Econometric Society Monographs

Editors:
Jean-Michel Grandmont *Centre d'Études Prospectives d'Économie Mathématique Appliquées à la Planification, Paris*
Charles F. Manski *University of Wisconsin, Madison*

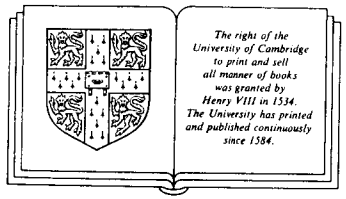
The Econometric Society is an international society for the advancement of economic theory in relation to statistics and mathematics. The Econometric Society Monograph Series is designed to promote the publication of original research contributions of high quality in mathematical economics and theoretical and applied econometrics.

Other titles in the series:

Werner Hildenbrand, Editor *Advances in economic theory*
Werner Hildenbrand, Editor *Advances in econometrics*
G. S. Maddala *Limited-dependent and qualitative variables in econometrics*
Gerard Debreu *Mathematical economics: twenty papers by Gerard Debreu*
Jean-Michel Grandmont *Money and value: a reconsideration of classical and neoclassical monetary economics*
Franklin M. Fisher *Disequilibrium foundations of equilibrium economics*
Bezalel Peleg *Game theoretic analysis of voting in committees*
Roger J. Bowden and Darrell A. Turkington *Instrumental variables*
Andreu Mas-Colell *The theory of general economic equilibrium: a differentiable approach*

Longitudinal analysis of labor market data

Edited by
JAMES J. HECKMAN and BURTON SINGER
University of Chicago *Yale University*



CAMBRIDGE UNIVERSITY PRESS

Cambridge
London New York New Rochelle
Melbourne Sydney

1985

CHAPTER 4

Alternative methods for evaluating the impact of interventions

James J. Heckman and Richard Robb, Jr.

1 The problem and an overview of solutions to it

This chapter considers the problem of estimating the impact of interventions in the presence of selection decisions by agents. For specificity we focus on the problem of estimating the impact of training on earnings when the enrollment of persons into training is the outcome of a 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.

This chapter investigates the prior restrictions needed to secure consistent estimators of the selection-bias-free impact of training on earnings. We examine their plausibility in the light of economic theory.

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 posttraining 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 the order of their availability and in inverse order of their cost of acquisition.[1] 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.

Longitudinal data are widely regarded as a panacea for selection and simultaneity problems. Yet to date there has been no systematic statement of conditions under which longitudinal data are required to isolate estimates of the impact of training on earnings free of selection bias. This chapter presents such conditions. En route to deriving these conditions we investigate the assumptions required to use less costly cross-section and repeated cross-section data to estimate the impact of training on earnings. Once this is done, it is possible to state what assumptions can be relaxed or tested if the analyst has access to longitudinal data.

Our conclusions are rather startling. Provided that conventional fixed effect specifications of earnings functions are adopted, there is no need to use longitudinal data to identify the impact of training on earnings. Estimators based on repeated cross-section data for unrelated persons identify the same parameter. 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. Some of these require longitudinal data, but for others longitudinal data are still not required. A major conclusion of our chapter 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. While some widely used cross-section estimators suffer from this defect, such assumptions are not an essential feature of the cross-section 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 nonparametric cross-section approach. Without prior restrictions, it is not possible to cross-classify observations on the basis of values assumed by explanatory variables in the earnings function and the enrollment rules and do a "regressor-free" estimation of the impact of training on earnings that is free of selection bias. A regressor is required in the enrollment rule, and for most cross-section estimators this requires specification of the functional form of the decision rule. Longitudinal and repeated cross-section estimators do not require this condition.

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").
   c. The absence of a control group of nontrainees or the contamination of the control group so that the training status is not known for a random sample of the population

Notably absent from this list of questions is any mention of the relative 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 chapter is on the tradeoffs in assumptions that must be imposed in order to estimate a common 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 different estimators, an efficiency comparison is often meaningless. Under the assumptions about an underlying model that justify one estimator, another estimator may not be applicable. Only by postulating a common assumption set that is unnecessarily large for any single estimator is it possible to compare the efficiency of alternative estimators. For the topic of this chapter – model identification – the efficiency issue is a red herring.

This chapter is divided into six sections. Section 1 presents a formal statement of the problem considered in this chapter and an intuitive statement of our principal results. Section 2 presents a set of prototypical decision rules that are assumed to characterize a person's decision to enroll in training. One contribution of this chapter to the literature on training is to make explicit the econometric implications of various decision processes that confront agents contemplating enrollment in training. The validity of many estimators hinges on the specification of decision rules. Previous work is statistical in nature and advocates various estimators on the basis of implicit enrollment rules that are never fully articulated. For this reason there has been little use of economic theory in previous work to guide the choice of appropriate estimators.

Section 3 examines cross-section estimators. Section 4 presents repeated cross-section estimators when the training status of persons is unknown, and Section 5 considers longitudinal estimators and other repeated cross-section estimators when the training status of persons is known. The chapter concludes with a summary and comparison of procedures and a brief discussion of the efficiency of alternative methods when that concept is well defined.

1.1 *The problem and some notation*

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 impact of training on earnings if people were randomly assigned to training?

Question 2:
How do the postprogram 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 people who would be selected as trainees 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 only when training has the same impact on earnings for everyone or when assignment to training is random.

In the presence of nonrandom 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 increment in the posttraining earnings of trainees when the same selection rule characterizes past and future trainees. Question 1 is intrinsically more difficult to answer because it asks to forecast the increment in the earnings of trainees over their pretraining earnings when no selection bias characterizes enrollment while selection bias characterizes the available data. It is really a special case of the more general question:

Question 3:
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?

It is important to note that the answer to the more modest question 2 is all that is required to estimate the future program impact if future selection criteria are like past criteria.

To focus on essential aspects of the problem, we initially assume that training has the same effect on everyone (so the three questions have the same answers). We further assume that training is offered only once (in period k) in a person's life cycle. If the training option is pursued, it takes one period to complete. During a period of training, trainees receive no employment income.

The earnings of individual i in period t (Y_{it}) are assumed to depend on characteristics ($\mathbf{X}_{it}$). Postprogram earnings depend on a dummy variable (d_i) that equals 1 if a person participated in training and is 0 otherwise. To account for unobserved characteristics, a mean zero disturbance (U_{it}) is assumed. A linear version of this specification may be written as

$$\begin{aligned} Y_{it} &= \mathbf{X}_{it}\boldsymbol{\beta} + d_i\alpha + U_{it}, \qquad t > k \\ Y_{it} &= \mathbf{X}_{it}\boldsymbol{\beta} \qquad\quad + U_{it}, \qquad t \leq k \end{aligned} \tag{1.1}$$

where $E(U_{it}) = 0$. U_{it} is assumed to be independently distributed across persons.

Boldfaced variables, denoted by Roman letters, are understood to be row vectors. Boldfaced parameters, denoted by Greek letters, are understood to be column vectors. We adopt this convention throughout this chapter.

For the moment we assume that α is the same for everyone. Later we alter (1.1) and substitute α_i for α to obtain a random coefficient model. In that case the three questions have different answers. We also abstract from likely depreciation or growth effects of training that would suggest writing α_t in place of α in (1.1). We indicate below how the analysis can be modified to account for temporal variation in program impact.

The decision to participate in training may be determined by a trainee, 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) of taking training. It is a function of observed ($\mathbf{Z}_i$) and unobserved (v_i) variables. Thus,

$$IN_i = \mathbf{Z}_i\gamma + V_i \tag{1.2}$$

In terms of this function

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

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

We normalize V_i so that $E(V_i) = 0$ and $\text{Var}(V_i) = 1$. The distribution function of V_i [$\Pr(V_i \leq v_i)$] is denoted as $F(v_i)$. V_i is assumed to be independently and identically distributed across persons.

At this point, nothing has been stated about the nature of the stochastic dependence among $\mathbf{X}_{it}$, d_i, U_{it}, V_i, and $\mathbf{Z}_i$. Our vagueness is deliberate. This chapter explores a variety of assumptions about stochastic dependence relations among these variables. Below we are very precise about our stochastic dependence assumptions.

The central problem considered in this chapter arises when the decision to take training is not random with respect to the disturbance term in

earnings function (1.1). More precisely, the problem of selection bias arises when

(A-1) $E(U_{it}d_i) \neq 0$

This may occur because of stochastic dependence between U_{it} and the unobservable V_i in (1.2) (selection on the unobservables) or because of stochastic dependence between U_{it} and $\mathbf{Z}_i$ in (1.2) (selection on the observables). When condition (A-1) occurs a standard least squares assumption is violated and α is not identified without invoking additional assumptions.

This is not the only potential source of nonidentifiability of α. For example, even if $E(U_{it}d_i) = 0$, α is not identified without further assumptions when $E(\mathbf{X}_{it}U_{it}) \neq 0$ unless $\mathbf{X}_{it}$ is orthogonal to d_i. This source of nonidentifiability for α is entirely conventional, and we have little to say about it here. Throughout most of this chapter we focus on models in which (A-1) characterizes the data and is the sole source of nonidentifiability of α. When $\mathbf{X}_{it}$ contains lagged values of Y_{it}, we assume that (1.1) can be solved for a reduced form expression of exogenous variables and we use that expression in place of (1.1).

An extreme form of model (1.1) that focuses on (A-1) as the principal source of nonidentifiability is one with no regressors other than the program dummy so that

$$\begin{aligned} Y_{it} &= \beta_t + d_i\alpha + U_{it}, \qquad t > k \\ Y_{it} &= \beta_t \qquad\quad + U_{it}, \qquad t \leq k \end{aligned} \tag{1.3}$$

If $\mathbf{X}_{it}$ assumes a finite number of possible values, this model is actually more general than (1.1) in that it can arise by stratifying (1.1) on the basis of values of the $\mathbf{X}_{it}$ regressors allowing β_t to differ across strata. In the remainder of this section, we adopt earnings specification (1.3) and decision rule (1.2). Although there is no regressor in the earnings equation, regressors may appear in enrollment equation (1.2).

1.2 *Estimators for models without selection bias*

To fix ideas it is useful to begin by considering estimation of the training effect in a model without selection bias so that $E(U_{it}d_i) = 0$ for all t, except possibly for $t = k$. We consider cross-section, repeated cross-section, and longitudinal estimators of α. We assume access to random samples of I_t individuals for each period t, and we assume that we know an individual's training status and age without error. We confine our attention to data on a single cohort.

One cross-section estimator for α is the difference between the earnings of participants and nonparticipants in a postprogram period ($t > k$). Define the mean earnings of trainees and nontrainees in period t as $\bar{Y}_t^{(1)}$ and $\bar{Y}_t^{(0)}$, respectively, so that

$$\bar{Y}_t^{(1)} = \frac{\sum Y_{it} d_i}{\sum d_i} \quad \text{and} \quad \bar{Y}_t^{(0)} = \frac{\sum Y_{it}(1 - d_i)}{\sum (1 - d_i)}$$

for $0 < \sum d_i < I_t$. The cross-section estimator of α is

$$\hat{\alpha}_C = \bar{Y}_t^{(1)} - \bar{Y}_t^{(0)} \quad \text{for} \quad t > k \tag{1.4}$$

Because $E(U_{it}|d_i) = 0$, we have

$$\text{plim}(\hat{\alpha}_C) = \alpha \quad \text{if} \quad 0 < \Pr(d_i = 1) < 1$$

Note that this estimator is robust to violations of the random sampling assumption provided that $E(U_{it}|d_i) = 0$.

One possible repeated cross-section estimator is $\hat{\alpha}_C$ applied to each cross section. Under the current assumptions, access to repeated cross sections only improves the capacity to estimate a separate $\alpha\ (=\alpha_t)$ for each period or if $\alpha_t = \alpha$ improves the efficiency of the estimator of α.

If the training status of persons is unknown, $\hat{\alpha}_C$ is useless. If, however, the analyst has access to the sample means of the earnings of cohorts in different years, it is sometimes possible to recover α without knowledge of an individual's training status. For a given cohort assume access to at least one postprogram and one preprogram mean of earnings, say $\bar{Y}_t$ and $\bar{Y}_{t'}$, where $t > k > t'$. In each cross section the individual identities of trainees or nontrainees need not be known. Assume that the postprogram population proportion that has taken training (p_t) is known or can be consistently estimated. Let $\hat{\alpha}_{RC}$ denote the repeated cross-section estimator defined as

$$\hat{\alpha}_{RC} = \bar{Y}_t - \bar{Y}_{t'} \quad \text{for} \quad t > k > t'$$

$$E(\hat{\alpha}_{RC}) = \beta_t - \beta_{t'} + p_t \alpha \tag{1.5}$$

If the environment is time-homogeneous (so that $\beta_t = \beta_{t'}$) and p_t is known or can be consistently estimated, then α can be estimated (without bias or consistently, respectively) by dividing $\hat{\alpha}_{RC}$ by p_t.

Unlike the cross-section estimator, the repeated cross-section estimator does not require that the identities of trainees be known in each cross section (provided that p_t is known). In this sense it is more robust. However, it is not robust to departures from the time homogeneity assumption, whereas the cross-section estimator is.

If the environment is sufficiently smooth, it is still possible to use $\hat{\alpha}_{RC}$ to consistently estimate α in a time-inhomogeneous environment. To do so

requires that β_t expressed as a function of t be described by at least one fewer parameter than the number of available cross sections. For example, from the mean earnings of three temporally distinct cross sections, if $\beta_t = \beta^0 + \beta_t^1$, then β^0, β^1, and α can be consistently estimated provided that (a) $p_t > 0$, (b) p_t can be consistently estimated, and (c) data for at least one year of postprogram earnings and one year of preprogram earnings are available.

If year effects are identical across cohorts, access to multiple-cohort data can aid in identifying α in a time-inhomogeneous environment. To see why this is so, suppose that three adjacent cohorts denoted A, B, C are followed. A is the oldest and C is the youngest. The mean of C's earnings in the period preceding their enrollment is an unbiased estimator of β_t—the period effect component of the mean earnings of A in their first postprogram year. Given p_t for cohort A, it is possible to consistently estimate α by subtracting the mean earnings of C from the mean earnings of A and dividing by p_t. Note further that with different (known or consistently estimable) proportions of trainees in posttraining cross sections t, t'' it follows that

$$\begin{aligned} \text{plim}\ \hat{\alpha}_{RC} &= \text{plim}(\bar{Y}_t - \bar{Y}_{t''}), \qquad t, t'' > k \\ &= \alpha(p_t - p_{t''}) \end{aligned}$$

so that α is consistently estimable solely from postprogram cross sections by dividing $\bar{Y}_t - \bar{Y}_{t''}$ by a consistent estimator of $p_t - p_{t''}$.

If the identity of trainees and nontrainees is known in each cross section, it is possible to estimate α consistently (using $\hat{\alpha}_C$) and hence β_t without any smoothness assumptions.

Abstracting from potential gains in devising more efficient estimators, longitudinal data provide no additional information beyond their use as cross-section or repeated cross-section data in models without selection bias.

1.3 *Models with selection bias*

All of the assumptions of Section 1.1 are retained except that we no longer assume that $E(U_{it}d_i) = 0$. We continue to assume that a random sampling scheme generates the data.

1.3.1. Cross-section estimators. The cross-section estimator $\hat{\alpha}_C$ is always biased for α. For each observation

$$\begin{aligned} E(Y_{it}|d_i = 1) &= \beta_t + \alpha + E(U_{it}|d_i = 1) \\ E(Y_{it}|d_i = 0) &= \beta_t + \qquad E(U_{it}|d_i = 0) \end{aligned}$$

Assuming that the data are generated by a random sampling scheme, we have

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

Since $E(U_{it}) = 0$ for all i, the expectation of each term inside the square brackets is nonzero and of opposite sign if $E(U_{it}d_i) \neq 0$, and so the expectation of the entire term is nonzero.

If there are no regressors in decision rule (1.2), so that $\Pr(d_i = 1) = p$ is the same constant for each observation and the U_{it} are i.i.d. (independent and identically distributed), then

$$E(U_{it}|d_i = 1) = \phi \quad \text{so} \quad E(U_{it}|d_i = 0) = -\frac{p}{1-p}\phi$$

and so then

$$\text{plim } \hat{\alpha}_C = \alpha + \frac{\phi}{1-p}$$

Even if p is known, using $\hat{\alpha}_C$, α cannot be separated from ϕ when data on means and variances are used.[2]

Another way to state this nonidentification result is to write the regression function

$$E(Y_{it}|d_i) = [\beta_t + E(U_{it}|d_i = 0)] + d_i(\alpha + E(U_{it}|d_i = 1) - E(U_{it}|d_i = 0)) \tag{1.6}$$

If

$$E(U_{it}|d_i = 1) = \phi \quad \text{and} \quad \Pr(d_i = 1) = p$$

then

$$E(Y_{it}|d_i) = \left(\beta_t - \frac{p}{1-p}\phi\right) + d_i\left(\alpha + \frac{\phi}{1-p}\right)$$

Unless $\beta_t = 0$ or other prior information is assumed, α is not identified.

If $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ is a nonconstant function of $\mathbf{Z}_i$, 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 (1.1) and decision rule (1.2) are estimated in each $(\mathbf{X}_{it}, \mathbf{Z}_i)$ stratum. For within each stratum, $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ is not a constant function of $\mathbf{Z}_i$ and α is not identified from cross-section data.

If $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ is a nonconstant function of $\mathbf{Z}_i$, 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 the alternative strategies. In Section 3 we present a systematic discussion of each approach.

(i) Suppose $\mathbf{Z}_i$ or a subset of $\mathbf{Z}_i$ is exogenous with respect to U_{it}. Under conditions specified more fully in Section 3, the exogenous subset may be used to construct an instrumental variable for d_i in equation (1.3), and α can be consistently estimated by instrumental variables methods. No distributional assumptions are required (Heckman, 1978).

(ii) Suppose that $\mathbf{Z}_i$ is distributed independently of V_i and the functional form of the distribution of V_i is known. Under standard conditions, γ in (1.2) can be consistently estimated using conventional methods in discrete choice analysis (Amemiya, 1981). If $\mathbf{Z}_i$ is distributed independently of U_{it}, $F(-\mathbf{Z}_i\hat{\gamma})$ can be used as an instrument for d_i in equation (1.2) (Heckman, 1978).

(iii) Under the same conditions as specified in (ii),

$$E(Y_{it}|\mathbf{Z}_i) = \beta_t + \alpha(1 - F(-\mathbf{Z}_i\gamma)) \tag{1.7}$$

β_t and α can be consistently estimated using $F(-\mathbf{Z}_i\hat{\gamma})$ in place of $F(-\mathbf{Z}_i\gamma)$ in equation (1.7) (Heckman, 1976, 1978) or else the equation can be fit by nonlinear least squares estimating β_t, α, and γ jointly (given the functional form of F) (Barnow, Cain, and Goldberger, 1980).

(iv) If the functional forms of $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ and $E(U_{it}|d_i = 0, \mathbf{Z}_i)$ as functions of $\mathbf{Z}_i$ are known up to a finite set of parameters, it is sometimes possible to consistently estimate β_t, α and the parameters of the conditional means from the (nonlinear) regression function

$$E(Y_{it}|d_i, \mathbf{Z}_i) = \beta_t + d_i\alpha + d_iE(U_{it}|d_i = 1, \mathbf{Z}_i) + (1 - d_i)E(U_{it}|d_i = 0, \mathbf{Z}_i) \tag{1.8}$$

One way to acquire information about the functional form of $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ 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 in principle. Note further that this procedure does not require that $\mathbf{Z}_i$ be distributed independently of V_i in (1.2) (Barnow, Cain, and Goldberger, 1980).

(v) Instead of (iv), it is possible to do a two-stage estimation procedure if the joint density of (U_{it}, V_i) is asssumed to be known up to a finite set of parameters. In stage 1, $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ and $E(U_{it}|d_i = 0, \mathbf{Z}_i)$ are determined up to some unknown parameters ψ by conventional discrete choice analysis. Then regression (1.8) is run using the estimated E in place of E on the right-hand side of the equation (Heckman, 1978, 1979).

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

Conventional selection bias approaches (iv)–(vi) rely on strong distributional assumptions, but in fact these are not required. Given that a regressor appears in decision rule (1.2), if it is exogenous with respect to

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 exogenous with respect to U_{it}. In practice, however, it is usually assumed to be so.

A regressor is not required in the decision rule if some other identifying information is invoked. For example, assuming joint normality of (U_{it}, V_i), Heckman (1978) establishes that α is identified in the absence of any regressor in selection rule (1.2). In Section 3, we demonstrate that if the marginal distribution of U_{it} satisfies some moment restrictions [e.g., $E(U_{it}^3) = 0$ and $E(U_{it}^5) = 0$ or other such conditions], α is identified from cross-section data without any regressor in (1.2) and without specifying the full joint distribution of (U_{it}, V_i).

