left( \hat{\beta} \right)_{IV} = \left( \sum_{i=1}^{I_t} \left( J_{2it}' J_{1it} / I_t \right) \right)^{-1} \sum_{i=1}^{I_t} \left( J_{2it}' Y_{it} / I_t \right),
$$

is consistent for $(\beta, \alpha)$ 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 non-linear function of variables appearing in $X_{it}$ as long as (9) is 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.

4.4. Identification through distributional assumptions about the marginal distribution of $U_{it}$

If no regressor appears in decision rule (2), the estimators presented so far in this section cannot be used to estimate $\alpha$ consistently unless additional
restrictions are imposed. Heckman (1978) demonstrates that if \((U_{it}, V_i)\) are jointly normally distributed, \(\alpha\) is identified even if there is no regressor in enrollment rule (2). His conditions are overly strong.

If \(U_{it}\) has zero third and fifth central moments, \(\alpha\) 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 \(\alpha\) 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 \(\alpha\).

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

\[
\begin{align*}
E[U_{it}^3] &= 0, \\
E[U_{it}^5] &= 0, \\
\{U_{it}, V_i\} &\text{ is iid.}
\end{align*}
\]

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 analogues 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 (1), so \(X_{it}\beta = \beta_i\). The proposed estimator finds the value of \(\hat{\alpha}\) that sets

\[
(1/I_t) \sum_{i=1}^{I_t} \left[ (Y_{it} - \bar{Y}) - \hat{\alpha}(d_i - \bar{d}) \right]^3
\]

and

\[
(1/I_t) \sum_{i=1}^{I_t} \left[ (Y_{it} - \bar{Y}) - \hat{\alpha}(d_i - \bar{d}) \right]^5
\]

as close to zero as possible in a suitably chosen metric where, as before, the overbar denotes sample mean. In our earlier paper, we establish the existence of a unique consistent root that sets (11a) and (11b) to zero in large samples.

5. 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 $\alpha$ (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. In the notation of eq. (1), $X_{it}\beta = \beta_i$. Then, assuming a
random sampling scheme generates the data.

$$\text{plim } \frac{\bar{Y}_t}{I_t} = \frac{\text{plim } \sum Y_{it} / I_t}{\text{plim } I_t} = \text{E}[\beta + \alpha_0, + U_{it}] = \beta_i + \alpha_0, \quad t > k,$$

$$\text{plim } \frac{\bar{Y}_{t'}}{I_{t'}} = \frac{\text{plim } \sum Y_{it'} / I_{t'}}{\text{plim } I_{t'}} = \text{E}[\beta_{t'} + U_{it'}] = \beta_{t'}, \quad t' < k.$$

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

$$\text{plim}(\bar{Y}_t - \bar{Y}_{t'})/\hat{\rho} = \alpha,$$

where $\hat{\rho}$ is a consistent estimator of $\rho = \text{E}[d_i]$.

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

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

From $L$ temporally distinct cross-sections, it is possible to estimate con-
sistently the $L - 1$ $\pi$-parameters and $\alpha$ provided that the number of observa-
tions in each cross-section becomes large, and there is at least one pre-program
and one post-program cross-section.

If the effect of training differs across periods, it is still possible to identify $\alpha_i$
provided that the environment changes in a 'sufficiently regular' way. For

\[^1\text{If regressors appear in the earnings function, the following procedure can be used. Rewrite (1) as } Y_{it} = \beta_i + X_{it} \sigma + d_i \alpha + U_{it}. \text{ It is possible to estimate } \sigma \text{ from pre-program data. Replace } Y_{it} \text{ by } \tilde{Y}_{it} - \hat{\sigma} \text{ and the analysis in the text goes through. Note that we are assuming that no } X_{it} \text{ variables become non-constant after period } k.\]
example, suppose
\[ \beta_t = \pi_0 + \pi_1 t \quad \text{for} \quad t > k, \]
\[ \alpha_t = \phi_0 (\phi_1)^{t-k} \quad \text{for} \quad t > k. \]

In this case, \( \pi_0, \pi_1, \phi_0, \phi_1 \) are identified from the means of four cross-sections, so long as at least one of these means comes from a pre-program period.