1.3.2. Repeated cross-section estimators. Assuming a time-homogeneous environment and random sampling, the repeated cross-section estimator identifies α (a) without any regressor in decision rule (1.2), (b) without need to specify the joint distribution of U_{it} and V_i, and (c) without any need to identify the training status of individuals in the cross sections (but the proportion of trainees must be known or consistently estimable).

To understand why this claim is true, it is useful to distinguish two different sampling schemes. Above we defined d_i as a random variable that equals 1 if a member of the population has received training and is 0 otherwise. When random sampling is assumed, the probability of sampling a trainee is $\Pr(d_i = 1)$. When some other sampling rule is assumed, the probability of sampling a trainee is not $\Pr(d_i = 1)$. In year t, let this probability be $\Pr_t(d_i = 1) = p_t^*$. If $p_t^* = \Pr_t(d_i = 1) = \Pr(d_i = 1)$ and if the sampling is only on the basis of training status, then the available data are a random sample of the population.[3]

With this notation and assuming that (A-1) characterizes the data so that there is selection into training and that successive observations are independently and identically distributed across individuals, we get

$$E(\hat{\alpha}_{RC}) = \beta_t - \beta_{t'} + p_t^*\alpha + [p_t^*E(U_{it}|d_i = 1) + (1 - p_t^*)E(U_{it}|d_i = 0)] - [p_{t'}^*E(U_{it'}|d_i = 1) + (1 - p_{t'}^*)E(U_{it'}|d_i = 0)] \quad \text{for} \quad t > k > t'$$

In data from a random sample, the terms in square brackets are zero. As a consequence of time homogeneity, $\beta_t = \beta_{t'}$. Thus

$$\operatorname{plim} \hat{\alpha}_{RC} = p_t^*\alpha = p_t\alpha$$

and if p_t^* is known or consistently estimable, α is identified by dividing $\hat{\alpha}_{RC}$ by p_t^* assuming $p_t^* \neq 0$.

Thus under the stated conditions, $\hat{\alpha}_{RC}$ can be used to identify α while $\hat{\alpha}_{\mathcal{C}}$ is inconsistent for α. If the environment is time-inhomogeneous, $\hat{\alpha}_{RC}$ does not identify α, for reasons already presented in Section 1.1. It is necessary to invoke a smoothness assumption in order to use $\hat{\alpha}_{RC}$ to identify α. Unlike the case considered in Section 1.1, access to the trainee identity of individuals does not break the identification problem raised by time inhomogeneity.

The random sampling requirement is overly strong. Abstracting from time inhomogeneity, all that is required for $\hat{\alpha}_{RC}$ to identify α given knowledge of p_t^* is that

$$\text{(A-2)} \quad p_t^*E(U_{it}|d_i = 1) + (1 - p_t^*)E(U_{it}|d_i = 0) = p_{t'}^*E(U_{it'}|d_i = 1) + (1 - p_{t'}^*)E(U_{it'}|d_i = 0)$$

that is, that the expectations of the mean disturbances be equal in t and t', not that they both equal zero. In principle, it is possible for $\Pr_t(d_i = 1) \neq \Pr(d_i = 1)$ and condition (A-2) to be satisfied. Further, if (A-2) is satisfied and $p_t^* \neq p_{t'}^*$, then it is possible to use two postprogram cross sections to identify α.[4] While it is possible that (A-2) is satisfied in nonrandom samples, there is little evidence that suggests that available samples are collected so that they satisfy (A-2).

The random sampling requirement or the weaker condition (A-2) is the Achilles' heel of the repeated cross-section estimator $\hat{\alpha}_{RC}$. If (A-2) is violated, the repeated cross-section estimator does not identify α.

Provided that the training status of persons is known in each cross section and specific assumptions are made about the time series process generating U_{it} and its stochastic dependence on V_i, it is possible to construct other repeated cross-section estimators that are robust to general forms of time inhomogeneity in the environment. The proposed estimators do not require that the data be generated by a random sampling scheme or even that the same sampling rules be used in any two cross sections. Knowledge of the proportion of the population taking training is not required.

To exhibit one such estimator, suppose that decision rule (1.2) contains no regressor ($\mathbf{Z}_i = 1$) so that in principle α cannot be identified using the cross-section methods previously discussed. Assume further that for each t the pair (U_{it}, V_i) is independently and identically distributed across persons. Suppose further that the following additional information is available:

$$\text{(A-3)} \quad E(U_{it}|d_i = 1) = E(U_{it'}|d_i = 1), \qquad t > k > t'$$

and hence

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

Then α is identified in a general time-inhomogeneous environment.

The proposed repeated cross-section estimator subtracts the mean of trainee earnings in t from the mean of earnings in t' of those who will subsequently be trained. This identifies $\beta_t - \beta_{t'} + \alpha$. Subtracting the mean of nontrainee earnings in t from the mean of nontrainee earnings in t' identifies $\beta_t - \beta_{t'}$. Hence α can be identified by subtracting the second difference from the first.

Define $\hat{\alpha}^*_{RC}$ as the proposed estimator

$$\hat{\alpha}^*_{RC} = (\bar{Y}^{(1)}_t - \bar{Y}^{(1)}_{t'}) - (\bar{Y}^{(0)}_t - \bar{Y}^{(0)}_{t'}), \qquad t > k > t'$$

As a consequence of (A-3),

$$\begin{aligned} \operatorname{plim} \hat{\alpha}^*_{RC} &= (\beta_t - \beta_{t'} + \alpha) - (\beta_t - \beta_{t'}) \\ &= \alpha \end{aligned}$$

Note that longitudinal data are not required to implement the estimator although if they are available and if (A-3) is satisfied, they can be used to form $\hat{\alpha}^*_{RC}$. All that is required are data on the means of period t and period t' earnings for trainees and nontrainees.

One way that (A-3) can arise is if (a) the disturbance in the earnings equation (1.3) of person i is of the permanent–transitory form

$$U_{it} = \phi_i + \varepsilon_{it} \tag{1.9}$$

where ϕ_i is a mean zero i.i.d. random variable across i and ε_{it} is a mean zero i.i.d. random variable for all i and t and is distributed independently of ϕ_i; and (b) $\mathbf{Z}_i\gamma + V_i$ in decision rule (1.2) is distributed independently of ε_{it}. To justify (A-3) in the general case requires adopting a specification of the time series process generating the unobservables in person i's earnings equation and its time series relationship to the unobservable in decision rule (1.2). Longitudinal data can be used to test the validity of the assumed time series process for earnings even though they are not required to estimate the model if (A-3) is true. Other assumptions produce (A-3). Examples are given in Section 5.

In one respect this example is contrived. It assumes that in preprogram cross section t' ($< k$) we know the identity of future trainees. Such data might exist (e.g., individuals in 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 α in the model of this example is that if the survey extends before k, the identity of prospective trainees is known.

The need for preprogram earnings to identify α is, however, only an artifact of assumption (A-3) or the permanent–transitory error structure (1.9). Suppose instead that U_{it} follows a first-order autoregressive process so that

$$U_{it} = \rho U_{it-1} + \varepsilon_{it}, \qquad \rho \neq 1 \tag{1.10}$$

where ε_{it} is an i.i.d. mean zero disturbance. For $t > k$ suppose that

(A-4) $\quad E(\varepsilon_{it}|d_i) = 0, \qquad t > k$

[In Section 2 we justify (A-4) for a broad class of models.] 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 (U_{it}, V_i) is i.i.d. across i and that no regressor appears in (1.2),

$$\begin{aligned} \operatorname{plim} \bar{Y}^{(1)}_j &= \beta_j + \alpha + E(U^{(1)}_j) \\ \operatorname{plim} \bar{Y}^{(0)}_j &= \beta_j \qquad + E(U^{(0)}_j), \qquad j = t, t+1, t+2 \end{aligned}$$

where $E(U^{(1)}_j)$ is shorthand notation for $E(U_{ij}|d_i = 1)$ and $E(U^{(0)}_j) = E(U_{ij}|d_i = 0)$. Assuming that (A-4) is true,

$$\begin{aligned} E(U^{(1)}_{t+1}) &= \rho E(U^{(1)}_t) \\ E(U^{(0)}_{t+1}) &= \rho E(U^{(0)}_t) \\ E(U^{(1)}_{t+2}) &= \rho^2 E(U^{(1)}_t) \\ E(U^{(0)}_{t+2}) &= \rho^2 E(U^{(0)}_t) \end{aligned}$$

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

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

is consistent for ρ (plim $\hat{\rho} = \rho$) and that $\hat{\alpha}$ defined by

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

is consistent for α (plim $\hat{\alpha} = \alpha$).

Thus, with autoregressive error structure (1.10) and assumption (A-4), it is possible to consistently estimate α in a general time-inhomogeneous environment using only three cross sections of postprogram data if the training status of individuals is known. Longitudinal data can also be used for this purpose, but they are not required. They are not even required to test assumption (1.10) if four or more cross sections are available.

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 recover the same parameter.[5]

To establish why two periods of posttraining longitudinal data suffice to identify α, use (1.10) to write

$$\begin{aligned} Y_{i,t+1} &= \beta_{t+1} + d_i\alpha + U_{i,t+1} \\ &= \beta_{t+1} + d_i\alpha + \rho U_{it} + \varepsilon_{i,t+1} \end{aligned}$$

and substitute for U_{it} from (1.3) and collect terms to reach

$$Y_{i,t+1} = \beta_{t+1} - \rho\beta_t + d_i\alpha(1-\rho) + \rho Y_{it} + \varepsilon_{i,t+1} \tag{1.11}$$

As a consequence of (A-4) and the serial independence of ε_{it}, regression estimators of (1.11) are consistent for ρ and $\alpha(1-\rho)$ so that α can be consistently estimated. Heckman and Wolpin (1976) invoke (A-4) and estimate a multivariate version of (1.11) in their study of the impact of affirmative action programs.

Note, however, that the repeated cross-section estimator based on sample means is robust to mean zero measurement error in income. The regression estimator of equation (1.11) is not. An instrument is required for Y_{it}. One natural candidate is $Y_{i,t-1}$ if the additional assumption is made that the measurement error is independently distributed across t. But if this instrument is used, three periods of panel data are required to consistently estimate α. Thus in the presence of measurement error in income, the clear advantage of longitudinal data disappears. The main point, however, is that cross-section and repeated cross-section estimators based on means are robust to mean zero measurement error whereas regression estimators are not.

1.3.3. Longitudinal estimators. Many longitudinal data estimators considered in this chapter use an individual's earnings path (future or retrospective) to construct a control function which, when inserted into earnings function (1.1) or (1.3), purges the equation of covariance between d_i and U_{it}. We define a control function for the more general equation (1.1) so that the concept will not have to be defined twice.

(D-1) K_{it} is a control function for (1.1) if it depends on variables, ... $Y_{i,t+1}, Y_{it}, Y_{i,t-1}, \ldots, \mathbf{X}_{i,t+1}, \mathbf{X}_{it}, \mathbf{X}_{i,t-1}, \ldots, d_i$ and parameters ψ and if
(a) $E(U_{it} - K_{it})d_i = 0$,
(b) $E(U_{it} - K_{it})\mathbf{X}_{it} = 0$,
(c) $E(U_{it} - K_{it})K_{it} = 0$, and
(d) ψ is identified.

For many models discussed below it is sometimes possible to use a weaker form of conditions (a)–(c) that requires that

$$\operatorname*{plim}_{I_t \to \infty} \frac{\sum (U_{it} - K_{it})M_i}{I_t} = 0$$

for M_i equal to d_i, $\mathbf{X}_{it}$, and K_{it}. We use the stronger form of the conditions because it facilitates the exposition.

The basic idea underlying the control function is that when it is inserted into (1.1) and therefore implicitly subtracted from U_{it}, the purged disturbance $\{U_{it} - K_{it}\}$ is orthogonal, at least in large samples, to all of the right-hand-side variables in the new equation

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + d_i\alpha + K_{it} + \{U_{it} - K_{it}\} \tag{1.12}$$

Requirement (d) is simply that ψ can be identified from (nonlinear) regression estimation of (1.12).

We have already encountered a control function. For the model satisfying (A-4) and (1.10),

$$K_{it} = \rho(Y_{it-1} - \beta_{t-1} - d_i\alpha), \qquad t > k+1$$

so $\psi = (\rho, \beta_{t-1}, \alpha)$. Other examples of control functions will be given in Sections 3 and 5.

1.4 *Random coefficient specifications*

We now consider a random coefficient version of (1.3) in which α varies in the population. The motivation for this model is that the impact of training may differ across persons and may even be negative for some people. To capture this idea we write in place of (1.3)

$$Y_{it} = \beta_t + d_i\alpha_i + U_{it} \quad \text{for} \quad t > k$$

We define $E(\alpha_i) = \bar{\alpha} < \infty$ and $\varepsilon_i = \alpha_i - \bar{\alpha}$. $E(\varepsilon_i) = 0$ and $\operatorname{Var}(\varepsilon_i) < \infty$. With this notation we can rewrite the preceding equation as

$$Y_{it} = \beta_t + d_i\bar{\alpha} + \{U_{it} + d_i\varepsilon_i\} \tag{1.13}$$

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

$$Y_{1it} = \beta_{1t} + U_{1it}$$

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

$$Y_{0it} = \beta_{0t} + U_{0it}$$

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

$$\begin{aligned} Y_i &= d_iY_{1it} + (1-d_i)Y_{0it} \\ &= \beta_{0t} + (\beta_{1t} - \beta_{0t})d_i + U_{0it} + (U_{1it} - U_{0it})d_i \end{aligned}$$

Letting $\bar{\alpha} = \beta_{1t} - \beta_{0t}$, $\varepsilon_i = (U_{1it} - U_{0it})$, $\beta_{0t} = \beta_t$, and $U_{0it} = U_{it}$ produces equation (1.13). The difference between fixed coefficient and random coefficient specifications has not been appreciated in studies of unionism by Chamberlain (1984) and Lewis (1982), among others. (See note 24.) Björklund and Moffitt (1983) consider random coefficient models of training.

In this model there is a fundamental nonidentification result when no regressors appear in decision rule (1.2). Without a regressor in (1.2) and in the absence of any further distributional assumptions it is not possible to identify $\bar{\alpha}$ unless $E(\varepsilon_i|d_i = 1, \mathbf{Z}_i) = 0$ or some other known constant.

To see this note that

$$E(Y_{it}|d_i = 1, \mathbf{Z}_i) = \beta_t + \bar{\alpha} + E(\varepsilon_i|d_i = 1, \mathbf{Z}_i) + E(U_{it}|d_i = 1, \mathbf{Z}_i)$$

and

$$E(Y_{it}|d_i = 0, \mathbf{Z}_i) = \beta_t \qquad + E(U_{it}|d_i = 0, \mathbf{Z}_i)$$

Unless $E(\varepsilon_i|d_i = 1, \mathbf{Z}_i)$ is known, it is impossible without invoking distributional assumptions to decompose $\alpha + E(\varepsilon_i|d_i = 1, \mathbf{Z}_i)$ into its constituent components unless there is independent variation in $E(\varepsilon_i|d_i = 1, \mathbf{Z}_i)$ across observations [i.e., a regressor appears in (1.2)]. Without a regressor, $E(\varepsilon_i|d_i = 1, \mathbf{Z}_i)$ is a constant that is indistinguishable from $\bar{\alpha}$.

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

$$Y_{it} = \beta_t + d_i\alpha^* + \{U_{it} + d_i(\varepsilon_i - E(\varepsilon_i|d_i = 1))\} \tag{1.14}$$

where

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

and content ourselves with the estimation of α^*. If everywhere in Sections 1.1 and 1.2 we replace α with α^*, the preceding analysis goes through as before.

The parameter α^* answers question 2 of Section 1. It addresses the question of determining the effect of training on the people selected as trainees. This parameter is useful in making forecasts only when the same selection rule operates in the future as has operated in the past. It may not answer question 1 or 3. Indeed, without regressors in decision rule (1.2), these questions cannot be answered unless specific distributional assumptions are invoked.

A major conclusion of this subsection is that structural question 1 or 3 cannot be answered in a fully nonparametric random coefficient model without invoking distributional assumptions. A shifter or regressor in the decision rule is required.

However, it is not obvious that $E(\varepsilon_i|d_i = 1) \neq 0$. In Section 2 we present a model in which enrollment decisions are made in the presence of uncertainty about ε_i (i.e., a person may not know his or her value of ε_i at the time decisions to enroll are made). In this case it is possible that $\alpha^* = \bar{\alpha}$, and if everywhere in Sections 1.2 and 1.3 we replace α with $\bar{\alpha}$, the preceding analysis goes through as before.

With a regressor in the decision rule, and under further conditions presented in Section 3 below, it is possible to estimate α even if $E(\varepsilon_i|d_i = 1) \neq 0$. To do so requires more a priori structure than is required in the fixed coefficient model for α. All of the known consistent estimators of $\bar{\alpha}$ work in a single cross section, but not all of the cross-section estimators listed in Section 1.3.1 will identify $\bar{\alpha}$.[6] For this reason, we do not discuss random coefficient models in Sections 4 and 5.

1.5 *Robustness to nonrandom sampling schemes and contamination bias*

This section discusses the problems of choice-based sampling and contamination bias and indicates their relevance to the problem of estimating the impact of training on the earnings of trainees. It is more productive to defer a detailed discussion of these problems to later sections of the chapter after specific estimators have been presented. Here these concepts are introduced for the fixed coefficient earnings function (1.3) and general approaches to solving these problems are presented.

1.5.1. Choice-based sampling plans. The data available for analyzing the impact of training on earnings are often nonrandom samples. More often they consist of pooled data from two sources: (a) a sample of trainees selected from program records and (b) a sample of nontrainees selected from some national sample. The sampling rule used to generate the nontrainee data is often (simple) random sampling. Typically, trainees are overrepresented in such samples 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).

The problem of choice-based sampling occurs if in the available data the probability of sampling a trainee is not the population probability that an individual is a trainee. In the population let the joint frequency of $(d, \mathbf{Z})$ be $f(d, \mathbf{Z})$. Let $f(d)$ and $f(\mathbf{Z})$ be the population marginal frequencies of d and $\mathbf{Z}$, respectively, and $f(d|\mathbf{Z})$ be the conditional frequency of d given $\mathbf{Z}$.

In a choice-based sample, the rule generating the available sample selects proportion $\phi(d) \neq f(d)$ (see Manski and McFadden, 1981) and selection depends only on training status. Thus as the sample becomes large

so that sampling fluctuations can be ignored, the frequency of the data is

$$h(\mathbf{Z}) = \sum_{d=0,1} f(\mathbf{Z}|d)\phi(d)$$

In a large sample the conditional probability of d given $\mathbf{Z}$ is

$$k(d|\mathbf{Z}) = \frac{f(\mathbf{Z}|d)\phi(d)}{h(\mathbf{Z})}$$

Using Bayes's rule we reach

$$k(d|\mathbf{Z}) = f(d|\mathbf{Z})\frac{\phi(d)}{f(d)}\frac{1}{\sum_{j=0,1} f(j|\mathbf{Z})\frac{\phi(j)}{f(j)}}$$

The sample regression function relating Y_{it} to d_i (sampled in year t) may be written as

$$E(Y_{it}|d_i) = \beta_t + d_i\alpha + \{d_iE(U_{it}|d_i = 1) + (1 - d_i)E(U_{it}|d_i = 0)\} \tag{1.15}$$

In the absence of random sampling, the term in braces does not have mean zero, because the sample proportion of d_i does not converge to $\Pr(d_i = 1)$. Estimators such as the instrumental variable estimator (i) in Section 1.3 for cross-section data that exploit the fact that this term has zero mean will be biased and inconsistent for α if they are mechanically applied to choice-based samples.

The repeated cross-section estimator $\hat{\alpha}_{RC}$ that exploits condition (A-2) is inconsistent when applied to choice-based samples if in different samples the means of the term in braces in (1.15) differ. In this case, (A-2) is not satisfied and $\hat{\alpha}_{RC}$ is inconsistent for α even in a time-homogeneous environment. It is necessary to know the identity of trainees in order to weight the sample back to a sample with proportions of trainees that would be produced by a random sample in order to obtain consistent estimators. Hence one of the advantages of $\hat{\alpha}_{RC}$ is lost if the data are generated by a choice-based sample.

Some of the cross-sectional and longitudinal estimators that are control function estimators are robust to choice-based sampling. Instead of (D-1) we define a subset of control functions such that K_{it} is in the subset if

(D-2) (a) $E(U_{it} - K_{it}|d_i, \mathbf{X}_{it}, K_{it}) = 0$,
(b) ψ is identified.