In Heckman and Robb (1985) we rigorously state the conditions required to consistently estimate \( \alpha \) using repeated cross-section data that does not record the training identity of individuals. Section 9 examines the sensitivity of this class of estimators to violations of the random sampling assumption.

6. 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 \( \alpha \) by assuming only that the error term in the earnings equation is covariance stationary.

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

6.1. The fixed effects method

This method was developed by Mundlak (1961, 1978) and refined by Chamberlain (1982). It is based on the following assumption:

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

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

\[ \mathbb{E}[Y_{it} - Y_{i't} | d_i, X_{it} - X_{i't}] = (X_{it} - X_{i't})\beta + d_i\alpha, \quad t > k > t'. \]

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

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

\[ U_{it} = \phi_t + \epsilon_{it}, \quad (13) \]

where \( \epsilon_{it} \) is a mean zero random variable independent of all other values of \( \epsilon_{i't} \).
and is distributed independently of $\phi_i$, a mean zero person-specific time-invariant random variable. Assuming that $S_i$ in decision rule (5) is distributed independently of all $\varepsilon_{i_i}$ except possibly for $\varepsilon_{i_k}$, then (12) will be satisfied.

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

$$E_{k-1}[\varepsilon_{i_k}] = 0,$$

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 $\varepsilon_{i_i}$ except possible for $\varepsilon_{i_k}$, this model also produces (12).

We investigate the plausibility of (12) with respect to more general decision rules and error processes in section 8.

6.2. $U_{it}$ follows a first-order autoregressive process

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

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

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

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

Assume further that the perfect foresight rule (5) determines enrollment, and the $v_{ij}$ are distributed independently of $S_i$ and $X_{ik}$ in (5). Heckman and Wolpin (1976) invoke similar assumptions in their analysis of affirmative action programs. As a consequence of these assumptions,

$$E[Y_{it}|X_{it}, X_{it'}, d_i, Y_{it'}] = \left( X_{it} - X_{it} \rho^{i-t'} \right) \beta + (1 - \rho^{i-t'}) d_i \alpha + \rho^{i-t'} Y_{it'},$$

so that (linear or non-linear) least squares applied to (16) consistently estimates
\(\alpha\) as the number of observations becomes large. (The appropriate non-linear regression increases efficiency by imposing the cross-coefficient restrictions.)

6.3. \(U_{it}\) is covariance-stationary

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

\[ U_{it} \text{ is covariance stationary so } \mathbb{E}[U_{it}U_{i,t-j}] = \mathbb{E}[U_{it}U_{i,t'-j}] = \sigma_j \text{ for } j \geq 0 \text{ for all } t, t', \]

Access to at least two observations on pre-program earnings in \(t'\) and \(t'-j\) as well as one period of post-program earnings in \(t\) where \(t-t'=j\).

\[ p\mathbb{E}[U_{it}|d_i = 1] \neq 0. \]

Unlike the two previous examples, we make no assumptions here about the appropriate enrollment rule or about the stochastic relationship between \(U_{it}\) and the cost of enrollment \(S_t\).

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

\[ Y_{it} = \beta_i + d_i \alpha + U_{it}, \quad t > k, \]
\[ Y_{i't'} = \beta_{i'} + U_{i't'}, \quad t' < k, \]

where \(\beta_i\) and \(\beta_{i'}\) are period-specific shifters.

From a random sample of pre-program earnings from periods \(t'\) and \(t' - j\), \(\sigma_j\) can be consistently estimated from the sample covariances between \(Y_{it}\) and \(Y_{i't' - j}\):

\[ m_1 = \left( \sum (Y_{it} - \bar{Y}_t)(Y_{i,t'-j} - \bar{Y}_{i'-j}) \right) / I, \quad \text{plim } m_1 = \sigma_j. \]

If \(t > k\) and \(t-t'=j\) so that the post-program 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 \(Y_{i't'}\):

\[ m_2 = \left( \sum (Y_{it} - \bar{Y}_t)(Y_{i't'} - \bar{Y}_{i'}) \right) / I, \]
which has the probability limit

\[ \text{plim} m_2 = \sigma + \alpha \rho \text{E}[U_{it} | d_i = 1], \quad t > k > t'. \]

Form the sample covariance between \(d_i\) and \(Y_{it}^{*}\),

\[ m_3 = \left( \sum (Y_{it}^{*} - \bar{Y}_{it}^{*})d_i \right) / I, \]

with probability limit

\[ \text{plim} m_3 = \rho \text{E}[U_{it} | d_i = 1], \quad t' < k. \]

Combining this information and assuming \( \rho \text{E}[U_{it} | d_i = 1] \neq 0 \) for \( t' < k \),

\[ \text{plim} \hat{\alpha} = \text{plim} ((m_2 - m_1) / m_3) - \alpha. \]

7. Repeated cross-section analogues 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 non-random attrition which plague panel data.

The previous section presented longitudinal estimators of \( \alpha \). In each case, however, \( \alpha \) can actually be identified with repeated cross-section data. Here we establish this claim. Our earlier paper gives additional examples of longitudinal estimators which can be implemented on repeated cross-section data.²

7.1. The fixed effect model

As in section 6.1, assume that (12) holds so

\[ \text{E}[U_{it} | d_i = 1] = \text{E}[U_{it} | d_i = 1], \]
\[ \text{E}[U_{it} | d_i = 0] = \text{E}[U_{it} | d_i = 0], \]

for all \( t > k > t' \). Let \( X_{it} \beta = \beta_i \) and define, in terms of the notation of section 4.1,

\[ \hat{\alpha} = \left[ \bar{Y}_{t}^{(1)} - \bar{Y}_{t}^{(0)} \right] - \left[ \bar{Y}_{t}^{(1)} - \bar{Y}_{t}^{(0)} \right]. \]

²We also produce one example of a longitudinal procedure which has no repeated cross-section analogue.
Assuming random sampling, consistency of $\hat{\alpha}$ follows immediately from (12):

$$\text{plim} \hat{\alpha} = [\alpha + \beta_i - \beta_i + E[U_{it}|d_i = 1] - E[U_{it}|d_i = 0]$$

$$- [\beta_i - \beta_i + E[U_{it}|d_i = 1] - E[U_{it}|d_i = 0]]$$

$$= \alpha.$$ 

7.2. $U_{it}$ follows a first-order autoregressive process

In one respect the preceding example is contrived. It assumes that in pre-program 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 pre-period $k$ earnings to see if they qualify for admission), but this seems unlikely. One advantage of longitudinal data for estimating $\alpha$ in the fixed effect model is that if the survey extends before period $k$, the identity of future trainees is known.

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

$$E[U_{it}|d_i] = 0, \quad t > k, \quad (18)$$

as in section 6.2. With three successive post-program cross-sections in which the identity of trainees is known, it is possible to identify $\alpha$.

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

$$\text{plim} \bar{Y}_{j}^{(1)} = \beta_j + \alpha + E[U_{ij}|d_i = 1],$$

$$\text{plim} \bar{Y}_{j}^{(0)} = \beta_j + E[U_{ij}|d_i = 0].$$

From (18),

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

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

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

$$E[U_{i,t+2}|d_i = 0] = \rho^2 E[U_{it}|d_i = 0].$$
Using these formulae, it is straightforward to verify that $\hat{\rho}$, defined by

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

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

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

is consistent for $\alpha$.