If (D-2) holds, then by construction the error term in braces in (1.12) has mean zero in any choice-based sample because it has mean zero for each subsample of d_i values, and the error term in any choice-based sample is orthogonal to all the right-hand-side variables in the equation because it is orthogonal to the regressors in each subsample of d_i values. More precisely, using (1.12),

$$\begin{aligned} E(Y_{it}|\mathbf{X}_{it}, d_i, K_{it}) = {} & \mathbf{X}_{it}\boldsymbol{\beta} + d_i\alpha + K_{it} \\ & + d_iE(U_{it} - K_{it}|d_i = 1, \mathbf{X}_{it}, K_{it}) \\ & + (1 - d_i)E(U_{it} - K_{it}|d_i = 0, \mathbf{X}_{it}, K_{it}) \end{aligned}$$

Using (D-2), condition (a),

$$E(Y_{it}|\mathbf{X}_{it}, d_i, K_{it}) = \mathbf{X}_{it}\boldsymbol{\beta} + d_i\alpha + K_{it}$$

Thus α can be consistently estimated under general conditions specified in Section 3. Models that satisfy (A-3) and those that satisfy (A-4) and (1.10) produce K functions satisfying (D-2), and so both repeated cross-section and longitudinal estimators proposed for those models may be applied without modification to data generated from choice-based samples. The robustness of the K function estimators to choice-based sampling plans is a very attractive feature of this class of estimators since most of the available data are choice-based samples.

1.5.2. Contamination bias. The problem of contamination bias arises when the training status of certain individuals is recorded with error. Many control samples such as the Current Population Survey (CPS) or Social Security Work History File do not reveal whether or not persons have received training.

The contamination bias problem is one of measurement error in d_i in equation (1.1) or (1.3). In the analysis of training programs, the population proportion of trainees is known or can be consistently estimated (Barnow, 1983). With this information in hand, consistent estimators for α can be constructed using methods developed by Cochran (1968), Aigner (1973), and the authors in Section 3.8 of this chapter.

2 Prototypical enrollment rules

2.1 *Introduction*

The nature of the stochastic dependence relationships among the regressors and unobservables in (1.1) and (1.2) is critical in designing consistent estimators for α. Of the estimators considered thus far, only $\hat{\alpha}_{RC}$ can produce a consistent estimator for α under any specification of stochastic

dependence among the variables in earnings and enrollment equations (1.1) and (1.2) (provided that the environment is time-homogeneous or sufficiently regular in a sense to be made precise in Section 4.1).

Previous work on evaluating the impact of training on earnings has not specified explicit choice rules that govern program participation. For this reason it is difficult to evaluate the plausibility of proposed estimation procedures when measured against economically appealing models of the life cycle evolution of earnings and of the decision to enroll in training.

This section of the chapter presents several prototypical decision rules that are motivated by economic theory and that are analytically and empirically tractable. The models presented here serve as a framework within which it is possible to evaluate the economic plausibility of the estimators proposed in the remaining sections of this chapter. We consider models in certain and uncertain environments with α both fixed and random.

2.2 *A perfect-foresight model*

A natural starting point is a model of trainee self-selection based on a comparison of the present value of earnings with and without training in an environment of perfect foresight. The earnings function is assumed to be (1.1). For simplicity we assume that training programs accept all applicants. This assumption is relaxed in Section 2.5.

The prospective trainee is assumed to discount all earnings streams by a common discount factor $1/(1+r)$. From (1.1) training raises trainee earnings by α per period. While in training, individual i receives subsidy S_i, which may be negative (so there may be costs of program participation.) Income in training period k is forgone for trainees. To simplify the expressions we assume that people live forever.

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

$$PV_i(0) = \sum_{j=0}^{\infty} \left(\frac{1}{1+r}\right)^j Y_{i,k+j}$$

(Recall that training is an option available only in period k.) The present value of earnings for a trainee is

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

The perfect-foresight present value 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 (1.2),

$$IN_i = PV_i(1) - PV_i(0) = S_i - Y_{ik} + \frac{\alpha}{r} \tag{2.1}$$

Thus

$$d_i = 1 \quad \text{iff} \quad S_i - Y_{ik} + \frac{\alpha}{r} > 0$$

$$d_i = 0 \quad \text{otherwise} \tag{2.2}$$

Recall that Y_{ik} is not observed for trainees. To make (2.2) empirically operational, substitute for Y_{ik} in (2.2) from (1.1) and write

$$S_i = \mathbf{W}_i\boldsymbol{\phi} + \tau_i \tag{2.3}$$

where $\mathbf{W}_i$ is observed by the econometrician and τ_i is not. Collecting terms, we reach

$$d_i = 1 \quad \text{iff} \quad \mathbf{W}_i\boldsymbol{\phi} + \frac{\alpha}{r} - \mathbf{X}_{ik}\boldsymbol{\beta} + \tau_i - U_{ik} > 0$$

$$d_i = 0 \quad \text{otherwise} \tag{2.4}$$

Now $(\tau_i - U_{ik}) = V_i$ in (1.2), and $(\mathbf{W}_i, \mathbf{X}_{ik})$ corresponds to $\mathbf{Z}_i$ in (1.2). Assuming that $(\mathbf{W}_i, \mathbf{X}_{ik})$ is distributed independently of V_i makes (2.4) a standard discrete choice model.

Maintaining the assumption that $E(\mathbf{X}_{it}' U_{it}) = 0$, if

$$E(U_{it} d_i) = 0 \tag{2.5}$$

under general conditions (see White, 1980, or Section 3) least squares consistently estimates $\boldsymbol{\beta}$ and α in (1.1).

If the costs of program participation are independent of U_{it} for all t (so both $\mathbf{W}_i$ and τ_i are independent of U_{it}), (2.5) is satisfied only if the unobservables in period t are (mean) independent of the unobservables in period k so that

$$E(U_{it} | U_{ik}) = 0 \quad \text{for} \quad t > k$$

Whether (2.5) is satisfied hinges on the serial correlation properties of U_{it}. If U_{it} is a moving average of order m, and so

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

where the $\varepsilon_{i,t-j}$ are i.i.d., then for $t - k > m$, (2.5) is satisfied. On the other hand, if U_{it} obeys a first-order autoregressive scheme, (2.5) is not satisfied.

2.3 *A perfect-foresight random coefficients model*

The assignment rule for the case of a perfect-foresight random coefficients model is the same as (2.4) except that α varies in the population. Recalling that $E(\alpha_i) = \bar{\alpha}$ and that $\varepsilon_i = \alpha_i - \bar{\alpha}$, we write

$$d_i = 1 \quad \text{iff} \quad \mathbf{W}_i\boldsymbol{\phi} - \mathbf{X}_{ik}\boldsymbol{\beta} + \frac{\bar{\alpha}}{r} + \tau_i - U_{ik} + \frac{\varepsilon_i}{r} > 0$$

$$d_i = 0 \quad \text{otherwise} \tag{2.6}$$

Even if U_{it} is (mean) independent of U_{ik} and τ_i, so that

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

the composite error term in (2.6) is not mean independent of the error term in earnings function (1.14) because of their common dependence on ε_i for $t > k$.

If there are no regressor variables in the enrollment equation (2.6) (so $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are zero), $E(\varepsilon_i|d_i = 1, \mathbf{X}_{it})$ is constant and $\bar{\alpha}$ cannot be identified but α^* in (1.14) might be identified. If there are regressor variables in $\mathbf{W}_i$ or $\mathbf{X}_{ik}$, $\bar{\alpha}$ might be identified as noted in Section 1.

The random coefficients model captures the key idea in Roy's model of self-selection (1951) that has been revived in recent work by Lee (1978) and Willis and Rosen (1979). In the Roy model it is solely the population variation in X_{ik}, α_i, and U_{ik} that sorts people into training status [so $\tau_i = 0$ and $\mathbf{0} = \mathbf{W}_i$ in (2.6)].

2.4 *Introducing uncertainty*

It is unlikely that prospective trainees know all components of future earnings and the costs and benefits of program participation at the time they make enrollment decisions. More likely the enrollment decision is made in an environment of uncertainty. When risk aversion is ignored the natural generalization of decision rules (2.2) and (2.6) assumes that a prospective trainee compares the expectation of $PV_i(0)$ evaluated at date $k-1$ with the expectation of $PV_i(1)$ evaluated at the same date.

We are thus led to write

$$d_i = 1 \quad \text{iff} \quad E_{k-1}\left[S_i - Y_{ik} + \frac{\alpha_i}{r}\right] > 0$$

$$d_i = 0 \quad \text{otherwise} \tag{2.7}$$

where E_{k-1} denotes the expectation of the argument in brackets conditional on the information available in period $k-1$, and α_i is set to α for the fixed coefficient model.

Introducing uncertainty can sometimes simplify the econometrics of a problem (see, e.g., Zellner et al., 1966). For example, in the random coefficients model suppose that at time $k-1$ individuals do not know the value of α_i they will draw upon completion of training but they know the population distribution of α_i. Suppose further that their best estimate of training impact is the population mean $\bar{\alpha}$. Then in equation (1.14),

$$E(\varepsilon_i|d_i = 1) = 0$$

and so

$$E(\varepsilon_i d_i) = 0$$

and the error component $\varepsilon_i d_i$ creates no new econometric problem that does not appear in the fixed coefficient model.[7] If the only source of uncertainty is in α_i, decision rule (2.2) is identical to decision rule (2.7) provided that all agents have the same estimate of α_i.[8]

In the general case in which future earnings are not known, the optimal forecast rule for Y_{ik} depends on the time series process that generates U_{it}. For example, suppose that

$$U_{it} = \theta_i + v_{it} \tag{2.8}$$

where

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

and where

$$E(\theta_i) = E(\xi_{it}) = 0, \qquad |\rho| < 1$$

$$\text{Var}(\theta_i) < \infty, \qquad \text{Var}(\xi_{it}) < \infty$$

$$E(\xi_{it}\theta_i) = 0 \quad \text{for all} \quad t$$

Suppose that in period $k-1$ agents know current earnings, θ_i, α_i, and all future values of $\mathbf{X}_{it}$, $t \geq k$, but they do not know future values of ξ_{it}. By an application of the standard Wiener–Kolmogorov prediction formula (see, e.g., Sargent, 1979),

$$\begin{aligned} E_{k-1}(Y_{ik}) &= \mathbf{X}_{ik}\boldsymbol{\beta} + \theta_i + \rho(Y_{i,k-1} - \mathbf{X}_{i,k-1}\boldsymbol{\beta} - \theta_i) \\ &= (\mathbf{X}_{ik} - \rho\mathbf{X}_{i,k-1})\boldsymbol{\beta} + \theta_i(1-\rho) + \rho Y_{i,k-1} \end{aligned}$$

If it is further assumed that S_i is known with certainty by the prospective trainee and the specification of S_i is given by (2.3), assignment rule

(2.7) may be written as

$$d_i = 1 \quad \text{iff} \quad \mathbf{W}_i\boldsymbol{\phi} - (\mathbf{X}_{ik} - \rho\mathbf{X}_{i,k-1})\boldsymbol{\beta} + \frac{\bar{\alpha}}{r} - \rho Y_{i,k-1} + \left[\tau_i + \frac{\varepsilon_i}{r} - (1-\rho)\theta_i\right] > 0$$

$$d_i = 0 \quad \text{otherwise} \tag{2.9}$$

In this case, the assignment rule contains $Y_{i,k-1}$ among the $\mathbf{Z}_i$ in (1.2) and $\mathbf{Z}_i$ is not independent of

$$V_i = \tau_i + \frac{\varepsilon_i}{r} - (1-\rho)\theta_i$$

unless $\rho = 1$ or $\theta_i = 0$ and τ_i is distributed independently of $U_{i,k-1}$. Hence cross-section method (ii) listed in Section 1.3 cannot be applied directly without invoking very strong and implausible assumptions.

Note that one implication of decision rule (2.9) is that if τ_i is distributed independently of v_{it}, transitory dips in preprogram earnings make participation in the training program more likely. This specification is thus consistent with the empirical evidence on program enrollment presented by Ashenfelter (1978).

In the absence of uncertainty regarding earnings in period k and with α a fixed known constant, inequality (2.4) characterizes the enrollment decision (setting $U_{ik} = \theta_i + v_{ik}$). In a world of perfect certainty $\mathbf{X}_{i,k-1}$ is not an argument of the decision rule.[9] A general feature of an uncertainty model is that it expands the candidate instrumental variable set [the $\mathbf{Z}_i$ in (1.2)] and so in this sense aids in identification.

With longitudinal earnings data of sufficient length it is possible to test for any assumed time series error structure for U_{it} (see, e.g., MaCurdy, 1982). However, such tests cannot reveal which components of U_{ik} are known to the agent and which are not, nor can they reveal the nature of the dependence between τ_i and U_{it}. Extending assignment rule (2.4) to account for uncertainty requires additional assumptions about forecasting rules used by prospective trainees and the information sets available to them. Knowledge of the time series process generating the unobservables in earnings data does not shed light on the information set confronting an agent. This injects an extra element of arbitrariness into the analysis that is not present in a certainty setting.

The analysis presented for error structure (2.8) carries over to more general models. In general, the error structure of V_i [in (1.2)] is induced in part by the expectations mechanism assumed to be used by the agent. In the example just presented V_i is correlated with future U_{it} via their common dependence on θ_i and $v_{i,k-1}$. The covariance structure of V_i induced by the assumed estimators is presented below in Section 5.

2.5 *Multiple selection rules*

For convenience we have assumed that only one assignment rule governs program participation; any individual who desires to enroll in training is free to do so. In fact, satisfaction of inequality (2.1) may be a necessary but not sufficient condition for program participation; a candidate trainee may also have to be selected by a program administrator.

In principle all of our analysis carries over to a more general case in which multiple selection governs program participation. Let

$$\{IN_l\}_{l=1}^{L}$$

be a set of L index functions all of which must be positive for a person to be enrolled in training. For example, IN_1 may be $PV(1) - PV(0)$, IN_2 may be the index function for a program administrator.

If we define

$$IN^* = \min(IN_1, \ldots, IN_L)$$

we may define

$$d_i = 1 \quad \text{if} \quad IN^* > 0$$

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

Replacing IN with IN^*, all of the preceding analyses go through as before so that with a suitable change in notation our analysis can be readily generalized to a multiple selection rules case.[10] However, for large L, explicit formulas for the components of IN^* are not available except in special cases. For simplicity we assume that $L = 1$.[11]

3 Cross-section methods

3.1 *Introduction*

In this section we present cross-section methods for consistently estimating α in (1.1) when (A-1) characterizes the data so that the assignment of persons to training is nonrandom. Our discussion proceeds in the following way.

We initially assume access to one postprogram cross section for a random sample (or exogenously stratified sample) of the population some fraction of which has participated in training. Six different consistent estimators of α are presented corresponding to six different types of assumptions about the earnings function (1.1) and the enrollment rule (1.2). The

assumption sets are presented in decreasing order of the generality of their content as far as this is possible. However, all of the estimators cannot be ranked in this fashion because they rely on nonoverlapping assumption sets.

We then consider consistent estimation of a random coefficients model. We next examine the robustness of the proposed estimators to choice-based sampling schemes and errors in measuring d_i. The section concludes with a summary and discussion of the results.

Except in our discussion of identification through assumptions about the distribution of U_{it}, throughout most of this section we make the following assumption:

(A-5) There is at least one nondegenerate regressor in $\mathbf{Z}_i$ in decision rule (1.2) with a nonzero coefficient.

As discussed in Section 1, without this assumption α cannot be identified from a cross section without invoking a distributional assumption.

We also make the following additional technical assumptions about the data available in cross section t:

(A-6)
(a) The earnings function is (1.1).
(b) The enrollment decision is governed by (1.2).
(c) $\{\mathbf{X}_{it}, \mathbf{Z}_i, V_i, U_{it}\}$ is an independent sequence with respect to i.
(d) $E(\mathbf{X}_{it}' U_{it}) = 0$ for all i.
(e) $E|X_{ijt} U_{it}|^{1+\delta} < \Delta < \infty$ for some $\delta > 0$, where $j = 1, \ldots, M$ denotes an element of the M vector, $\mathbf{X}_{it}$, for all i.
(f) $E|X_{ijt}^2|^{1+\delta} < \Delta < \infty$ for some $\delta > 0, j = 1, \ldots, M$ for all i.
(g) $E|d_i U_{it}|^{1+\delta} < \Delta < \infty$ for some $\delta > 0$ for all i.
(h) Array $\mathbf{X}_{it}$, d_i into vector $\mathbf{J}_i = (\mathbf{X}_{it}, d_i)$. In this notation,

$$\bar{J}_{I_t} \equiv E\left(\sum_{i=1}^{I_t} \frac{\mathbf{J}_i' \mathbf{J}_i}{I_t}\right)$$

has determinant $\det \bar{J}_{I_t} > \delta > 0$ for all I_t sufficiently large.

Assumption (A-6) coupled with the assumption $E(d_i U_{it}) = 0$ ensures that ordinary least squares is (strongly) consistent for α in (1.1) (see *White*, 1980). Assumption (d) rules out lagged values of the dependent variable in $\mathbf{X}_{it}$ if the U_{it} are serially correlated. If (1.1) includes lagged values, we write out the reduced form expression for (1.1) and require that it satisfy (A-6). The reduced form is used in the ensuing analysis. The underlying samples are not restricted to be simple random samples – samples stratified on the basis of exogenous variables may also satisfy (A-6).

If the analyst has access to simple random samples, assumptions (A-6e) and (A-6f) may be eliminated and δ can be set to 0 in (A-6g). The stronger conditions (A-6) produce sufficient conditions for strong consistency for models estimated on samples in which the observations are independently

but not identically distributed. Such samples are in wide use (e.g., exogenously stratified samples). Note further that conditions (A-6), although conventional and familiar (see, e.g., White, 1984), are overly strong. Weak consistency is all that is required in econometric analysis. For the sample sizes likely to be encountered in practice (500 or more independent observations), asymptotic theory should produce a reliable guide to the performance of estimators.

3.2 *The instrumental variables estimator*

The instrumental variables estimator is the least demanding in the a priori conditions that must be satisfied for its use. It requires in addition to (A-5) and (A-6) the following assumptions:

(A-7)
(a) There is at least one variable in $\mathbf{Z}_i$, Z_i^e, with a nonzero γ coefficient in (1.2), such that for some known transformation of Z_i^e, $g(Z_i^e)$, $E[U_{it} g(Z_i^e)] = 0$.
(b) Array $\mathbf{X}_{it}$, $g(Z_i^e)$ into a vector $\mathbf{J}_{it}^* = [\mathbf{X}_{it}, g(Z_i^e)]$. In this notation, $E[\sum_{i=1}^{I_t} (\mathbf{J}_{it}^{*\prime} \mathbf{J}_{it}/I_t)]$ has full rank uniformly in I_t for I_t sufficiently large.
(c) Replacing $\mathbf{X}_{it}$ by $\mathbf{J}_{it}^*$, (A-6) holds except for (A-6h).

With these assumptions, the instrumental variable estimator

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\alpha} \end{pmatrix}_{IV} = \left(\sum_{i=1}^{I_t} \frac{\mathbf{J}_{it}^{*\prime} \mathbf{J}_{it}}{I_t}\right)^{-1} \sum_{i=1}^{I_t} \frac{\mathbf{J}_{it}^{*\prime} Y_{it}}{I_t}$$

is consistent for $\begin{pmatrix} \boldsymbol{\beta} \\ \alpha \end{pmatrix}$. Thus α is identified if there is a regressor in (1.2) that satisfies (A-7).

It is important to notice how weak these conditions are. The functional form of the distribution of V_i need not be known. $\mathbf{Z}_i$ need not be distributed independently of V_i. Only some function of one of the nondegenerate arguments of $\mathbf{Z}_i$ is required to satisfy (A-7a). Moreover, in principle, $g(Z_i^e)$ may be a nonlinear function of variables appearing in $\mathbf{X}_{it}$ as long as (A-7b) is satisfied. Except for the linear probability model, $E(d_i | \mathbf{Z}_i)$ is nonlinear in the arguments of $\mathbf{Z}_i$, and so rank condition (b) is likely to be satisfied. Assuming that (A-7) is satisfied, we can conduct a test for the endogeneity of d_i in (1.1) using the Durbin–Wu–Hausman test (Durbin, 1954; Wu, 1973, 1983; Hausman, 1978).

In certainty decision rule (2.4) the list of potential instruments includes (transformations of) the costs of participation ($\mathbf{W}_i$) and the regressors explaining earnings in the training period ($\mathbf{X}_{ik}$). Only one variable in this list is required to identify α in (1.1). If there are a variety of instruments so that α is overidentified, standard methods can be used to produce more efficient estimators (see, e.g., White, 1984).

3.3 *Procedures when the functional form of F is known or can be consistently estimated*

Procedures when the functional form of F is known or can be consistently estimated require that assumptions (A-5) and (A-6) be strengthened in the following way:

(A-8) (a) $\mathbf{Z}_i$ is distributed independently of V_i.
(d) $E(U_{it}|\mathbf{Z}_i, \mathbf{X}_{it}) = 0$.
(c) Array $\mathbf{X}_{it}$, $E(d_i|\mathbf{Z}_i)$ into vector $\mathbf{J}_i = [\mathbf{X}_{it}, E(d_i|\mathbf{Z}_i)]$. In this notation,

$$\tilde{\tilde{J}}_{I_t} \equiv E\left(\sum_{i=1}^{I_t} \frac{\mathbf{J}_i'\mathbf{J}_i}{I_t}\right)$$

has determinant $\det \tilde{\tilde{J}}_{I_t} > \delta > 0$, as $I_t \to \infty$.
(d) $E(U_{it}|\mathbf{Z}_i, \mathbf{X}_{it}) = 0$.
(e) Distribution function F is known (up to a finite number of parameters) or can be consistently estimated, and γ is identified.

Cosslett (1983) demonstrates that if (a) is satisfied and $\mathbf{Z}_i$ includes at least one "continuous valued regressor" that takes values in an interval and if the $\mathbf{Z}_i$ are i.i.d. nondegenerate random variables, F can be consistently nonparametrically estimated.

Note that assumption (A-8d) is not implied by (d) in (A-6). Assumption (b) in (A-8) is not implied by any set of conditions in (A-6).

From assumptions (A-8b) and (A-8d), the conditional expectation of Y_{it} given $\mathbf{X}_{it}$ and $\mathbf{Z}_i$ may be written as

$$E(Y_{it}|\mathbf{X}_{it}, \mathbf{Z}_i) = \mathbf{X}_{it}\boldsymbol{\beta} + E(d_i|\mathbf{Z}_i)\alpha \tag{3.1}$$

From assumption (A-8a), this expectation may be written as

$$E(\mathbf{Y}_{it}|\mathbf{X}_{it}, \mathbf{Z}_i) = \mathbf{X}_{it}\boldsymbol{\beta} + [1 - F(-\mathbf{Z}_i\gamma)]\alpha \tag{3.2}$$

Given (A-5), (A-6), and (A-8), α can be consistently estimated by the following two-stage procedure.

(A) Use discrete choice analysis to estimate γ from data on the training decision (see Amemiya, 1981). Form $F(-\mathbf{Z}_i\hat{\gamma})$ and run a linear regression of Y_{it} on $\mathbf{X}_{it}$ and $F(-\mathbf{Z}_i\hat{\gamma})$.

Standard errors must be adjusted to account for estimation error in $\hat{\gamma}$ (see Heckman, 1979, and Amemiya, 1983). If Cosslett's procedure can be applied, no parametric position need be taken with regard to F.

An alternative procedure that identifies α under the same assumptions is:

(B) Estimate (3.2) directly by nonlinear least squares.

Method (B) can be implemented without assuming a parametric form for F using, for example, Gallant's (1981) procedures expanding F in terms of a Fourier series in $\mathbf{Z}_i\gamma$.

Method (B) is more general than (A) in the following sense. It is possible to relax (A-8a, b, and e) and still recover α if (A-8c) is satisfied replacing $E(d_i|\mathbf{Z}_i)$ by $E(d_i|\mathbf{Z}_i, \mathbf{X}_{it})$ and if $\boldsymbol{\beta}$, α and the independent parameters of $E(d_i|\mathbf{X}_{it}, \mathbf{Z}_i)$ can be recovered from the regression function

$$E(Y_{it}|\mathbf{X}_{it}, \mathbf{Z}_i) = \mathbf{X}_{it}\boldsymbol{\beta} + E(d_i|\mathbf{Z}_i, \mathbf{X}_{it})\alpha \tag{3.3}$$

Note further that even if (A-8a) is violated so that γ cannot be consistently estimated, $F_i = F(-\mathbf{Z}_i\hat{\gamma})$ is a valid instrument for d_i in (1.1) provided that the conditions of (A-5)–(A-7) are satisfied. Even if γ can be consistently estimated, use of F_i as an instrument is an alternative to methods (A) and (B) of this section.

Because the assumptions embodied in (A-8) are more stringent than those presented in (A-7), they are less likely to be satisfied by models generated by the decision rules in Section 2.1. For example, (A-8a) requires that $\mathbf{W}_i$ and $\mathbf{X}_{ik}$ be jointly independent of $\tau_i - U_{ik}$ in perfect-foresight decision rule (2.4), whereas the instrumental variable estimator requires no such assumption. (A-8b) is stronger than anything required for the instrumental variable estimator to be consistent for α. Assumption (A-8b) will not be satisfied for decision rules (2.2) and (2.7) if data on $\mathbf{X}_{ik}$ are not available. Assumption (A-8d) is not satisfied by uncertainty decision rule (2.9) because unless $\rho = 1$ or $\theta_i = 0$, $Y_{i,k-1}$ is not distributed independently of θ_i. This assumption can be satisfied if $Y_{i,k-1}$ is replaced with a reduced expression in terms of lagged $\mathbf{X}_{ik}$.

3.4 *Cross-section control function estimators*

We now consider the cross-section version of the control function estimators introduced in definitions (D-1) and (D-2). Here we use the strong form of these conditions – (D-2). As noted in Section 1, the strong form is robust to choice-based sampling.

Using (1.1) we write

$$E(Y_{it}|\mathbf{X}_{it}, d_i, \mathbf{Z}_i) = \mathbf{X}_{it}\boldsymbol{\beta} + d_i\alpha + E(U_{it}|\mathbf{X}_{it}, d_i, \mathbf{Z}_i) \tag{3.4}$$

If $E(U_{it}|\mathbf{X}_{it}, \mathbf{Z}_i)$ is known or its parameters can be consistently estimated from regression (3.4) applied to the available cross-section sample, then

$$E(U_{it}|\mathbf{X}_{it}, d_i, \mathbf{Z}_i) = K_{it}$$

is a control function for (1.1).

In the literature (Heckman, 1976, 1979) the following *sufficient* conditions in addition to (A-5) and (A-6) are invoked to produce a control function:

(A-9) (a) (A-8a).
(b) $E(U_{it}|\mathbf{X}_{it}, d_i, \mathbf{Z}_i) = E(U_{it}|d_i, \mathbf{Z}_i)$.
(c) The joint density of (U_i, V_i) denoted $h_i(U_{it}, V_i|\chi)$ is known up to a finite set of parameters χ. Typically it is assumed to be bivariate normal.[12] Elements of χ are not functionally dependent on (α, β, γ).
(d) (A-8e).
(e) In the population $[\mathbf{X}_{it}, d_i, \partial \mathbf{E}(U_{it}|d_i, \mathbf{Z}_i)/\partial\chi]$ is a vector of nondegenerate random variables; that is, the joint distribution is nonsingular for the true value χ that generates $E(U_{it}|d_i, \mathbf{Z}_i)$.

As a consequence of the assumptions, it is possible to fit (3.4) by nonlinear least squares and secure identification of α.[13]

Note that the role of (A-9c) is to produce an explicit functional form for $E(U_{it}|d_i, \mathbf{Z}_i)$:

$$E(U_{it}|d_i = 0, \mathbf{Z}_i) = \frac{\int_{-\infty}^{\infty} t_1 \int_{-\infty}^{-\mathbf{Z}_i\gamma} h_i(t_1, t_2|\chi)\, dt_1\, dt_2}{\int_{-\infty}^{-\mathbf{Z}_i\gamma} h_i(t_2)\, dt_2} \tag{3.5}$$

In principle, one could dispense with (A-9c) and postulate a functional form for (3.5) directly. Any assumed functional form for $E(U_{it}|d_i = 0, \mathbf{Z}_i)$ should reflect the fact that the conditional mean of U_{it} is a function of $\Pr(d_i = 0|\mathbf{Z}_i)$ and χ.

Thus by virtue of (A-8a),

$$\Pr(d_i = 0|\mathbf{Z}_i) = F(-\mathbf{Z}_i\gamma)$$

and so

$$\mathbf{Z}_i\gamma = -F^{-1}(\Pr(d_i = 0|\mathbf{Z}_i))$$

Substituting in (3.5),

$$E(U_{it}|d_i = 0, \mathbf{Z}_i) = \frac{\int_{-\infty}^{\infty} t_1 \int_{-\infty}^{F^{-1}[\Pr(d_i = 0|\mathbf{Z}_i)]} h_i(t_1, t_2|\chi)\, dt_1\, dt_2}{\Pr(d_i = 0|\mathbf{Z}_i)} \tag{3.6}$$

Assuming that h_i is differentiable to all orders, it is possible to express $E(U_{it}|d_i = 0, \mathbf{Z}_i)$ as a power series in $\Pr(d_i = 0|\mathbf{Z}_i)$ and the relevant components of χ. Given (A-9e), $\Pr(d_i = 0|\mathbf{Z}_i)$ can be estimated from a discrete choice analysis of the training decision.[14]

There are three distinct estimators that exploit the additional information assumed in (A-9).

3.4.1. The two-stage method. This is developed in Heckman (1976, 1978, 1979).

(i) Estimate $E(d_i|\mathbf{Z}_i)$ by discrete choice analysis.
(ii) Exploit (3.6) using estimated $\Pr(d_i = 0|\mathbf{Z}_i)$ in place of actual values. If $h_i(U_{it}, V_i)$ is bivariate normal, then the corresponding expression for $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ from (3.6) is linear in certain ratios of elements of χ.
(iii) Regress Y_{it} on $\mathbf{X}_{it}$, d_i, and $E(U_{it}|d_i, \mathbf{Z}_i)$, where the final expression is known up to a finite set of parameters in χ.

The standard errors must be adjusted to account for the estimated regressor (see Heckman, 1979, Amemiya, 1983).

3.4.2. Direct nonlinear regression. This method is suggested in Barnow, Cain, and Goldberger (1980). Estimate (3.4) directly without estimating $E(U_{it}|d_i, \mathbf{Z}_i)$ in a first stage.

3.4.3 Maximum likelihood. Application of this method to models with dummy endogenous variables is discussed in Heckman (1976, 1978) for models with normal error terms. Provided that (A-9c) is true, this procedure produces more efficient estimators.

Note that 3.4.1 and 3.4.2 do not require (A-9c). All that need be known is $E(U_{it}|d_i, \mathbf{Z}_i)$ up to a finite set of parameters provided that the other conditions are satisfied. In this sense 3.4.1 and 3.4.2 are less demanding methods. Note further that assumptions (A-9a) and (A-9b) can be relaxed. If $E(U_{it}|d_i, \mathbf{X}_{it}, \mathbf{Z}_i)$ is known up to a finite set of parameters and it is the case that when $E(U_{it}|d_i, \mathbf{X}_{it}, \mathbf{Z}_i)$ replaces $E(U_{it}|d_i, \mathbf{Z}_i)$ (A-9e) is satisfied, α is still identified. Mincer and Jovanovic (1981) invoke such assumptions in their analysis of job turnover.

In principle, if $E(U_{it}|d_i, \mathbf{X}_{it}, \mathbf{Z}_i)$ is known up to a finite set of parameters and *only* (A-5), (A-6), and (A-9e) (suitably modified) are true, α can be identified without an instrument (in the sense of Section 3.2) and without knowledge of $E(d_i|\mathbf{Z}_i)$ as is assumed in Section 3.3. In such a case identification is secured solely by the assumed a priori functional form. However, given such assumptions, it is in principle possible to devise consistent control function estimators of α for all of the fixed coefficient enrollment models considered in Section 2.

The additional assumptions invoked in this section are purely statistical in nature. Accordingly appeal to the prototypical decision rules of Section 2 offers little guidance in the choice of estimator.

3.5 *Controlling for selection on observables*

The condition for selection bias

(A-1) $E(U_{it}d_i) \neq 0$

can arise for one of two distinct reasons: (a) dependence between V_i and U_{it} or (b) dependence between $\mathbf{Z}_i$ and U_{it}. Under assumptions stated in (3.2), the instrumental variables (IV) estimator is consistent for α in either case. The procedures suggested in Section 3.3 abstract from dependence (b) [see assumption (A-8d)] whereas in principle the control function estimators of Section 3.4 are consistent for α when (A-1) occurs for either reason. Many of the commonly used estimators implicitly assume that there is no stochastic dependence between $\mathbf{Z}_i$ and U_{it}.

In this subsection, we consider the case when (A-1) occurs solely because of (b). Barnow, Cain, and Goldberger (BCG) (1980) first analyzed this case. Ashenfelter (1978) presents empirical evidence that enrollment into training depends on earnings in period $k-1$. If $\rho = 1$ in uncertainty decision rule (2.9) (or if $\theta_i = 0$ for all i), if τ_i is distributed independently of U_{it}, and if ε_i is unknown with mean zero at the time of enrollment, then selection occurs because of reason (b) provided that $Y_{i,k-1}$ is observed.

The BCG procedure is based on the following assumptions: (A-5), (A-6), and

(A-10)
(a) $E(U_{it}|\mathbf{Z}_i, \mathbf{X}_{it}, d_i) = E(U_{it}|\mathbf{Z}_i) \neq 0$ for some $\mathbf{Z}_i$.
(b) The functional form of $E(U_{it}|\mathbf{Z}_i)$ is known up to a finite vector of parameters ω. [Thus we write $E(U_{it}|\mathbf{Z}_i, \omega)$.]
(c) In the population $[\mathbf{X}_{it}, d_i, \partial \mathbf{E}(U_{it}|\mathbf{Z}_i, \omega)/\partial\omega]$ is a nondegenerate vector of random variables (i.e., the joint distribution is nonsingular for the true values of ω).

As a consequence of (A-10a) we may write

$$E(Y_{it}|\mathbf{X}_{it}, d_i, \mathbf{Z}_i) = \mathbf{X}_{it}\boldsymbol{\beta} + d_i\alpha + E(U_{it}|\mathbf{Z}_i, \omega)$$

As a consequence of the assumptions, α is consistently estimated by (possibly nonlinear) regression methods.[15] BCG assume the special functional form

$$E(U_{it}|\mathbf{Z}_i, \omega) = \mathbf{Z}_i\omega$$

although this particular specification is not essential to their approach. Assumption (A-10c) in this case rules out perfect collinearity among $\mathbf{X}_{it}$, d_i, and $\mathbf{Z}_i$ [and thus excludes a linear probability model for $\dot{E}(d_i|\mathbf{Z}_i)$ if $\mathbf{Z}_i$ lies in the column space of $\mathbf{X}_{it}$].

This estimator is another example of a control function estimator:

$$K_{it} = E(U_{it}|\mathbf{Z}_i, \omega)$$

The natural generalization of this control function and the one presented in Section 3.4 is

$$K_{it} = E(U_{it}|d_i, \mathbf{Z}_i, \mathbf{X}_{it}, \omega)$$

Provided (A-5) and (A-6) are true and (A-10) is appropriately modified {in particular, (A-10a) is dropped, (A-10b) is modified to include $\mathbf{X}_{it}$ and d_i in the conditioning set, and (A-10c) is replaced with the requirement that $[\mathbf{X}_{it}, d_i, \partial \mathbf{E}(U_{it}|d_i, \mathbf{Z}_i, \mathbf{X}_{it}, \omega)/\partial\omega]$ is a vector of full column rank in the population}, α is identified.

Selection solely on the basis of observables is implausible in the light of the decision rules presented in Section 2. In both certainty and uncertainty settings, this occurs only if unmeasured cost component τ_i is distributed independently of U_{it} and U_{it} is independent of U_{ik} (the certainty case) or U_{it} is independent of $E_{k-1}(U_{ik})$ (the uncertainty case).

3.6 *Identification through distributional assumptions about U_{it}*

If no regressor appears in decision rule (1.2), the estimators presented in the preceding sections do not consistently estimate α unless additional restrictions are imposed. Heckman (1978) demonstrates that if (U_{it}, V_i) is normally distributed, α is identified even if there is no regressor in enrollment rule (1.2). His conditions are overly strong.

In this section we demonstrate that if U_{it} has zero third and fifth 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} (assuming the first five moments exist) but it is weaker than either. The fact that α can be identified by invoking distributional assumptions about U_{it} is an instance of 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 establish that under the following assumptions, α in (1.3) is identified even if (A-5) is not satisfied:

(A-11)
(a) $E(U_{it}^3) = 0$.
(b) $E(U_{it}^5) = 0$.
(c) The earnings function is (1.3), so there are no regressors.
(d) $\{U_{it}, V_i\}$ is i.i.d.

The assumption that no regressor appears in the enrollment rule or in the earnings function is made only to simplify the initial analysis. We relax this assumption below. The i.i.d. assumption is made only to simplify the proofs. Its relaxation involves only minor changes but adds to the notational complication and so is not done here.

A method-of-moments estimator exploits these assumptions. We find $\hat{\alpha}$ that equates the sample analogs of $E(U_{it}^3)$ and $E(U_{it}^5)$ to zero. The proposed estimator solves

$$\frac{1}{I_t}\sum_{i=1}^{I_t}[(Y_{it}-\bar{Y}_t)-\hat{\alpha}(d_i-\bar{d}_t)]^3=0$$

and

$$\frac{1}{I_t}\sum_{i=1}^{I_t}[(Y_{it}-\bar{Y}_t)-\hat{\alpha}(d_i-\bar{d}_t)]^5=0$$

where, as before, "$^-$" denotes sample mean. By virtue of Slutsky's theorem we may replace sample means with population means in these expressions for the purpose of establishing consistency. Thus we propose the criterion

$$\frac{1}{I_t}\sum_{i=1}^{I_t}[(Y_{it}-\mu_t)-\hat{\alpha}(d_i-p)]^3=0 \tag{3.7a}$$

and

$$\frac{1}{I_t}\sum_{i=1}^{I_t}[(Y_{it}-\mu_t)-\hat{\alpha}(d_i-p)]^3=0 \tag{3.7b}$$

where $E(Y_{it})=\mu_t$ and $E(d_i)=p$.

For fixed $\hat{\alpha}$, (3.7a) converges in probability to

$$\begin{aligned}&\operatorname*{plim}_{I_t\to\infty}\frac{1}{I_t}\sum_{i=1}^{I_t}[(Y_{it}-\mu_t)-\hat{\alpha}(d_i-p)]^3\\&=\{\alpha^3(p)(1-p)(1-2p)\quad+3\alpha^2E(U_{it}|d_i=1)(1-2p)p+3\alpha[E(U_{it}^2|d_i=1)p-p\sigma_u^2]\}\\&\quad-3\tilde{\alpha}[\alpha^2(p)(1-p)(1-2p)+\ 2\alpha E(U_{it}|d_i=1)(1-2p)p+\quad E(U_{it}^2|d_i=1)p-p\sigma_u^2]\\&\quad+3\tilde{\alpha}^2[\alpha(p)(1-p)(1-2p)+\quad E(U_{it}|d_i=1)(1-2p)p]\\&\quad-\tilde{\alpha}^3[p(1-p)(1-2p)]\end{aligned} \tag{3.7a'}$$

where $E(U_{it}^2)=\sigma_u^2$ and where $\tilde{\alpha}=\operatorname{plim}\hat{\alpha}$. Setting $\tilde{\alpha}=\alpha$ and adding up the elements in each column establishes that there exists one root of (3.7a) that is consistent for α.

The other two roots of equation (3.7a) converge to

$$\operatorname{plim}\hat{\alpha}=\alpha+\frac{3}{2}\frac{E(U_{it}|d_i=1)}{1-p}\pm\frac{1}{2}\sqrt{9\left[\frac{E(U_{it}|d_i=1)}{1-p}\right]^2+\frac{12[\sigma^2-E(U_{it}^2|d_i=1)]}{(1-p)(1-2p)}}$$

for $p\neq\frac{1}{2}$. When $p=\frac{1}{2}$, (3.7a) converges to a linear equation in $\tilde{\alpha}$ that is consistent for α provided that $E(U_{it}^2|d_i=1)\neq E(U_{it}^2)$.

The fact that a consistent root of (3.7a) exists is empirically useless. Except in the case when $p=\frac{1}{2}$, we do not know which of the three roots is the consistent root unless the argument inside the square root is negative so that the inconsistent roots are complex conjugates. Nothing in the problem restricts this expression to be negative. A sufficient condition for the existence of complex roots is that selection increases the variance of U_{it} [i.e., $\operatorname{Var}(U_{it}|d_i=1)>\operatorname{Var}(U_{it})$].

In order to pick the consistent root in the general case it is necessary to use the higher moment restriction (A-11b). We now establish that (a) there is one consistent root of (3.7b) and (b) the inconsistent roots of (3.7b) do not converge to the inconsistent roots of (3.7a). Thus in large samples it is possible to detect the consistent root of these equations. A consistent estimator of α is the value of $\hat{\alpha}$ that sets (3.7a) and (3.7b) as close to 0 as possible in a suitably defined metric.