For this model, the advantage of longitudinal data is clear. Only two time periods of longitudinal data are required to identify $\alpha$, 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 6.2 does not identify $\alpha$ unless the analyst observes earnings without error. Given three years of longitudinal data and assuming that measurement error is serially uncorrelated, one could instrument (15) using earnings in the earliest year as an instrument. Thus one advantage of the longitudinal estimator disappears in the presence of measurement error.

7.3. Covariance stationarity

For simplicity we suppress regressors in the earnings equation and let $X_{it}\beta = \beta_i$. Assume that conditions (17) are satisfied. Before presenting the repeated cross-section estimator, it is helpful to record the following facts:

\[
\text{var}(Y_{it}) = \sigma_u^2(1 - p) + 2\alpha E[U_{it}|d_i = 1]p + \sigma_u^2, \quad t > k, \tag{19a}
\]

\[
\text{var}(Y_{it'}) = \sigma_u^2, \quad t' < k, \tag{19b}
\]

\[
\text{cov}(Y_{it}, d_i) = \alpha p(1 - p) + p E[U_{it}|d_i = 1]. \tag{19c}
\]
Note that \( E[U_{it}^2] = E[U_{it}'^2] \) by virtue of assumption (17a). Then

\[
\hat{\alpha} = \left( p(1-p) \right)^{-1} \left[ \frac{\sum(Y_{it} - Y_i)}{I_t} \right. \\
- \sqrt{ \left( \frac{\sum(Y_{it} - Y_i) d_i}{I_t} \right)^2 - p(1-p) \left( \frac{\sum(Y_{it} - Y_i)^2}{I_t} - \frac{\sum(Y_{it}' - Y_i')^2}{I_t'} \right) } \right]
\]

(20)

is consistent for \( \alpha \).

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

Notice that the estimators of sections 6.3 and 7.3 exploit different features of the covariance stationarity assumptions. The longitudinal procedure only requires that \( E[U_{it+1} U_{it-1}] = E[U_{it-1} U_{it+1}] \) for \( j > 0 \); variances need not be equal across periods. The repeated cross-section analogue presented above only requires that \( E[U_{it} U_{it-j}] = E[U_{it-j} U_{it}] \) for \( j = 0 \); covariances may differ among equispaced pairs of the \( U_{it} \).

8. First difference methods

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

8.1. Models which justify condition (12)

Whenever condition (12) holds, \( \alpha \) can be estimated consistently from the difference regression method described in section 6.1. Section 6.1 presents a model which satisfies condition (12): the earnings residual has a permanent-transitory structure, decision rule (4) or (5) determines 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 (12). For example, even if (13) characterizes \( U_{it} \), if \( S_i \) in (5) does not have same joint (bivariate) distribution with respect to all \( t \), except for \( e_{it} \), (12) may be violated.

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

\[
U_{it} = \rho U_{it-1} + \nu_{it},
\]

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

$$E[U_{it}d_i = 1] = E[U_{it}U_{ik} + X_{it} \beta - \alpha/r < S_i] = \rho^{t-k} E[U_{ik}|d_i = 1]$$

$$\neq E[U_{it}d_i = 1] = E[U_{it}|U_{ik} + X_{ik}\beta - \alpha/r < S_i],$$

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

$$E[U_{it}d_i = 1] = \rho^{t-k} E[U_{ik}|d_i = 1],$$

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

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

8.2. More general first difference estimators

Instead of (12), assume that

$$E[(U_{it} - U_{it'}) (X_{it} - X_{it'})] = 0 \text{ for some } t, t', \quad t > k > t',$$

$$E[(U_{it} - U_{it'})d_i] = 0 \text{ for some } t > k > t'. \quad (21)$$

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 $\alpha$.