In order to establish this claim, it is helpful to simplify the notation by defining

$$\begin{aligned}c_1&=E(d_i-p)^5&&=p(1-5p+10p^2-10p^3+4p^4)\\c_2&=E[(d_i-p)^4U_{it}]&&=p(1-(4p-6p^2+4p^3)E(U_{it}|d_i=1))\\c_3&=E[(d_i-p)^3U_{it}^2]&&=p(E(U_{it}^2|d_i=1)(1-3p+3p^2)-p^2\sigma_u^2)\\c_4&=E[(d_i-p)^2U_{it}^3]&&=p(E(U_{it}^3|d_i=1)(1-2p))\\c_5&=E[(d_i-p)U_{it}^4]&&=p(E(U_{it}^4|d_i=1)-E(U_{it}^4))\end{aligned}$$

In this notation, it is straightforward but tedious to establish that

$$\begin{aligned}&\operatorname*{plim}_{I_t\to\infty}\frac{1}{I_t}\sum_{i=1}^{I_t}[(Y_{it}-\mu_t)-\hat{\alpha}(d_i-p)]^5\\&=\quad\{\alpha^5c_1+5\alpha^4c_2+10\alpha^3c_3+10\alpha^2c_4+5\alpha c_5\}\\&\quad-5\tilde{\alpha}\{\alpha^4c_1+4\alpha^3c_2+\ 6\alpha^2c_3+\ 4\alpha c_4+\ c_5\}\\&\quad+10\tilde{\alpha}^2\{\alpha^3c_1+3\alpha^2c_2+\ 3\alpha c_3+\ c_4\ \}\\&\quad-10\tilde{\alpha}^3\{\alpha^2c_1+\ 2\alpha c_2+\ c_3\ \}\\&\quad+\ 5\tilde{\alpha}^4\{\ \alpha c_1+\ c_2\ \}\\&\quad-\ \tilde{\alpha}^5\{\ c_1\ \}\end{aligned} \tag{3.7b'}$$

Setting $\tilde{\alpha}=\alpha$ and adding down each column demonstrates the existence of a consistent root (3.7b'). However, there are four other roots of this

equation, all of which may be distinct and real. We now establish that the other four roots of (3.7b′) are distinct from the inconsistent roots of (3.7a′).

To demonstrate that this is so it is helpful to define some additional notation.

$$b_3 = \alpha - 5\left(\alpha + \frac{c_2}{c_1}\right)$$

$$b_2 = 10\left(\alpha^2 + 2\frac{\alpha c_2}{c_1} + \frac{c_3}{c_1}\right) + \alpha\left[\alpha - 5\left(\alpha + \frac{c_2}{c_1}\right)\right]$$

$$b_1 = \alpha\left\{10\left(\alpha^2 + 2\frac{\alpha c_2}{c_1} + \frac{c_3}{c_1}\right) + \alpha\left[\alpha - 5\left(\alpha + \frac{c_2}{c_1}\right)\right]\right\} - 10\left(\alpha^3 + 3\frac{\alpha^2 c_2}{c_1} + 3\frac{\alpha c_3}{c_1} + \frac{c_4}{c_1}\right)$$

$$b_0 = \alpha^4 + 5\frac{\alpha^3 c_2}{c_1} + 10\frac{\alpha^2 c_3}{c_1} + 10\frac{\alpha c_4}{c_1} + 5\frac{c_5}{c_1}$$

With this notation in hand, factor the right-hand side of (3.7b′) into a lower-order quartic multiplied by the factor $\tilde{\alpha} - \alpha$:

$$(\tilde{\alpha} - \alpha)\{\tilde{\alpha}^4 + \tilde{\alpha}^3 b_3 + \tilde{\alpha}^2 b_2 + \tilde{\alpha} b_1 + b_0\}$$

Let r denote either of the roots of the quadratic equation given below (3.7a)′. Here r is a root of the quartic if and only if

$$r^4 + r^3 b_3 + r^2 b_2 + r b_1 + b_0 = 0$$

In general, this equation cannot be satisfied because r does not depend on $E(U_{it}^3 | d_i = 1)$, $E(U_{it}^4 | d_i = 1)$, or $E(U_{it}^4)$ whereas b_1 and b_0 do. The formal condition for distinct roots for the two equations is that the parameters of the model are such that for both possible values of r (given the parameters of the model)

$$r^4 + r^3 b_3 + r^2 b_2 + r b_1 + b_0 \neq 0$$

In general, this condition will be satisfied if the third and fourth moments of the U_{it} can be freely specified.

The operational version of the estimator selects a common α that makes (3.7a) and (3.7b) as close to zero in a suitable metric. One obvious choice of metric is a least squares criterion summing squared deviations of the left-hand sides of (3.7a) and (3.7b) from equality with zero.

If regressors appear in the earnings function, the method-of-moments procedure proposed above can be modified in the following way. In place of (A-11) we write

(A-11′) (a) $E(U_{it}^3) = 0$.
(b) $E(U_{it}^5) = 0$.
(c) Assumptions (A-6) are satisfied except (A-6c).
(d) Assumption (A-6c) is strengthened to read that the vectors of variables in that assumption are i.i.d.

For each value of $\hat{\alpha}$, compute $\boldsymbol{\beta}(\hat{\alpha})$ by regressing

$$(Y_{it} - \hat{\alpha} d_i) \quad \text{on} \quad \mathbf{X}_{it}$$

Define $\tilde{Y}_{it}$ as the value of Y_{it} with the value of $\mathbf{X}_{it}$ removed:

$$\tilde{Y}_{it} = (Y_{it} - \bar{Y}_t) - (\mathbf{X}_{it} - \bar{\mathbf{X}}_t)\boldsymbol{\beta}(\hat{\alpha})$$

Then, under the conditions of (A-11′), there is a unique consistent root that solves

$$\frac{1}{I_t}\sum_{i=1}^{I_t}[\tilde{Y}_{it} - \hat{\alpha}(d_i - p)]^3 = 0 \tag{3.7a″}$$

and

$$\frac{1}{I_t}\sum_{i=1}^{I_t}[\tilde{Y}_{it} - \hat{\alpha}(d_i - p)]^5 = 0 \tag{3.7b″}$$

The proof is straightforward and is omitted.

The moment estimation proposed in this section is just one example of an entire class of consistent estimators for α that do not require a regressor in enrollment rule (1.2). Other restrictions on moments (e.g., functional restrictions between the second and fourth moments of U_{it} that are implied by a normality assumption) can be exploited to devise consistent estimators for α.

3.7 *Estimation in the random coefficients model*

In place of random coefficients model (1.13) we write the more general model with regressors

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + d_i\alpha_i + U_{it} \tag{3.8}$$

where $E(\alpha_i) = \bar{\alpha} < \infty$, $\varepsilon_i = \alpha_i - \bar{\alpha}$ and $E(\varepsilon_i) = 0$. To focus on essential issues we continue to assume that $\boldsymbol{\beta}$ is a fixed parameter vector. We strengthen assumption (A-6c) to read

(A-6c′) $\{\mathbf{X}_{it}, \mathbf{Z}_i, V_i, U_{it}, \varepsilon_i\}$ is an independent sequence with respect to i

and (A-6a) is modified appropriately. In addition, $U_{it} + d_i\varepsilon_i$ replaces U_{it} in (A-6d)–(A-6g). If (A-5) does not hold and (A-6) (as augmented) holds and

if the (V_i, ε_i) are i.i.d., then it is not possible to identify $\bar{\alpha}$ without further prior information. Instead, only

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

can be identified without invoking distributional assumptions or other prior information.

If there is a regressor in (1.2), it is in principle possible to identify $\bar{\alpha}$. We write

$$E(\varepsilon_i | d_i = 1, \mathbf{Z}_i) = \phi(\mathbf{Z}_i) \tag{3.9}$$

A new econometric problem arises that has not previously appeared. To state it most clearly, rewrite (3.8) as

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + d_i\bar{\alpha} + [U_{it} + d_i\varepsilon_i] \tag{3.10}$$

In view of (3.9), the error term in brackets in (3.10) does not have mean zero, and in fact has a mean that depends on $\mathbf{Z}_i$.

Accordingly, the IV method presented in Section 3.2 applied to (3.10) does not produce consistent estimators because (A-7a) fails with respect to the composite error term (i.e., the term in brackets has a nonzero mean). The methodology of Section 3.3 breaks down because

$$\begin{aligned} E(Y_{it} | \mathbf{X}_{it}, \mathbf{Z}_i) &= \mathbf{X}_{it}\boldsymbol{\beta} + [\bar{\alpha} + \phi(Z_i)] \Pr(d_i = 1 | \mathbf{Z}_i) \\ &\neq \mathbf{X}_{it}\boldsymbol{\beta} + \bar{\alpha} \Pr(d_i = 1 | \mathbf{Z}_i) \end{aligned}$$

The control function methodology of Sections 3.4 and 3.5 continues to apply in the random coefficient case except now

$$E(d_i\varepsilon_i + U_{it} | d_i, \mathbf{X}_{it}, \mathbf{Z}_i)$$

must be specified. Provided that certain conditions are satisfied, $\bar{\alpha}$ can be identified from a (possibly nonlinear) regression:

$$E(Y_{it} | \mathbf{X}_{it}, d_i, \mathbf{Z}_i) = \mathbf{X}_{it}\boldsymbol{\beta} + d_i\bar{\alpha} + E(d_i\varepsilon_i + U_{it} | d_i, \mathbf{X}_{it}, \mathbf{Z}_i) \tag{3.11}$$

The required conditions are (A-5), (A-6) (as strengthened above), and

(A-12) (a) The functional form of $E(U_{it} + d_i\varepsilon_i | d_i, \mathbf{X}_{it}, \mathbf{Z}_i)$ is known up to a finite vector ω.

(b) In the population $[\mathbf{X}_{it}, d_i, \partial \mathbf{E}(U_{it} + d_i\varepsilon_i | d_i, \mathbf{X}_{it}, \mathbf{Z}_i, \omega)/\partial\omega]$ is a nondegenerate vector of random variables (i.e., for true ω, the joint distribution is nondegenerate).

This contrast between the consistency of the IV method and the consistency of the control function estimator for the random coefficient model suggests a formal statistical test between random and fixed coefficient models using a Durbin (1954)–Wu (1973, 1983)–Hausman (1978) statistic.

Lee's model of unionism, which is a random coefficient model, generates (A-12a) by assuming that

(a) $(U_{it}, V_i, \varepsilon_i)$ are joint normal random variables with zero mean and a finite nondegenerate covariance matrix functionally independent of $(\alpha, \boldsymbol{\beta}, \gamma)$.

(b) $E(\varepsilon_i d_i + U_{it} | d_i, \mathbf{X}_{it}, \mathbf{Z}_i, \omega) = E(d_i\varepsilon_i + \mathbf{U}_{it} | d_i, \mathbf{Z}_i, \omega)$.

(c) $E(U_{it} | \mathbf{X}_{it}, d_i, \mathbf{Z}_i) = E(U_{it} | d_i, \mathbf{Z}_i)$.

(d) $\mathbf{Z}_i$ is distributed independently of V_i.

(e) The distribution of $\mathbf{Z}_i$ is nondegenerate.[16]

3.8 *Accounting for choice-based sampling and contamination bias*

The preceding analysis assumes access to simple random samples or samples stratified on the basis of exogenous variables. As noted in Section 1.5, the available data on training programs often are not random samples. More frequently the following types of data are available:

(i) Earnings, earnings characteristics, and enrollment characteristics (Y_{it}, $\mathbf{X}_{it}$ and $\mathbf{Z}_i$, respectively) for a sample of trainees ($d_i = 1$).

(ii) Earnings, earnings characteristics, and enrollment characteristics for a sample of nontrainees ($d_i = 0$).

(iii) 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 types (i) and (ii) data are combined and the sample proportion of trainees does not converge to the population proportion of trainees, then the combined sample is a choice-based sample as defined in Section 1.5. If types (i) and (iii) data are combined with or without type (ii) data, there is contamination bias, because the training status of certain persons is not known. In this subsection we examine the robustness of the estimators presented in Sections 3.2–3.6 to choice-based and contaminated samples, and we discuss how certain nonrobust estimators can be modified to produce consistent estimators.

Throughout this subsection we assume

(A-13) (a) There is access to type (i) data;

and we frequently invoke

(A-13) (b) The population proportion of trainees p is known:

$$\Pr(d_i = 1) = p$$

Assumption (a) is essential to any evaluation using cross-section data. Assumption (b) is satisfied for data on most training programs (Barnow, 1983).

We discuss each estimator in turn starting with the instrumental variable estimator presented in Section 3.2. The format of each discussion is

identical. We first assume access to samples (i) and (ii) (i.e., a choice-based sample). Then we assume access to samples (i) and (iii) (i.e., a contaminated sample). Finally we assume access to pooled choice-based contaminated samples.

3.8.1. The IV estimator (Section 3.2)

3.8.1.A. Choice-based sampling [samples (i) and (ii) pooled]. If conditions (A-6d) and (A-7a) are strengthened to read

(A-6d') $E(\mathbf{X}_{it}'U_{it}|d_i) = 0$
(A-7a') $E[g(Z_i^e)U_{it}|d_i] = 0$

and the other conditions of Section 3.2 are 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:

$$\begin{bmatrix} \frac{\sum \mathbf{X}_{it}'\mathbf{X}_{it}}{I_t} & \frac{\sum \mathbf{X}_{it}'d_i}{I_t} \\ \frac{\sum g(Z_i^e)\mathbf{X}_{it}}{I_t} & \frac{\sum g(Z_i^e)d_i}{I_t} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\alpha} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\sum \mathbf{X}_{it}'Y_{it}}{I_t} \\ \frac{\sum g(Z_i^e)Y_{it}}{I_t} \end{pmatrix} = \begin{bmatrix} \frac{\sum \mathbf{X}_{it}'\mathbf{X}_{it}}{I_t} & \frac{\sum \mathbf{X}_{it}'d_i}{I_t} \\ \frac{\sum g(Z_i^e)\mathbf{X}_{it}}{I_t} & \frac{\sum g(Z_i^e)d_i}{I_t} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \alpha \end{pmatrix} + \begin{pmatrix} \frac{\sum \mathbf{X}_{it}'U_{it}}{I_t} \\ \frac{\sum g(Z_i^e)U_{it}}{I_t} \end{pmatrix} \tag{3.12}$$

Since (A-6d') and (A-7a') guarantee that

$$\operatorname*{plim}_{I_t\to\infty} \frac{\sum X_{it}'U_{it}}{I_t} = 0 \quad \text{and} \quad \operatorname*{plim}_{I_t\to\infty} \frac{\sum g(Z_i^e)U_{it}}{I_t} = 0 \tag{3.13}$$

and rank condition (A-7b) is satisfied, the IV estimator is consistent.

In a choice-based sample for period t, the probability that individual i is enrolled in training is p^* in the notation of Section 1.5 and is not the population proportion of trainees. Thus, even if (A-6d) and (A-7a) are satisfied along with the other conditions, there is no guarantee that the conditions (3.13) are met. This is so because[17]

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

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

In general, the terms inside the braces are not zero and so the IV estimator is inconsistent.[18]

In a random sampling environment, $p^* = \Pr(d_i = 1) = p$ and the terms inside the braces are identically zero. They are also zero if (A-6d') and (A-7a') are satisfied. However, it is not necessary to invoke conditions (A-6d') and (A-7a'). Since p is known, it is possible to reweight the data to secure consistent estimators under the assumptions of Section 3.2. Multiplying equation (1.3) by 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 (3.13). The proof is straightforward and hence is omitted. The intuition underlying this procedure is clear. By weighting the sample at hand back to random sample proportions, the IV estimator of Section 3.2 produces a consistent estimator (see Manski and Lerman, 1977).

3.8.1.B. Contamination bias [samples (i) and (iii)]. By assumption, d_i is not observed for observations in random sample (iii). Applying the IV estimator to pooled samples (i) and (iii), assuming that observations in (iii) have $d_i = 0$, produces an inconsistent estimator.

In terms of the IV equations (3.13), from sample (iii) data it is possible to generate the cross-products from the $I_{(iii)}$ observations

$$\frac{\sum \mathbf{X}_{it}'\mathbf{X}_{it}}{I_{(iii)}}, \quad \frac{\sum g(Z_i^e)\mathbf{X}_{it}}{I_{(iii)}}, \quad \frac{\sum \mathbf{X}_{it}'Y_{it}}{I_{(iii)}}, \quad \frac{\sum g(Z_i^e)Y_{it}}{I_{(iii)}} \tag{3.14}$$

which under the conditions stated in (A-6) converge to the desired population counterparts. Equations (3.13) are satisfied in this sample. Missing is information on the cross-products

$$\frac{\sum \mathbf{X}_{it}'d_i}{I_{(iii)}}, \quad \frac{\sum g(Z_i^e)d_i}{I_{(iii)}} \tag{3.15}$$

Notice that if d_i were accurately measured in sample (iii), then[19]

$$\operatorname*{plim}_{I_{(iii)}\to\infty} \frac{\sum \mathbf{X}_{it}'d_i}{I_{(iii)}} = p\frac{\sum E(\mathbf{X}_{it}'|d_i=1)}{I_{(iii)}}$$

$$\operatorname*{plim}_{I_{(iii)}\to\infty} \frac{\sum g(Z_i^e)d_i}{I_{(iii)}} = p\frac{\sum E[g(Z_i^e)|d_i=1]}{I_{(iii)}}$$

But the means of $\mathbf{X}_{it}$ and $g(Z_i^e)$ in sample (i) converge to

$$\frac{\sum E(\mathbf{X}_{it}|d_i=1)}{I_{(i)}} \quad \text{and} \quad \frac{\sum E[g(Z_i^e)|d_i=1]}{I_{(i)}}$$

respectively. Hence inserting the sample (i) means of $\mathbf{X}_{it}$ and $g(Z_i^s)$ multiplied by p in the second column of the matrix of IV equations (3.12) produces a consistent IV estimator, provided that in the limit the sizes of samples (i) and (iii) both approach infinity at the same rate.

3.8.1.C. Choice-based sampling plus contamination bias [samples (i), (ii), and (iii)]. Samples (i) and (ii) can be pooled using the weights ω_i defined above. Sample (iii) can be used to improve the efficiency of the procedure by combining the moments in (3.14) constructed from sample (ii) with the corresponding moments for the weighted observations to form the normal questions for the IV estimator. In combining moments formed from different samples, weight the moments for each sample by the relative sample size.

3.8.2. Procedures based on known or estimated F (Section 3.3)

3.8.2.A. Choice-based sampling [samples (i) and (ii)]. Procedures that exploit knowledge of F and the other conditions listed in Section 3.3 are of three types: (1) nonlinear regression estimators of equation (3.2); (2) two-stage estimators that estimate F_i in stage 1 and use $\hat{F}_i$ as a regressor in the second stage; (3) a procedure that uses $\hat{F}_i$ as an instrument for d_i. In this subsection we consider how the estimators proposed for random samples can be adapted to yield consistent estimators for choice-based samples.

Manski and McFadden (1981) demonstrate that if (A-13b) is true, it is possible to consistently estimate γ and hence F_i from choice-based samples provided that standard regularity conditions are met. [In fact, they demonstrate that in certain cases (A-13b) is stronger than is required.] If their conditions are met, they propose several consistent estimators for F_i.

Provided that the data are appropriately reweighted, it is possible to apply estimators (1)–(3) listed above to data generated from choice-based samples. The appropriate weight for observation i in cross section t is

$$\phi_{it} = \frac{\Pr(d_i = 1|\mathbf{Z}_i)}{p^*(d_i = 1, \mathbf{Z}_i)} d_i + \frac{\Pr(d_i = 0|\mathbf{Z}_i)}{p^*(d_i = 0|\mathbf{Z}_i)} (1 - d_i)$$

where $\Pr(d_i = 1|\mathbf{Z}_i)$ is the probability that $d_i = 1$ in the population and $p^*(d = 1|\mathbf{Z}_i)$ is the probability that $d_i = 1$ in the population generated by a choice-based sample. The terms $\Pr(d_i = 0|\mathbf{Z}_i)$ and $p^*(d_i = 0|\mathbf{Z}_i)$ are the corresponding conditional probabilities for the event $d_i = 0$. Consistent empirically feasible values of these weights are produced from the Manski–McFadden procedure (for the numerator expressions) and by direct estimation from the choice-based sample (for the denominator expressions).

Direct calculation using the law of iterated expectation reveals (for known ϕ_{it}) that in a reweighted choice-based sample[20]

$$E(\phi_{it} Y_{it}|\phi_{it}\mathbf{X}_{it}, \mathbf{Z}_i) = \phi_{it}\mathbf{X}_{it}\boldsymbol{\beta} + [1 - F(-\mathbf{Z}_i\gamma)]\alpha$$

Methods (1) and (2) listed above apply without modification to the reweighted data. Method (3) uses the $\hat{F}_i$ obtained from the Manski–McFadden procedure in the IV equations described in the preceding subsection. Conditional weights are not required in method 3 because unconditional moments are used in the IV method.

3.8.2.B. Contamination bias [samples (i) and (iii)]. By construction, estimation of equation (3.2) does not require knowledge of d_i, so that under the assumptions of Section 3.3 direct nonlinear estimation of (3.2) on sample (iii) produces consistent estimators of the parameters of that equation. Note that sample (i) is nowhere needed, nor is the population proportion of trainees [so assumption (A-13) is not required].

The two-stage estimator is not directly feasible. Assuming that it is possible to consistently estimate $p = \Pr(d_i = 1)$, it may be possible to consistently estimate γ and hence $F(-\mathbf{Z}_i\gamma)$. The proposed procedure posits the existence of two vector-valued functions $\mathbf{g}^*(\mathbf{Z}_i|\gamma)$ and $\mathbf{g}(\mathbf{Z}_i|\gamma)$, which have the property that

$$\int \mathbf{g}^*(\mathbf{Z}_i|\gamma) f(\mathbf{Z}_i|d = 1, \gamma_0)\, d\mathbf{Z}_i = \int \mathbf{g}(\mathbf{Z}_i|\gamma) f(\mathbf{Z}_i)\, d\mathbf{Z}_i \tag{3.16}$$

when

$$\gamma = \gamma_0$$

where γ_0 is the true value of the parameter and where for notational convenience we suppress the dependence of $f(\mathbf{Z}_i|d = 1, \gamma_0)$ and $f(\mathbf{Z}_i)$ on other parameters of the model.

The postulated functions are also assumed to possess the property that for all values of ρ sufficiently small

$$\{\gamma \,|\, |\gamma - \gamma_0| < \rho, \qquad \gamma \neq \gamma_0\}$$

implies

$$\int \mathbf{g}^*(\mathbf{Z}_i|\gamma) f(\mathbf{Z}_i|d = 1, \gamma_0)\, dZ_i \neq \int \mathbf{g}(\mathbf{Z}_i|\gamma) f(\mathbf{Z}_i)\, d\mathbf{Z}_i$$

so that γ is at least locally identified. If this condition is satisfied for all $\gamma \neq \gamma_0$, γ is globally identified. Given this identifiability condition, it is possible to consistently estimate γ from the means of $\mathbf{g}$ and $\mathbf{g}^*$ formed in samples (iii) and (i), respectively. The proposed estimator $\hat{\gamma}$ solves for the γ that equates the means of $\mathbf{g}^*(\mathbf{Z}_i|\gamma)$ and $\mathbf{g}(\mathbf{Z}_i|\gamma)$ in the two samples:

$$\frac{1}{I_{(i)}} \sum_{i=1}^{I_{(i)}} \mathbf{g}^*(\mathbf{Z}_i|\hat{\gamma}) = \frac{1}{I_{(iii)}} \sum_{i=1}^{I_{(iii)}} \mathbf{g}(\mathbf{Z}_i|\hat{\gamma}) \tag{3.17}$$

where $I_{(i)}$ and $I_{(iii)}$ are, respectively, the sample sizes in samples (i) and (iii). As $I_{(i)}$ and $I_{(iii)} \to \infty$, $\hat{\gamma}$ is consistent for γ_0 given standard assumptions that justify the uniform strong law of large numbers for stratified samples. For any value of γ that satisfies these conditions, the left-hand side of (3.17) converges to the left-hand side of (3.16) and the right-hand side of (3.17) converges to the right-hand side of (3.16).

One set of $\mathbf{g}$ and $\mathbf{g}^*$ functions is produced from the following intuitive argument. In sample (i) choose γ to maximize the average of the logs of the probability of enrollment

$$\operatorname{Max} \frac{1}{I_{(i)}} \sum_{i=1}^{I_{(i)}} \ln[1 - F(-\mathbf{Z}_i\gamma)] \tag{3.18}$$

subject to the constraint that in sample (iii)

$$\frac{1}{I_{(iii)}} \sum_{i=1}^{I_{(iii)}} [1 - F(-\mathbf{Z}_i\gamma)] = p \tag{3.19}$$

where p is known. A routine calculation reveals that, in large samples at $\hat{\gamma} = \gamma_0$, the Lagrange multiplier associated with the constraint has value $\lambda = -1/p$.

The first-order conditions for this problem are (asymptotically)

$$0 = \frac{1}{I_{(i)}} \sum_{i=1}^{I_{(i)}} \left(-\frac{F'(-\mathbf{Z}_i\hat{\gamma})\mathbf{Z}_i}{1 - F(\mathbf{Z}_i\hat{\gamma})} \right) + \frac{1}{I_{(iii)}} \sum_{i=1}^{I_{(iii)}} \left(\frac{1}{p}\right) F'(-\mathbf{Z}_i\hat{\gamma})\mathbf{Z}_i$$

It is easily verified that by defining

$$\mathbf{g}^*(\mathbf{Z}_i|\gamma) = \frac{F'(-\mathbf{Z}_i\gamma)\mathbf{Z}_i}{1 - F(\mathbf{Z}_i\gamma)}$$

and

$$\mathbf{g}(\mathbf{Z}_i|\gamma) = \frac{F'(-\mathbf{Z}_i\gamma)\mathbf{Z}_i}{p}$$

we produce $\mathbf{g}^*$, $\mathbf{g}$ functions that satisfy the definition. Thus, solving the synthetic optimization problem produces one pair of appropriate $\mathbf{g}$, $\mathbf{g}^*$ functions. With consistent estimators for γ it is possible to estimate $F(-\mathbf{Z}_i\gamma)$ and hence perform the two-stage procedure.

3.8.2.C. Choice-based sampling plus contamination bias [samples (i), (ii), and (iii)]. Data from the contaminated sample can be pooled with data from the ω_i weighted choice-based sample to create a pooled sample on which nonlinear regression (3.2) – method 1 – consistently estimates α. Data from samples (i) and (ii) can be pooled to consistently estimate F_i. Now $\hat{F}_i$ can be constructed for each observation in sample (iii). Reweighting samples (i) and (ii) and pooling with sample (iii) produces a sample in which the two-

stage estimator (method 2) consistently estimates α. Alternatively, $\hat{F}_i$ can be used as an instrument in the manner already described in the preceding subsection.

3.8.3. Control function estimators (Section 3.4)

3.8.3.A. Choice-based sampling [samples (i) and (ii) pooled]. If $E(U_{it}|\mathbf{X}_{it}, d_i, \mathbf{Z}_i)$ is known or can be consistently estimated (up to a finite set of parameters) and any one of the sets of conditions stated in Section 3.4 is satisfied, the conditional mean can be used as a control function. As already noted in Section 1.5, control function estimators are robust to choice-based sampling.

Method 3.4.2 of Section 3.4 (direct nonlinear regression) can be applied without modification to produce consistent estimators. The two-stage estimator (method 3.4.1) is also consistent if the Manski–McFadden procedures are used to estimate $\Pr(d_i = 0|\mathbf{Z}_i)$ and equation (3.6) is used to generate an estimate of $E(U_{it}|d_i = 0, \mathbf{Z}_i)$ up to a finite set of parameters. Maximum likelihood (method C) is consistent using the Manski–McFadden reweighted estimator for the joint frequency of (d_i, Y_{it}).

3.8.3.B. Contamination bias [samples (i) and (iii) pooled]. Using sample (i) it is possible to consistently estimate $f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it}|d_i = 1)$ (see, e.g., Cosslett, 1981, for a discussion of nonparametric frequency estimation).[21] From sample (iii) it is possible to consistently estimate $f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it})$. With these two frequencies and assumption A-13(b) [which gives $\Pr(d_i = 1) = p$] it is possible to solve for $f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it}|d_i = 0)$ from the equation

$$\begin{aligned} f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it}) = {} & f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it}|d_i = 1) \Pr(d_i = 1) \\ & + f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it}|d_i = 0) \Pr(d_i = 0) \end{aligned}$$

so that it is possible to consistently estimate

$$\begin{aligned} & f(\mathbf{Z}_i, Y_{it}, \mathbf{Z}_{it}, d_i) \\ & = [f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it}|d_i = 1) \Pr(d_i = 1)]^{d_i} [f(\mathbf{Z}_i, Y_{it}, \mathbf{X}_{it}|d_i = 0) \Pr(d_i = 0)]^{1-d_i} \end{aligned}$$

and hence

$$\Pr(d_i = 0|\mathbf{Z}_i) \quad \text{and} \quad f(Y_{it}, d_i|\mathbf{Z}_i, \mathbf{X}_{it})$$

Adopting parametric functional forms for the conditional frequency functions, $\Pr(d_i = 0|\mathbf{Z}_i)$ and $f(Y_{it}, d_i|\mathbf{Z}_i, \mathbf{X}_{it})$, standard maximum likelihood procedures can be used to estimate α consistently. Sample (i) provides information on $f(Y_{it}|\mathbf{X}_{it}, d_i = 1, \mathbf{Z}_i)$ whereas sample (iii) provides information on $f(Y_{it}|\mathbf{X}_{it}, \mathbf{Z}_i)$. The samples can be pooled to form a likelihood

function which, when maximized with respect to β, α, ψ, produces a consistent estimator (method 3.4.3) under the assumptions of Section 3.3.

Although d_i is not known for any observation in sample (iii), the normal equations associated with nonlinear regression (3.4) can be formed by pooling moments from samples (i) and (iii) in a manner to be described next. To understand how the proposed method works, it is helpful to write out the appropriate normal equations for a case in which d_i is observed and the analyst has access to a random sample.

The required equations are obtained by minimizing

$$\frac{1}{I_t}\sum_{i=1}^{I_t}\left[Y_{it} - \mathbf{X}_{it}\beta - d_i\alpha - E(U_{it}|d_i = 1, \mathbf{Z}_i, \psi)d_i \right.$$
$$\left. - E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)(1 - d_i)\right]^2$$

with respect to β, α, ψ. Note that we have made explicit the dependence of the conditional mean of the U_{it} on ψ. Let

$$K_i(d_i, \mathbf{Z}_i, \psi) = E(U_{it}|d_i = 1, \mathbf{Z}_i, \psi)d_i + E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)(1 - d_i)$$

to simplify notation. The first-order conditions for the minimization problem (suppressing the arguments of K_i) are

$$\frac{1}{I_t}\sum \mathbf{X}'_{it}Y_{it} = \left(\frac{1}{I_t}\sum \mathbf{X}'_{it}\mathbf{X}_{it}\right)\beta + \left(\frac{1}{I_t}\sum \mathbf{X}'_{it}d_i\right)\alpha + \frac{1}{I_t}\sum \mathbf{X}'_{it}K_i$$

$$\frac{1}{I_t}\sum d_iY_{it} = \left(\frac{1}{I_t}\sum d_i\mathbf{X}_{it}\right)\beta + \left(\frac{1}{I_t}\sum d_i\right)\alpha + \frac{1}{I_t}\sum d_iK_i \qquad (3.20)$$

$$\frac{1}{I_t}\sum\left(\frac{\partial K'_i}{\partial\psi}Y_{it}\right) = \left(\frac{1}{I_t}\sum \frac{\partial K'_i}{\partial\psi}\mathbf{X}_{it}\right)\beta + \left(\frac{1}{I_t}\sum \frac{\partial K'_i}{\partial\psi}d_i\right)\alpha + \frac{1}{I_t}\sum \frac{\partial K'_i}{\partial\psi}K_i$$

Under the assumptions of Section 3.4, there exists a consistent root of these equations.

All of the moments needed to form equation system (3.20) are not directly available from sample (iii). But the required moments can be formed by pooling sample (i) and sample (iii) information in the following way.

When we discussed the IV estimator, we showed how to use sample (i) and (iii) data to estimate the following moments consistently:

$$\operatorname{plim}\frac{1}{I_t}\sum \mathbf{X}'_{it}Y_{it}, \qquad \operatorname{plim}\frac{1}{I_t}\sum \mathbf{X}'_{it}\mathbf{X}_{it}, \qquad \operatorname{plim}\frac{1}{I_t}\sum \mathbf{X}'_{it}d_i,$$

$$\operatorname{plim}\frac{1}{I_t}Y_{it}d_i, \qquad \operatorname{plim}\frac{1}{I_t}\sum d_i = p$$

The term

$$\operatorname{plim}\frac{1}{I_t}\sum \frac{\partial K'_i}{\partial\psi}Y_{it}$$

can be estimated in a simple way. In sample (iii), form

$$\frac{1}{I_{(\text{iii})}}\sum Y_{it}\frac{\partial E}{\partial\psi}[U_{it}|d_i = 0, \mathbf{Z}_i, \psi]$$

In sample (i), form

$$\frac{p}{I_{(\text{i})}}\sum Y_{it}\left[\frac{\partial E}{\partial\psi}(U_{it}|d = 1, \psi, \mathbf{Z}_i) - \frac{\partial E}{\partial\psi}(U_{it}|d = 0, \psi, \mathbf{Z}_i)\right]$$

The sum of the preceding two quantities converges in probability to

$$(1 - p)E\left[Y_{it}\frac{\partial E}{\partial\psi}(U_{it}|d_i = 0, \psi, \mathbf{Z}_i)|d_i = 0\right]$$
$$+ pE\left[Y_{it}\frac{\partial E}{\partial\psi}(U_{it}|d_i = 1, \psi, \mathbf{Z}_i)d_i = 1\right]$$
$$= E\left[Y_{it}\frac{\partial K_i}{\partial\psi}\right] = \operatorname{plim}\frac{1}{I_t}\sum Y_{it}\frac{\partial K_i}{\partial\psi}$$

A parallel argument produces consistent estimators of

$$\operatorname{plim}\frac{1}{I_t}\sum \mathbf{X}'_{it}K_i, \qquad \operatorname{plim}\frac{1}{I_t}\sum d_iK_i, \qquad \operatorname{plim}\frac{1}{I_t}\sum \frac{\partial K'_i}{\partial\psi}\mathbf{X}_{it}, \qquad \text{and}$$

$$\operatorname{plim}\frac{1}{I_t}\sum \frac{\partial K'_i}{\partial\psi}d_i$$

Finally, we need to estimate

$$\operatorname{plim}\frac{1}{I_t}\sum \frac{\partial K'_i}{\partial\psi}K_i$$

To this end it is helpful to notice that

$$\frac{1}{I_t}\sum \frac{\partial K_i}{\partial\psi}K_i = \sum \frac{d_i}{I_t}\frac{\partial E(U_{it}|d_i = 1, \mathbf{Z}_i, \psi)}{\partial\psi}E(U_{it}|d_i = 1, \mathbf{Z}_i, \psi)$$
$$+ \sum \frac{(1 - d_i)}{I_t}\frac{\partial E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)}{\partial\psi}E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)$$

so

$$\operatorname{plim} \frac{1}{I_t} \sum \frac{\partial K_i}{\partial \psi} K_i = pE\left[\frac{\partial E(U_{it}|d_i = 1, \mathbf{Z}_i, \psi)}{\partial \psi} E[U_{it}|d_i = 1, \mathbf{Z}_i, \psi]|d_i = 1\right] + (1-p)E\left[\frac{\partial E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)}{\partial \psi} E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)|d_i = 0\right]$$

From samples (i) and (iii), we can construct the following two moments:

$$\frac{p}{I_{(i)}} \sum \left[\frac{\partial E(U_{it}|d_i = 1, \mathbf{Z}_i, \psi)}{\partial \psi} E(U_{it}|d_i = 1, \mathbf{Z}_i, \psi) - \frac{\partial E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)}{\partial \psi} E(U_{it}|d_i = 0, \mathbf{Z}_i, \psi)\right] \tag{3.21}$$

and

$$\frac{1}{I_{(iii)}} \sum \frac{\partial E(U_{it}|d = 0, \mathbf{Z}_i, \psi)}{\partial \psi} E(U_{it}|d = 0, \mathbf{Z}_i, \psi) \tag{3.22}$$

Adding (3.21) and (3.22) together produces a consistent estimator of

$$\operatorname{plim} \frac{1}{I_t} \sum \frac{\partial K_i}{\partial \psi} K_i$$

An alternative nonlinear regression estimator that is available when the functional form of $\Pr(d = 1|\mathbf{Z})$ is known up to a finite number of parameters pools samples (i) and (iii). Let $\tilde{d}_i = 1$ if a person is drawn from sample (i). Otherwise the observation is from sample (iii) ($\tilde{d}_i = 0$). In place of equation (3.4) we write

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + [\tilde{d}_i + (1 - \tilde{d}_i) \Pr(d_i = 1|\mathbf{Z}_i, \mathbf{X}_{it})]\alpha + E(U_{it}|\mathbf{Z}_i, d_i = 1, \psi)\tilde{d}_i + \chi_{it} \tag{3.4'}$$

where

$$\chi_{it} = U_{it} + (d_i - \tilde{d}_i)\alpha - (1 - \tilde{d}_i) \Pr(d_i = 1|\mathbf{Z}_i, \mathbf{X}_{it})\alpha - \tilde{d}_i E(U_{it}|\mathbf{Z}_i, \tilde{d}_i = 1, \psi)$$

Under the assumptions of Section 3.4, and assuming that U_{it} is mean independent of $\mathbf{Z}_i$, $E(\chi_{it}|\mathbf{X}_{it}, \mathbf{Z}_i, d_i) = 0$. To show this, note that

$$E(\chi_{it}|\tilde{d}_i = 1, \mathbf{X}_{it}, \mathbf{Z}_i) = E(U_{it}|\mathbf{Z}_i, \tilde{d}_i = 1, \psi) - E(U_{it}|\mathbf{Z}_i, d_i = 1, \psi) = 0$$

(Note that we use the fact that $\tilde{d}_i = 1 \Rightarrow d_i = 1$.) Note further that

$$\begin{aligned} E(\chi_{it}|\tilde{d}_i = 0, \mathbf{X}_{it}, \mathbf{Z}_i) &= E(U_{it}|\mathbf{Z}_i, \psi) + \Pr(d_i = 1|\mathbf{Z}_i, \mathbf{X}_{it})\alpha \\ &\quad - \Pr(d_i = 1|\mathbf{Z}_i, \mathbf{X}_{it})\alpha \\ &= E(U_{it}|\mathbf{Z}_i, \psi) = 0 \end{aligned}$$

Under the assumptions of Section 3.4 nonlinear regression estimators of the parameters of (3.4') are consistent for $(\alpha, \boldsymbol{\beta}, \psi)$ provided that the number of observations in both samples (i) and (iii) becomes large.

3.8.3.C. Choice-based sampling plus contamination bias [samples (i), (ii), and (iii)]. Data from the three samples may be pooled directly and the parameters consistently estimated by maximum likelihood. The choice-based sample data should be appropriately weighted as described above. The data from the three samples can be combined to form the elements of the normal equations (3.20).

3.8.4. Selection on observables and random coefficient models (Sections 3.5 and 3.7). Since the estimators proposed in Sections 3.5 and 3.7 are control function estimators, the preceding discussion applies to those estimators with obvious modifications for the change in the nature of the control functions.

3.8.5. Distributional assumptions invoked about U_{it} (Section 3.6)

3.8.5.A. Choice-based sampling [samples (i) and (ii) pooled]. The method-of-moments estimator of Section 3.6 consistently estimates α in choice-based samples provided that the data are appropriately reweighted. The appropriate weights are identical to the weights proposed in our discussion of the instrumental variables estimator. The moments formed in sample (i) should be weighted by p/p^*. The moments formed in sample (ii) should be weighted by $(1 - p)/(1 - p^*)$. The weighted sum of the moments converges to the desired random sample moments.

3.8.5.B. Contamination bias [samples (i) and (iii)]. Provided that the population proportion of trainees is known or can be consistently estimated, a consistent method-of-moments estimator can be devised for contaminated samples. We demonstrate the modifications required in (3.7a). The required modification for (3.7b) is a straightforward application of the principles used to modify (3.7a).

We write out (3.7a) in full:

$$\frac{1}{I_t}\sum (Y_{it} - \mu_t)^3 - 3\hat{\alpha}\frac{1}{I_t}\sum (Y_{it} - \mu_t)^2(d_i - p) + 3\hat{\alpha}^2\frac{1}{I_t}\sum (Y_{it} - \mu_t)(d_i - p)^2 - \frac{1}{I_t}\sum (d_i - p)^3 \tag{3.23}$$

The probability limit of the first term can be consistently estimated from sample (iii) data. The probability limit of the final term can be consistently estimated since p is known or is assumed to be consistently estimable.

The normalized sum in the second term consists of two components:

$$\frac{1}{I_t}\sum_{i=1}^{I_t}(Y_{it}-\mu_t)^2 d_i \quad \text{and} \quad \frac{p}{I_t}\sum_{i=1}^{I_t}(Y_{it}-\mu_t)^2$$

The probability limit of the second component can be consistently estimated from sample (iii) (given p). The probability limit of the first component can be estimated from sample (i) data using the estimate of μ_t obtained from sample (iii) data.

The sample (i) moment converges to

$$\operatorname*{plim}_{I_{(i)}\to\infty}\frac{1}{I_{(i)}}\sum_{i=1}^{I_{(i)}}(Y_{it}-\mu_t)^2 = E[(Y_{it}-\mu_t)^2 | d_i = 1]$$

Weighting the sample (iii) moment by p and combining with the second component produces an expression that converges to the probability limit of the normalized sum in the second term in (3.23). Thus the probability limit of the normalized sum in the second term can be consistently estimated. A parallel procedure can be used to consistently estimate the probability limit of the normalized sum in the third term of (3.23). By combining sample moments in this fashion we produce an estimating equation that converges to the same limit as the estimating equation formed in a random sample. A parallel procedure can be used for equation (3.7b). Thus it is possible to solve the problems raised by contamination bias.