Our earlier paper presents three examples of models that satisfy (21) but not (12). Here we discuss one of them. Suppose that

$U_{it}$ is covariance stationary, \hspace{1cm} (22a)

$U_{it}$ has a linear regression on $U_{ik}$ for all $t$ (i.e., $E[U_{it}|U_{ik}] = \beta_{it}U_{ik}$), \hspace{1cm} (22b)

$U_{it}$ is mutually independent of $(X_{ik}, S_i)$ for all $t$, \hspace{1cm} (22c)

$\alpha$ is common to all individuals (so the model is of the fixed coefficient form), \hspace{1cm} (22d)

The environment is one of perfect foresight where decision rule (5) determines participation. \hspace{1cm} (22e)

Under these assumptions, condition (21) characterizes the data.
To see this note that (22a) and (22b) imply there exists a $\delta$ such that

$$U_{it} = U_{i,k+j} = \delta U_{ik} + \omega_{it}, \quad j > 0, \quad t > k,$$

$$U_{i'i'} = U_{i,k-j} = \delta U_{ik} + \omega_{i'i'}, \quad j > 0.$$

and

$$E[\omega_{it}|U_{ik}] = E[\omega_{i'i'}|U_{ik}] = 0.$$

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 (22c),

$$E[\omega_{it}|d_i = 1] = 0,$$

since $E[\omega_{it}] = 0$ and because (22c) guarantees that the mean of $\omega_{it}$ does not depend on $X_{ik}$ and $S_i$. Similarly,

$$E[\omega_{i'i'}|d_i = 1] = 0,$$

and thus (21) 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.

8.3 Anomalous features of first difference estimators

Nearly all of the estimators require a control group (i.e., a sample of non-trainees). The only exception is the fixed effect estimator in a time homogeneous environment. In this case, if condition (12) or (21) holds, if we let $X_{it} = \beta_t$ to simplify the exposition, and if the environment is time homogeneous so $\beta_t = \beta_t$, then

$$\hat{\alpha} = \bar{Y}_{t}^{(1)} - \bar{Y}_{t}^{(1)}$$

consistently estimates $\alpha$. 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 post-program data. The covariance stationary estimators of sections 6.3 and
7.3, certain repeated cross-section estimators and first difference methods constitute an exception to this rule. In this sense, these estimators are anomalous.

Fixed effect estimators are also robust to departures from the random sampling assumption. For instance, suppose condition (12) or (21) is satisfied, but that the available data oversample or undersample trainees (i.e., the proportion of sample trainees does not converge to $p = \mathbb{E}[d_i]$). Suppose further that the analyst does not know the true value of $p$. Nevertheless, a first difference regression continues to identify $\alpha$. Most other procedures do not share this property.

9. Non-random 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 non-random samples. Frequently they consist of pooled data from two sources: (a) a sample of trainees selected from program records and (b) a sample of non-trainees 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 File 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 ($Y_i, X_i,$ and $Z_i$) for a sample of trainees ($d_i = 1$),
(B) Earnings, earnings characteristics, and enrollment characteristics for a sample of non-trainees ($d_i = 0$),
(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 $\alpha$ 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 earlier paper gives explicit case by case treatment of these issues for each estimator developed there.

### 9.1. The I.V. estimator: Choice-based sampling

If condition (9a) is strengthened to read

$$E[X_{it}U_{it}|d_i] = 0, \quad E[g(Z^c_{it})U_{it}|d_i] = 0,$$

and (9b) is also met, the I.V. estimator is consistent for $\alpha$ in choice-based samples.