If there are regressors in the earnings equations, a small modification in the procedure of Section 3.6 is required. Using an obvious and conventional matrix notation, we get

$$\hat{\boldsymbol{\beta}}(\hat{\alpha}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y}-\hat{\alpha}\mathbf{d}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{d})\hat{\alpha} \quad (3.24)$$

where $\mathbf{X}$ is the sample $\mathbf{X}_{it}$ vectors arrayed in a matrix, and $\mathbf{Y}$ and $\mathbf{d}$ are defined in a similar fashion. In a random sample

$$\operatorname*{plim}_{I_t\to\infty}\hat{\boldsymbol{\beta}}(\hat{\alpha}) = \operatorname*{plim}_{I_t\to\infty}\left(\frac{\mathbf{X}'\mathbf{X}}{I_t}\right)^{-1}\operatorname*{plim}_{I_t\to\infty}\left(\frac{\mathbf{X}'\mathbf{Y}}{I_t}\right) - \operatorname*{plim}_{I_t\to\infty}\left(\frac{\mathbf{X}'\mathbf{X}}{I_t}\right)^{-1}\left(\operatorname*{plim}_{I_t\to\infty}\frac{\mathbf{X}'\mathbf{d}}{I_t}\right)\hat{\alpha} \quad (3.25)$$

The first term on the right-hand side of (3.24) can be formed in sample (iii) and converges to the first term on the right-hand side of (3.25). Clearly $(\mathbf{X}'\mathbf{X}/I_t)^{-1}$ formed in sample (iii) converges to $\operatorname{plim}(\mathbf{X}'\mathbf{X}/I_t)^{-1}$.

The mean of the $\mathbf{X}_{it}$ in sample (i), $\bar{\mathbf{X}}^{(i)}$, converges to

$$E(\mathbf{X}_{it}|d_i = 1)$$

Multiplying $\bar{\mathbf{X}}^{(i)}$ by p produces an expression that converges to plim $(\mathbf{X}'\mathbf{d}/I_t)$. Then

$$\hat{\hat{\boldsymbol{\beta}}}(\hat{\alpha}) = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}]_{(iii)} - [(\mathbf{X}'\mathbf{X})^{-1}]_{(iii)}\hat{\alpha}\bar{\mathbf{X}}^{(i)}p$$

where $[\ ]_{iii}$ denotes that the term inside the brackets is formed in sample (iii), converges to the probability limit of $\hat{\boldsymbol{\beta}}(\hat{\alpha})$. Inserting $\hat{\hat{\boldsymbol{\beta}}}(\hat{\alpha})$ in place of $\hat{\boldsymbol{\beta}}(\hat{\alpha})$ in the estimator proposed in Section 3.5 produces an estimator that is consistent for α.

3.8.5.C. Choice-based sampling plus contamination bias [samples (i), (ii), and (iii)]. Data from the three samples may be pooled directly. The choice-based sample data should be appropriately weighted as described above. The data from sample (iii) may be used to form moments which when pooled with their weighted sample (i) and (ii) data will improve the efficiency of the estimator.

3.9 *Summary of cross-section procedures*

A variety of cross-sectional estimators have been presented. All of these estimators with the exception of the estimator presented in Section 3.6 share the common feature that in order for α in (1.1) to be identified, a regressor must appear in decision rule (1.2). However, little more than this is required if the regressor is exogenous with respect to U_{it} (in the sense of Section 3.2). For the fixed coefficient model, the regressor can be used as a valid instrument for d_i provided that the technical conditions specified in Section 3.2 are satisfied. No specific distributional assumptions with respect to (U_{it}, V_i) are required, nor is it necessary to specify the functional form of the conditional mean of U_{it} given d_i and $\mathbf{Z}_i$.

The same cannot be said about the random coefficients model (Section 3.7). Unless $E(\varepsilon_i|d_i, \mathbf{Z}_i) = 0$ or some other known constant, it is necessary to specify the functional form of $E(\varepsilon_i d_i + U_{it}|d_i, \mathbf{Z}_i)$ up to a finite number of unknown parameters in order to identify $\bar{\alpha}$. One way to do this is to assume that the joint distribution of (U_{it}, V_i) is known up to a finite set of parameters. This observation reiterates a main point of this section – that the random coefficient model requires much stronger identifying assumptions than the fixed coefficient model.

A variety of alternative cross-section estimators for the fixed coefficient model are presented in Sections 3.3–3.6. These estimators differ in the amount of prior information assumed to be available to the analyst. The

regression estimators of Section 3.3 require information on the functional form of the distribution of V_i. The control function estimators of Sections 3.4 and 3.5 require that the functional form of the conditional expectation of U_{it} and d_i, $\mathbf{Z}_i$, and $\mathbf{X}_{it}$ be known up to a finite number of parameters. In principle, no regressor variable in the decision rule needs to be exogenous with respect to V_i or U_{it} to identify α. However, in this case, identification of α is secured strictly as a consequence of assumed functional forms. Provided that one is willing to accept identification via functional form, all of the fixed coefficient training models of Section 3.2 can be consistently estimated irrespective of the exogeneity of regressors.

Our discussion in Section 3.6 demonstrates that if certain restrictions are imposed on the density of U_{it}, no regressor need appear in the enrollment rule in order to consistently estimate α.

The control function estimators are robust to choice-based sampling. Provided that the population proportion of trainees is known, the remaining estimators can be modified to produce consistent estimators of α in the presence of contamination bias and choice-based sampling.

If there is access to multiple cross sections, a separate α can be fit for each sample. It is not necessary to assume that α is constant over the trainee's life cycle. The cross-sectional estimator is thus seen to be robust to aging and decay effects in training.

4 Repeated cross-section methods for the case when the training identity of individuals is unknown

4.1 *Time homogeneity*

As already noted in Section 1, there are two types of repeated cross-section estimators: (1) those that do not require that the training status of individuals be known and (2) those that do. This section considers the first type of estimator. The second type is discussed in Section 5 along with the longitudinal estimators. That section demonstrates a major conclusion of this chapter: Virtually all longitudinal estimators can be implemented using repeated cross-section data provided that the training status is known for each person in each cross section.

We first present an explicit statement of the conditions required to consistently estimate α that are implicitly presented in Section 1. Assumption (A-1) is assumed to characterize the data in each cross section. In addition it is further assumed that:

(A-14)
(a) The earnings function is (1.3) (so there are no $\mathbf{X}_{it}$ variables in the earnings function and the model is of the fixed coefficient type).
(b) The environment is time-homogeneous (so $\beta_t = \beta_{t'}$ for all t, t').

(c) Condition (A-2) is satisfied.
(d) There is at least one preprogram cross section t' and one postprogram cross section t, $t > k > t'$.[22]
(e) The population proportion of trainees is known or can be consistently estimated in each cross section, and for each t, $p_t > 0$.

Conspicuously absent from this list is any statement about the enrollment rule. No regressor is required in enrollment rule (1.2), and any of the fixed coefficient enrollment rules of Section 2 can generate program participation. The repeated cross-section estimator is thus very robust to alternative specifications of the enrollment rule. In addition, it is robust to mean zero measurement error in the variables.

4.2 *Relaxing the time homogeneity assumption*

As already noted in Section 1, it is possible to relax assumption (A-14b) and still preserve consistency of the repeated cross-section estimator $\hat{\alpha}_{RC}$. It is useful to distinguish two cases. The first case assumes access to a cohort of persons all of whom have one opportunity to train in period k. The second case assumes access to multiple-cohort data in which individuals of different cohorts have identical period effects.

4.2.1 Single-cohort data. Provided that the available repeated cross-section data satisfy (A-14) except for (A-14b), α is identified provided that the environment is sufficiently regular. A more precise definition of sufficient regularity is as follows:

(D-3) Assume access to L temporally distinct cross sections of earnings data on a cohort of persons all of whom have access to training in period k. None of the cross sections is for period k. Now β_t in (1.3) is generated by a sufficiently regular environment if

$$\beta_t = g(t, \gamma)$$

where $g(t, \gamma)$ is a function of at most $L - 1$ independent parameters, arrayed in vector γ, and unique solutions for α and $g(t, \gamma)$ can be obtained from the probability limits of the L cross-section mean earnings.

More formally, if we array the L values of mean earnings in a $L \times 1$ vector $\bar{\mathbf{Y}}$, the environment is sufficiently regular if the equation system

$$\operatorname*{plim}_{I \to \infty} \bar{\mathbf{Y}} = [g(\gamma)] + (\mathbf{p})\alpha$$

has a unique solution, where γ is a vector of parameters in the g function, and where the tth element of $[g(\gamma)]$ is the value of $g(t, \gamma)$ for the appropriate year and the tth element of $\mathbf{p}$ is p_t for the appropriate year where $p_t = 0$, $t < k$. Here I denotes sample size.

From the L means it is possible to estimate consistently the $\beta_t + p_t\alpha$ for $t > k$ and β_t for $t < k$, where by virtue of assumption (A-14e) the p_t are assumed known. From the definition of sufficient regularity, it is possible to solve for $\mathbf{g}(t, \gamma)$ and hence α from the L means. A specimen $\mathbf{g}(t, \gamma)$ function is

$$\mathbf{g}(t, \gamma) = \gamma_0 + \gamma_1 t + \cdots + \gamma_{L-2} t^{L-2} \tag{4.1}$$

From L temporally distinct cross sections from a sufficiently regular environment it is possible to consistently estimate the $L - 1$ γ parameters and α provided that the number of observations in each cross section becomes large and there is at least one preprogram cross section.

If the effect of training differs across periods, it is still possible to identify α_t provided that the environment is more regular than is implied in (D-3). Define

$$\alpha_t = h(t), \qquad t > k$$

Then it is sometimes possible to identify $h(t)$ and $\mathbf{g}(t, \gamma)$ from L temporally distinct cross sections. For example, suppose

$$h(t) = \phi_0(\phi_1)^{t-k} \quad \text{for} \quad t > k \quad \text{and} \quad \mathbf{g}(t, \gamma) = \gamma_0 + \gamma_1 t$$

and $L = 4$. Then $h(t)$ and $\mathbf{g}(t, \gamma)$ are both identified, so long as there is at least one preprogram cross section or else $p_{t'} \neq p_t$. On the other hand, if $p_t = p_{t'}$, for all t, $t' > k$ and

$$h(t) = \phi_0 + \phi_1 t \quad \text{and} \quad \mathbf{g}(t, \gamma) = \gamma_0 + \gamma_1 t$$

it is not possible to identify $h(t)$ unless the analyst has two or more years of preprogram data.

4.2.2. Multiple-cohort data. We next assume access to samples of two or more cohorts satisfying (A-14) [except for (b)] that are spaced at least two periods apart. In place of (A-14b) we assume the following:

(A-14b′) All cohorts experience a common period effect.

With this assumption, it is possible to dispense with sufficient regularity assumption (D-3) altogether and still secure consistent estimators of α from repeated cross-section data. If each cohort experiences a unique period effect, there is no gain in identifying α from having access to multiple-cohort data.

To show how (A-14b′) aids in securing identification, it is instructive to consider Table 1. There we display the population mean earnings histories of three adjacent cohorts. To simplify the argument we assume that $p_t = p_{t'}$ for all t, t' for each cohort and that α_t is identical in all

Table 1. *Mean earnings history of three adjacent cohorts assuming that the training effect is common across time periods*

	Earnings in year before training	Earnings in first year after training	Earnings in second year after training
Cohort 1	β_{t^*-1}	$p^{(1)}\alpha + \beta_{t^*+1}$	$p^{(1)}\alpha + \beta_{t^*+2}$
Cohort 2	β_{t^*}	$p^{(2)}\alpha + \beta_{t^*+2}$	$p^{(2)}\alpha + \beta_{t^*+3}$
Cohort 3	β_{t^*+1}	$p^{(3)}\alpha + \beta_{t^*+3}$	$p^{(3)}\alpha + \beta_{t^*+4}$

Table 2. *Mean earnings history of three adjacent cohorts assuming that the training effect differs across time periods*

	Earnings in year before training	Earnings in first year after training	Earnings in second year after training
Cohort 1	β_{t^*-1}	$\beta_{t^*+1} + p^{(1)}\alpha_1$	$\beta_{t^*+2} + p^{(1)}\alpha_2$
Cohort 2	β_{t^*}	$\beta_{t^*+2} + p^{(2)}\alpha_1$	$\beta_{t^*+3} + p^{(2)}\alpha_2$
Cohort 3	β_{t^*+1}	$\beta_{t^*+3} + p^{(3)}\alpha_1$	$\beta_{t^*+4} + p^{(3)}\alpha_2$

periods. Now $p^{(i)}$ is the proportion of cohort i that takes training. We establish the convention that the first cohort takes its training in period t^*.

From the population mean earnings of cohort 3 in the last preprogram year (β_{t^*+1}) and the population mean of the first postprogram year earnings of cohort 1 [$p^{(1)}\alpha + \beta_{t^*+1}$], it is possible to solve for α since $p^{(1)}$ (> 0) is assumed to be known. This estimator is consistent for α if the number of observations in each cross section goes to infinity.

Access to multiple-cohort data allows the analyst to estimate separate training effects for each year of posttraining earnings. Assume that each cohort experiences the same earnings impact α_j when it is j periods removed from training. Then population mean earnings streams of cohorts before and after training are as displayed in Table 2. From the mean preprogram earnings of cohort 3 and the mean first year postprogram earnings of cohort 1 it is possible to consistently estimate α_1. [Recall $p^{(1)}$ is known.] Given α_1, use the first year postprogram earnings of cohort 2 to consistently estimate β_{t^*+2} and use the second year postprogram earnings of cohort 1 to consistently estimate α_2. A second consistent estimator of α_2 uses α_1 in the first year postprogram earnings of cohort 3 to estimate

β_{t^*+3} and then uses this information coupled with the second year post-program earnings of cohort 2 to estimate α_2. The fact that α_1 and α_2 are overidentified suggests that it is possible to permit limited interactions among the α_j and the period effects.

4.3 *Allowing for regressors*

The methods described above can be modified to apply to data in which (1.1) rather than (1.3) characterizes the earnings function so that $\mathbf{X}_{it}$ appears in the earnings function. To see how this can be done, we rewrite equation (1.1) to isolate β_t from the other components of $\boldsymbol{\beta}$:

$$\begin{aligned} Y_{it} &= \beta_t + \mathbf{X}_{it}\boldsymbol{\pi} + d_i\alpha + U_{it}, \qquad t > k \\ Y_{it} &= \beta_t + \mathbf{X}_{it}\boldsymbol{\pi} \qquad\quad + U_{it}, \qquad t \leq k \end{aligned} \tag{4.2}$$

Two methods can be used to extend the preceding analysis to account for regressors.

4.3.1. Method I: regression on preprogram earnings. Provide that (A-6) is satisfied [replacing $\boldsymbol{\beta}$ with $(\beta_t, \boldsymbol{\pi})$ in those conditions] for one year of preprogram data, it is possible to consistently estimate $\boldsymbol{\pi}$ and adjust the means of postprogram earnings appropriately (i.e., transform Y_{it} to $Y_{it} - \mathbf{X}_{it}\boldsymbol{\pi}$). The adjusted means can be used to consistently estimate the population means used above. This strategy fails if there are elements of $\mathbf{X}_{it}$ that become nonconstant only after period k.

4.3.2. Method II: use of sufficiently long postprogram repeated cross sections. In place of method I or as a supplement, from the means of earnings of successive cross sections it is possible to solve out for $\boldsymbol{\pi}$ using standard method-of-moments procedures. (If the mean of the $\mathbf{X}_{it}$, $\bar{\mathbf{X}}_t$ does not change over time, then $\bar{\mathbf{X}}_t\boldsymbol{\pi}$ is absorbed into β_t.) This method is just a variant of the procedure already proposed for sufficiently regular environments. Given $\hat{\boldsymbol{\pi}}$ so obtained it is possible to adjust the population mean earnings in the manner described in method I. Note that method II is robust to mean zero measurement error in $\mathbf{X}_{it}$ but that method I is not.

4.4 *Robustness to contamination bias and other measurement error*

Suppose the analyst has access to sample (iii) data as described in Section 3.8–random samples of the population in which the training status of persons is not known. Given (A-13b) so that the population proportion of trainees in each cohort is known, such contaminated samples can be used in all of the procedures proposed in Section 4. In none of these procedures is it necessary to know the training status of individuals. In addition, all of the repeated cross-section estimators based on sample means are robust to mean zero measurement error in *all* of the variables of the model.

4.5 *Robustness to choice-based sampling*

As noted in Section 1, unless (A-2) is satisfied, the repeated cross-section estimator that does not exploit knowledge of the training status of individuals is not robust to choice-based sampling. However, it is sometimes possible to satisfy (A-2) even in the presence of choice-based sampling.

Suppose, for example, that (A-3) strengthened to include the appropriate $\mathbf{X}_{it}$ in the conditioning set characterizes the earnings equation and that $p_t^* = p_{t'}^*$ in the notation of condition (A-2) (also strengthened to include the exogenous variables). Then the strengthened form of (A-2) is satisfied and $\hat{\alpha}_{RC}$ can be used to produce a consistent estimator of α provided one of the sets of conditions about the environment that have been stated above is satisfied for a model written in differences. For example, a permanent–transitory error structure for the unobservable in earnings equation (1.9) coupled with an independence assumption between S_t in decision rule (2.4) and ε_{it} in (1.9) produces a model that satisfies the strengthened form of (A-3). Assuming the sampling rule is such that trainees have the same chance of being observed in t and t', strengthened condition (A-2) is satisfied, and it is possible to consistently estimate α from differences in means between preprogram and postprogram earnings data even if the available samples are choice-based samples provided that the sample proportion of trainees is known for each sample.

This example is special. In general, repeated cross-section estimators do not consistently estimate α in choice-based samples even if $\beta_t = \beta_{t'}$. If the training status of persons were known, it would be possible to reweight the data in the manner described in Section 3.8.

5 Longitudinal and repeated cross-section estimators – methods that exploit information about the training status of individuals

5.1 *Introduction*

In this section of the chapter we present longitudinal methods for consistently estimating α. A major conclusion established here is that many longitudinal estimators have repeated cross-section counterparts provided that the training status of individuals is known.

All of the estimators presented in this section share the following features: (1) They are robust to general forms of time inhomogeneity of the environment provided that there are data on the earnings of both trainees and nontrainees. (2) They require an explicit characterization of the time series process of the unobservables in the earnings equation. (3) Except for one estimator, the proposed estimators require specification of an enrollment rule and the stochastic relationship of the observables and unobservables in the enrollment rule with the observables and unobservables in the earnings equation. (4) No regressor need appear in the enrollment rule (1.2). (5) Many of the estimators are control function estimators in the sense of (D-2). All of these estimators are robust to choice-based sampling. All of the estimators proposed here can be modified, when necessary, to control for choice-based sampling and contamination bias. In addition to these features, some of the estimators do not require data on preprogram earnings.

The plan of this section is as follows. Starting with the conventional fixed effect estimator, we present a variety of assumptions about the relationship between the unobservables in the earnings equation and the enrollment dummy d_i. These assumption sets are presented in increasing level of generality as far as this is possible (some sets of assumptions do not properly contain nor are contained in other sets). We present the estimators in this fashion because the "fixed effect" or "first-difference" estimator is also the most widely used estimator. For each assumption set we first assume access to simple random or stratified samples and give the longitudinal estimator. The plausibility of the identifying assumptions is evaluated in the light of the decision rules given in Section 2. We then present the repeated cross-section estimator where possible. Finally we discuss robustness to choice-based sampling and contamination bias.

5.2 *First-difference or fixed effect methods*

The first-difference (or fixed effect) method was developed by Mundlak (1961, 1978) and refined by Chamberlain (1982). It is based on the following assumptions:

(A-15) (a) (A-6) holds, stated in terms of differences in variables rather than levels.
(b) $E(U_{it} - U_{it'} | d, \mathbf{X}_{it} - \mathbf{X}_{it'}) = 0$ for all $t, t', t > k > t'$.
(c) There is access to at least one year of preprogram and postprogram earnings.

Notice that regressors are not required to appear in the enrollment equation (1.2). As a consequence of (A-15) we may write the difference

regression as

$$E(Y_{it} - Y_{it'} | \mathbf{X}_{it} - \mathbf{X}_{it'}, d_i) = (\mathbf{X}_{it} - \mathbf{X}_{it'})\boldsymbol{\beta} + d_i\alpha, \qquad t > k > t'$$

Here $\boldsymbol{\beta}$ is defined to include the coefficients of year dummies. 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 for α.

5.2.1. Economic models producing (A-15b). The statistical logic justifying this procedure is impeccable. We question the plausibility of (A-15) in the light of the prototypical enrollment rules presented in Section 2.