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

$$\begin{pmatrix}
\frac{\sum X_{it}X_{it}}{I_t} & \frac{\sum X_{it}d_i}{I_t} \\
\frac{\sum g(Z^c_{it})X_{it}}{I_t} & \frac{\sum g(Z^c_{it})d_i}{I_t}
\end{pmatrix}\begin{pmatrix}
\hat{\beta} \\
\hat{\alpha}
\end{pmatrix} = \begin{pmatrix}
\frac{\sum X_{it}Y_{it}}{I_t} \\
\frac{\sum g(Z^c_{it})Y_{it}}{I_t}
\end{pmatrix}$$

$$= \begin{pmatrix}
\frac{\sum X_{it}X_{it}}{I_t} & \frac{\sum X_{it}d_i}{I_t} \\
\frac{\sum g(Z^c_{it})X_{it}}{I_t} & \frac{\sum g(Z^c_{it})d_i}{I_t}
\end{pmatrix}\begin{pmatrix}
\beta \\
\alpha
\end{pmatrix} + \begin{pmatrix}
\frac{\sum X_{it}U_{it}}{I_t} \\
\frac{\sum g(Z^c_{it})U_{it}}{I_t}
\end{pmatrix}.$$  (24)

Since (23) guarantees that

$$\text{plim}_{I_t \to \infty} \frac{\sum X_{it}U_{it}}{I_t} = 0 \quad \text{and} \quad \text{plim}_{I_t \to \infty} \frac{\sum g(Z^c_{it})U_{it}}{I_t} = 0,$$

and the rank condition (9b) holds, the I.V. estimator is consistent.

In a choice based sample, let the probability that an individual has enrolled in training be $p^*$. Even if (9a) and (9b) are satisfied, there is no guarantee that
condition (25) will be met without invoking (23). This is so because

$$\lim_{t \to \infty} \sum_{i} \frac{X_{it}U_{it}}{I_t} = \mathbb{E}[X_{it}U_{it}|d_i = 1] p^* + \mathbb{E}[X_{it}U_{it}|d_i = 0](1 - p^*)$$

and

$$\lim_{t \to \infty} \sum_{i} g(Z_{it}^e)U_{it} = \mathbb{E}[g(Z_{it}^e)U_{it}|d_i = 1] p^* + \mathbb{E}[g(Z_{it}^e)U_{it}|d_i = 0](1 - p^*)$$

These expressions are not generally zero, so the I.V. 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 (23) is satisfied. However, it is not necessary to invoke (23). Provided $p$ is known, it is possible to reweight the data to secure consistent estimators under the assumptions of section 4. Multiplying eq. (1) by the weight

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

and applying I.V. to the transformed equation produces an estimator that satisfies (25). It is straightforward to check that weighting the sample at hand back to random sample proportions causes the I.V. method to consistently estimate $\alpha$ and $\beta$. [See Heckman and Robb (1985).]

### 9.2. The I.V. estimator: Contamination bias

For data of type (C), $d_i$ is not observed. Applying the I.V. estimator to pooled samples (A) and (C), assuming that observations in (C) have $d_i = 0$, produces an inconsistent estimator.

In terms of the I.V. eq. (24) from sample (C) it is possible to generate the cross-products

$$\frac{\sum X_{it}X_{it}}{I_C}, \quad \frac{\sum g(Z_{it}^e)X_{it}}{I_C}, \quad \frac{\sum X_{it}Y_{it}}{I_C}, \quad \frac{\sum g(Z_{it}^e)Y_{it}}{I_C},$$

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

\[
\sum_{i} X_{it}d_i \quad , \quad \sum_{i} g(Z_{it}^*)d_i.
\]

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

\[
\operatorname{plim}_{Ic \to \infty} \frac{\sum X_{it}d_i}{I_c} = p\mathbb{E}[X_{it}|d_i = 1],
\]

\[
\operatorname{plim}_{Ic \to \infty} \frac{\sum g(Z_{it}^*)d_i}{I_c} = p\mathbb{E}[g(Z_{it}^*)|d_i = 1].
\]

But the means of \(X_{it}\) and \(g(Z_{it}^*)\) in sample (A) converge to

\[
\mathbb{E}[X_{it}|d_i = 1] \quad \text{and} \quad \mathbb{E}[g(Z_{it}^*)|d_i = 1],
\]

respectively. Hence, inserting the sample (A) means of \(X_{it}\) and \(g(Z_{it}^*)\) multiplied by \(p\) in the second column of the matrix I.V. eq. (24) produces a consistent I.V. estimator provided that in the limit the size of samples (A) and (C) both approach infinity at the same rate.

9.3. Repeated cross-section methods with unknown training status and choice-based sampling

The repeated cross-section estimators discussed in section 5 are inconsistent when applied to choice-based samples unless additional conditions are assumed. For example, when the environment is time-homogeneous and (12) also holds, \((\bar{Y}_t - \bar{Y}_t')/p\) remains a consistent estimator of \(\alpha\) in choice-based samples as long as the same proportion of trainees are sampled in periods \(t'\) and \(t\). If a condition such as (12) is not met, it is necessary to know the identity of trainees in order to weight the sample back to the proportion of trainees that would be produced by a random sample in order 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.

9.4. 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:

\[ K_{it} \text{ depends on variables } \ldots, Y_{i,t+1}, Y_{it}, Y_{i,t-1}, \ldots, X_{i,t+1}, X_{it}, X_{i,t-1}, \ldots, d_i \text{ and parameters } \psi, \text{ and} \]

\[ \mathbb{E}[U_{it} - K_{it}|d_i, X_{it}, K_{it}, \psi] = 0, \quad (26a) \]

\( \psi \) is identified. \quad (26b)

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

\[ Y_{it} = X_{it}\beta + d_i\alpha + K_{it} + \{U_{it} - K_{it}\}. \quad (27) \]

The purged disturbance \( \{U_{it} - K_{it}\} \) is orthogonal to the right-hand-side variables in the new equation. Thus (possibly non-linear) regression applied to (27) consistently estimates the parameters \((\alpha, \beta, \psi)\). Moreover, (26) implies that \( \{U_{it} - K_{it}\} \) is orthogonal to the right-hand-side variables conditional on \( d_i, X_{it}, \) and \( K_{it} \):

\[ \mathbb{E}[Y_{it}|X_{it}, d_i, K_{it}] = X_{it}\beta + d_i\alpha + K_{it}. \]

Thus if type (A) and (B) data are combined in any proportion, least squares performed on (27) produces consistent estimates of \((\alpha, \beta, \psi)\) provided the number of trainees and non-trainees in the sample both approach infinity. The class of control function estimators which satisfy (26) can be implemented without modification in choice-based samples.

We encountered a control function in section 6. For the model satisfying (14) and (18),

\[ K_{it} = \rho(Y_{i,t-1} - X_{i,t-1} \beta - d_i\alpha), \quad t > k + 1, \]

so \( \psi = (\rho, \beta, \alpha) \). The sample selection bias methods (d)-(e) described in section 4.2 exploit the control function principle. Our longer paper gives further examples of control function estimators.

10. Conclusion

This paper presents alternative methods for estimating the impact of training on earnings when non-random selection characterizes the enrollment of persons into training. We have explored the benefits of cross-section, repeated cross-section and longitudinal data for addressing this problem by considering the assumptions required to use a variety of new and conventional estimators given access to various commonly encountered types of data. We also investigate the plausibility of assumptions needed to justify econometric procedures when viewed in the light of prototypical decision rules determining enrollment into training. Because many of the available samples are choice-based samples
and because the problem of measurement error in training status is pervasive in many available control samples, we examine the robustness of the estimators to choice-based sampling and contamination bias.

A key conclusion of our analysis is that the benefits of longitudinal data have been overstated in the recent econometric literature on training because a false comparison has been made. A cross-section selection bias estimator does not require the elaborate and unjustified assumptions about functional forms often invoked in cross-sectional studies. Repeated cross-section data can often be used to identify the same parameters as longitudinal data. The uniquely longitudinal estimators require assumptions that are different from and often no more plausible than the assumptions required for cross-section or repeated cross-section estimators.

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