Some decision rules and error processes for earnings justify (A-15). For example, consider a certainty environment in which α is a fixed coefficient and error structure (1.9) characterizes the earnings residual so that

$$U_{it} = \phi_i + \varepsilon_{it} \tag{5.1}$$

where ε_{it} is a mean zero finite variance random variable independent of all other values of $\varepsilon_{it'}$, for all i, t, and t', and is distributed independently of ϕ_i, a mean zero finite variance person-specific time-invariant random variable. Assuming that S_i in decision rule (2.4) and $\mathbf{X}_{ij}$ are distributed independently of all ε_{it} for all i and j except possibly for ε_{ik}, (A-15b) is satisfied.

Assumption (A-15b) may also be satisfied in an environment of uncertainty. Suppose in error structure (5.1) that

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

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 and $\mathbf{X}_{it}$ are distributed independently of all ε_{it} except possibly for ε_{ik}, this model also produces (A-15b).

Another example of a model that satisfies (A-15b) is a random coefficient model (1.14) in an environment of perfect certainty in which no regressor appears in decision rule (1.2) so that only α^* is estimable. Provided that the error structure for U_{it} is given by (5.1) and S_i in (2.6) and ε_i are distributed independently of ε_{it} (except possibly for ε_{ik}) and the $\mathbf{X}_{ij}$ are distributed independently of ε_{it} for all j and t, (A-15b) is satisfied.[23]

A final example of a model that satisfies (A-15b) is a random coefficient model in which ε_i is unknown at the time enrollment decisions are made

with mean zero, (5.1) characterizes the error process for earnings, and S_i in (2.6) and $\mathbf{X}_{ij}$ are distributed independently of ε_{it} for all t and j (except for ε_{ik}). In this case $\bar{\alpha}$ is identifiable from the first-difference estimator.

However, these examples are rather special. It is very easy to produce plausible models that do not satisfy (A-15b). For example, even if (5.1) characterizes U_{it}, if S_i in (2.4) or (2.6) does not have the same joint (bivariate) distribution with respect to all ε_{it}, except for ε_{ik}, (A-15b) is violated.

Even if S_i in (2.2) is distributed independently of U_{it} for all t, it is still not the case that (A-15b) is satisfied in a general model. For example, assume $\mathbf{X}_{ij}$ is distributed independently of all U_{it} for all t and j and let

$$U_{it} = \rho U_{i,t-1} + \varepsilon_{it} \tag{5.2}$$

where ε_{it} is a mean zero, i.i.d. random variable and $|\rho| < 1$.[24] If $\rho \neq 0$ and perfect-foresight decision rule (2.4) characterizes enrollment, (A-15b) is not satisfied because, for $t > k > t'$,

$$\begin{aligned} E(U_{it}|d_i = 1) &= E\left(U_{it}|U_{ik} + \mathbf{X}_{ik}\boldsymbol{\beta} - \frac{\alpha}{r} < S_i\right) \\ &= \rho^{t-k}E(U_{ik}|d_i = 1) \\ &\neq E(U_{it'}|d_i = 1) = E\left(U_{it'}|U_{ik} + \mathbf{X}_{ik}\boldsymbol{\beta} - \frac{\alpha}{r} < S_i\right) \end{aligned}$$

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

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

and so $E(U_{it} - U_{it'}|d_i = 1) = 0$ only for t', t such that $k - t' = t - k$. Thus (A-15b) is not satisfied for all $t > k > t'$.

With more general specifications of U_{it} and the stochastic dependence between S_i and U_{it}, (A-15b) will not be satisfied.[25]

5.2.2. The repeated cross-section version. Longitudinal data are not required to implement the fixed effect estimator. However, as noted in Section 1, the training status of individuals sampled in a preprogram year is more likely to be known in longitudinal data than in repeated cross-section data. The repeated cross-section version of the estimator was discussed in Section 1.3. Required modifications when regressors appear in earnings function (1.1) are presented in Section 4.3. If the random sampling assumption of that section is dropped and there is a choice-based sample, the regression required to implement the required modifications must be weighted as discussed in Section 3.8. Note that the repeated cross-section

estimator based on means is robust to mean zero measurement error in all of the variables.

5.2.3. Robustness to choice-based sampling and contamination bias. Since (A-15b) is satisfied, the fixed effect estimator is robust to choice-based sampling when the data are transformed to differences. Samples (i) and (ii) defined in Section 3.8 may be combined freely without affecting the consistency of the estimator.

Note further that if conditions (A-15) are satisfied, it is possible to consistently estimate α using only trainee data [sample (i)] provided that (a) the environment is time-homogeneous or (b) there is sufficient regularity in the environment (in the sense of Section 4.2) or (c) there is access to multiple-cohort data. Either longitudinal or repeated cross-section methods can be used. The benefit of having data on nontrainees is that they allow the analyst to control for more general forms of time inhomogeneity if (b) is not satisfied and multiple-cohort data are not available. If, in addition to the sample (i) data on trainee earnings, the analyst has access to a contaminated control sample (iii), it is unnecessary to use these data if the sample (i) data satisfy one of the three conditions listed above.

If the sample (i) data do not satisfy these conditions and the model contains no regressor, it is still possible to consistently estimate $\beta_t - \beta_{t'} + \alpha$, $t > k > t'$ by fitting a difference regression on sample (i). The probability limit of the mean of sample (iii) differences converges to

$$\text{plim}(\bar{Y}_t^{(\text{iii})} - \bar{Y}_{t'}^{(\text{iii})}) = \beta_t - \beta_{t'} + p\alpha.$$

Given knowledge of p, it is thus possible to combine these two pieces of information to consistently estimate α using single-cohort data without invoking any assumption about time homogeneity of the environment. These procedures can be modified in a straightforward way if there are regressors in equation (1.1).[26]

5.2.4. An unconditional version. In place of (A-15b) it is possible to state weaker conditions and still secure a consistent first-difference estimator for α:

(A-15b′) (i) $E[(\mathbf{X}_{it} - \mathbf{X}_{it'})'(U_{it} - U_{it'})] = 0, \quad t > k > t'.$
(ii) $E[d_i(U_{it} - U_{it'})] = 0, \quad t > k > t'.$

In stating these conditions, it is to be understood that time-invariant variables are deleted from the model. Under these conditions, the first-difference estimator consistently estimates α. All of the models that rationalize (A-15b) also rationalize (A-15b′).

5.3 *More general first-difference methods*

In place of (A-15) we assume the following:

(A-16) (a) (A-6) holds, modified to allow $\mathbf{Z}_i$, the regressor in the enrollment rule, to be a constant, and written in terms of differences of variables rather than levels.
(b) $E(U_{it} - U_{it'} | d_i, \mathbf{X}_{it} - \mathbf{X}_{it'}) = 0$ for *some* t and t' such that $t > k > t'$.
(c) There is access to at least one year of preprogram and postprogram earnings.

The only new idea embodied in this assumption is that in place of (A-15b), in which $E(U_{it} - U_{it'} | d_i, \mathbf{X}_{it} - \mathbf{X}_{it'}) = 0$ for all $t > k > t'$, the conditional expectation need be zero only for some $t > k > t'$. For the appropriate values of t and t',

$$E(Y_{it} - Y_{it'} | \mathbf{X}_{it} - \mathbf{X}_{it'}, d_i) = (\mathbf{X}_{it} - \mathbf{X}_{it'})\boldsymbol{\beta} + d_i\alpha$$

so that least squares applied to the differenced data consistently estimates α.

An alternative unconditional version of these conditions, analogous to (A-15b'), writes

(A-16b') (i) $E[(\mathbf{X}_{it} - \mathbf{X}_{it'})'(U_{it} - U_{it'})] = 0$ for some $t > k > t'$.
(ii) $E[d_i(U_{it} - U_{it'})] = 0$ for some $t > k > t'$.

Under these conditions, the generalized first-difference regression estimator consistently estimates α.

5.3.1. Examples of economic models producing (A-16b). We present three examples of models that satisfy (A-16b) but not (A-15b).

5.3.1.A. A multiple-selection-rules model. Suppose that the error structure of the earnings equation is given by (5.1) and that S_i and $\mathbf{X}_{ij}$ are distributed independently of U_{it} for all t and j. Assume an environment of perfect certainty and a fixed coefficient model so that α is common to all individuals.

Individuals use enrollment rule (2.2) to determine whether they would like to enroll in training. Administrators let a person enroll only if his income in period $k - 1$ is less than Y_c. Using the index function notation of Section 2.5,

$$I_{1i} = \frac{\alpha}{r} + S_i - Y_{ik}$$

$$I_{2i} = Y_c \quad - Y_{i,k-1}$$

so that

$$d_i = 1 \quad \text{if and only if} \quad I_{1i} > 0 \quad \text{and} \quad I_{2i} > 0$$
$$E(U_{it} - U_{it'} | d_i, \mathbf{X}_{it} - \mathbf{X}_{it'}) = 0$$

for $t > k$ and $t' < k - 1$. Thus (A-16b) is satisfied if $t > k$ and $t' < k - 1$ but (A-15b) is not satisfied for $t > k$ and $t' = k - 1$. Clearly (A-16b') is also satisfied.

5.3.1.B. Linearity of the regression. Suppose that

(i) U_{it} is covariance stationary [so $E(U_{it}U_{i,t-j}) = E(U_{it'}U_{i,t'-j})$ for all t, t', $j \geq 0$].
(ii) U_{it} has a linear regression on U_{ik} for all t [$E(U_{it}|U_{ik}) = \beta_{tk}U_{ik}$].[27]
(iii) The U_{it} are mutually independent of $(\mathbf{X}_{it}, S_i)$ for all t.
(iv) α is common to all individuals (so the model is of a fixed coefficient form).
(v) The environment is one of perfect foresight, so decision rule (2.2) determines participation.

Under these conditions assumption (A-16b') characterizes the data.

To see this note that (i) and (ii) imply that there exists δ such that

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

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

and

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

Note that

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

But

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

since $E(\omega_{it}) = 0$ and because (iii) ensures that the mean of ω_{it} does not depend on $\mathbf{X}_{ik}$ and S_i.[28]

Similarly,

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

and thus (A-16b') holds.

It is straightforward to show that if, in addition, the $\mathbf{X}_{it}$ are mutually independent of all t and for each i and are independent of S_i for all t, (A-16b) holds.

5.3.1.C. Pivotal symmetry. Conditions (i)–(ii) in the previous section are not required to justify (A-16b) or (A-16b'). Pivotal symmetry also implies these conditions.

(D-4) A collection of continuous random variables $\{U_{it}\}_{t=-\infty}^{\infty}$ is pivotally symmetric with respect to U_{ik} if for all values of $(U_{i,k+j}, U_{ik}, U_{i,k-j})$, $f_j(U_{i,k+j}, U_{ik}) = f_j(U_{i,k-j}, U_{ik})$ for all j.

Replacing (i) and (ii) of the previous subsection with a pivotal symmetry assumption and retaining the rest of the assumptions produces a model that satisfies (A-16b'). If, in addition, it is assumed that the $\mathbf{X}_{it}$ are i.i.d. for all t and for each i, (A-16b) is satisfied. Condition (iii) is stronger than is required to generate (A-16b'). Provided that $(\mathbf{X}_{ik}, S_i)$ have the same joint distribution with respect to $U_{i,k+j}$ as with respect to $U_{i,k-j}$, (A-16b) and (A-16b') are satisfied. Independence is not required.

Because the more general first-difference method is merely a variant of the first-difference method, the discussion of repeated cross-section versions and robustness to contamination bias and choice-based sampling presented in Section 5.2 applies to the methods discussed in Section 5.3. For the sake of brevity we do not repeat it here.

5.4 *Control function estimators*

In this subsection we propose three longitudinal control function estimators. The first two estimators satisfy the following conditions:

(A-17) (a) There exists a control function K_{it} that satisfies definition (D-2).
(b) Including the control function among the regressors $(\mathbf{X}_{it})$ in (1.1), (A-6) is satisfied.

Under these conditions the control function estimator is consistent for α.

The third estimator replaces (A-17a) with the following:

(A-17a') There exists a control function that satisfies definition (D-1).

5.4.1. U_{it} follows a generalized first-order autoregressive process. We suppose that

(i) U_{it} is a first-order autoregression $U_{it} = \rho U_{i,t-1} + v_{it}$, where $E(v_{it}) = 0$, $\mathrm{Var}(v_{it}) < \infty$ and the v_{it} are mutually independently (not necessarily identically) distributed random variables with $\rho \neq 1$.
(ii) Enrollment is determined by perfect-foresight rule (2.4), and α is common to all individuals.
(iii) The v_{ij}, $t' < j \leq t$, are distributed independently of S_i and $\mathbf{X}_{ik}$ in (2.4).

Heckman and Wolpin (1976) invoke these assumptions in their analysis of affirmative action programs.

Then

$$K_{it} = \rho^{t-t'} U_{it'}, \qquad t > t' > k, \qquad \rho \neq 1$$

is a valid control function which satisfies (A-17). The proof is by direct substitution using (1.1) to solve for $U_{it'}$ to obtain

$$Y_{it} = [\mathbf{X}_{it} - (\mathbf{X}_{it'}\rho^{t-t'})]\boldsymbol{\beta} + (1 - \rho^{t-t'})d_i\alpha + \rho^{t-t'}Y_{it'} + \left\{\sum_{j=0}^{t-(t'+1)} \rho^j v_{i,t-j}\right\} \tag{5.3}$$

As a consequence of conditions (i)–(iii),

$$E(Y_{it}|\mathbf{X}_{it}, \mathbf{X}_{it'}, d_i, Y_{it'}) = [\mathbf{X}_{it} - (\mathbf{X}_{it}'\rho^{t-t'})]\boldsymbol{\beta} + (1 - \rho^{t-t'})d_i\alpha + \rho^{t-t'}Y_{it'} \tag{5.4}$$

so that (nonlinear) least squares applied to (5.3) consistently estimates α as the number of observations becomes large. (The appropriate nonlinear regression imposes the implied cross-coefficient restrictions.)

Notice that (iii) is overly strong. The v_{ij}, $t' < j < t$, need not be distributed independently of S_i and $\mathbf{X}_{ik}$ in (2.4). They need only satisfy the condition that (5.4) is the conditional expectation of (5.3) [so that $E(\sum_{j=0}^{t-(t'+1)} \rho^j v_{i,t-j} | \mathbf{X}_{it}, \mathbf{X}_{it'}, d_i, Y_{it'}) = 0$]. Note further that (ii) is not required either. If agents are uncertain about future v_{it} at the time they make their enrollment decision, (ii) may be replaced with decision rule (2.7). In this case $K_{it} = \rho^{t-t'}U_{it'}$ is a valid control function for $t > t' > k$ or $t > k$, $t' = k - 1$.

5.4.2. U_{it} follows a higher-order autoregression. The estimator considered in Section 5.4.1 may be extended to the case where U_{it} follows a higher-order autoregression. Assume in addition to (A-17) that

(i) $U_{it} = \sum_{j=1}^{N} \rho_j U_{i,t-j} + v_{it}$, where $E(v_{it}) = 0$, $\mathrm{Var}(v_{it}) < \infty$, and the v_{it} are mutually independently (not necessarily identically) distributed and $\sum_{j=1}^{N} \rho_j \neq 1$.
(ii) Enrollment is determined by perfect-foresight rule (2.4) and α is common to all individuals.
(iii) v_{it} is distributed independently of S_i and $\mathbf{X}_{ik}$ in (2.4), for $t > k$.
(iv) $t \geq k + N + 1$.

Then

$$K_{it} = \rho_1(Y_{i,t-1} - \mathbf{X}_{i,t-1}\boldsymbol{\beta} - d_i\alpha) + \cdots + \rho_N(Y_{i,t-N} - \mathbf{X}_{i,t-N}\boldsymbol{\beta} - d_i\alpha)$$

is a control function in the sense of (D-2) because

$$E(v_{it}d_i) = 0$$

The appropriate estimating equation is

$$Y_{it} = \left(\mathbf{X}_{it} - \sum_{j=1}^{N} \rho_j \mathbf{X}_{i,t-j}\right)\boldsymbol{\beta} + \sum_{j=1}^{N} \rho_j Y_{i,t-j} + \left(1 - \sum_{j=1}^{N} \rho_j\right)d_i\alpha + v_{it} \tag{5.5}$$

Nonlinear least squares applied to (5.5) consistently estimates $\rho_1, \ldots, \rho_N$, β, and α as the number of persons in the longitudinal sample becomes large.

Assumption (iii) is overly strong. All that is required is that $E(v_{it}|\mathbf{X}_{it}, \mathbf{X}_{i,t-1}, \ldots, d_i, Y_{it-1}, \ldots, Y_{i,t-N}) = 0$. The estimator can be adapted to an uncertain environment using an argument directly analogous to that presented for the first-order autogressive model.

5.4.3. 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

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

that is, 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:

(ii) The relevant earnings history goes back N periods before period k.
(iii) The enrollment decision is characterized by rule (2.7).
(iv) S_i and $\mathbf{X}_{ik}$ are known as of period $k-1$ when the enrollment decision is being made.
(v) $\mathbf{X}_{it}$ is distributed independently of U_{ij} for all t and j.
(vi) S_i is distributed independently of U_{ij} for all j.

Defining

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

and

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

then assuming conditions (i)–(v) above,

$$K_{it} = c(G(\psi_i) - p)$$

is a valid control function in the sense of definition (D-1), where

$$p = E(d_i)$$

and

$$c = \frac{E[U_{it}(G(\psi_i) - p)]}{E[G(\psi_i) - p]^2} \tag{5.6}$$

To establish that $c(G(\psi_i) - p)$ is a valid control function, it is helpful to write (1.1) in the following way:

$$Y_{it} = \mathbf{X}_{it}\beta + d_i\alpha + c(G(\psi_i) - p) + [U_{it} - c(G(\psi_i) - p)] \tag{5.7}$$

In the transformed equation

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

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

Thus it remains to show that

$$E[d_i(U_{it} - c(G(\psi_i) - p))] = 0 \tag{5.8}$$

Before proving this it is helpful to notice that as a consequence of (i), (v), and (vi),[29]

$$E(d_i|U_{it}, U_{i,t-1}, \ldots, U_{i,k-1}, \ldots, U_{i,k-N}) = E(d_i|U_{i,k-1}, \ldots, U_{i,k-N}), \qquad t > k \tag{5.9}$$

Since only preprogram innovations determine participation, and because U_{it} is distributed independently of $\mathbf{X}_{ik}$ and S_i in decision rule (2.7), the conditional mean of d_i does not depend on postprogram values of U_{it} given all preprogram values.

Intuitively, (5.8) follows from (5.9): 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 (5.9).

The proof of the proposition uses the fact from equation (5.9) that $E(d_i|\psi_i, U_{it}) = G(\psi_i)$ in computing the expectation

$$\begin{aligned} E[d_i(U_{it} - c(G(\psi_i) - p))] &= E[E(d_i(U_{it} - c(G(\psi_i) - p))|\psi_i, U_{it})] \\ &= E[(U_{it} - c(G(\psi_i) - p))E(d_i|\psi_i, U_{it})] \\ &= E[(U_{it} - c(G(\psi_i) - p))G(\psi_i)] \\ &= E[(U_{it} - c(G(\psi_i) - p))(G(\psi_i) - p)] \\ &= 0 \end{aligned} \tag{5.10}$$

as a consequence of the definition of c in (5.6).

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

5.4.4. Repeated cross-section versions. The repeated cross-section version of method I has already been discussed in Section 1.3. Preprogram earnings data can be used to consistently estimate $\boldsymbol{\beta}$ to adjust the means. The repeated cross-section version of method II is defined in an analogous fashion. We have been unable to produce a repeated cross-section version of method III. Note that the repeated cross-section estimators, when defined, are robust to mean zero measurement error.

5.4.5. Robustness to choice-based sampling and contamination bias. Methods I and II are control function estimators in the sense of definition D-2. Accordingly, they are robust to choice-based sampling. Methods for applying control function estimators to contaminated data are described in Section 3.8 (see the subsection on contamination bias for control function estimators).

Method III is not robust to choice-based sampling. This is so because

$$E[X'_{it}(U_{it} - c(G(\psi_i) - p))|\psi_i, d_i, \mathbf{X}_{it}] \neq 0 \tag{5.11}$$

and

$$E[(G(\psi_i) - p)(U_{it} - c(G(\psi_i) - p))|\psi_i, d_i, \mathbf{X}_{it}] \neq 0$$

Reweighting the data by weight ω_i introduced in the discussion of the IV method in Section 3.8 and fitting (5.7) using weighted variables produces a consistent estimator as the number of observations becomes large. Method III is also not robust to contamination bias. The moments of samples (i) and (iii) may be combined in a fashion analogous to that described in the discussion of the IV estimator in Section 3.8. The appropriate procedure for combining choice-based and contaminated samples closely parallels the method presented for the IV estimator in Section 3.8, so that a repetition of that discussion is not necessary.

5.5 *Partial K functions*

A partial K function $\tilde{K}_{it}$ satisfies (a) and (b) of definition (D-1) but fails condition (c) and as a consequence fails (d) unless additional prior restrictions are available.

An example of a partial K function can be constructed for the permanent–transitory model satisfying (A-3). In that case, assuming no regressors appear in the earnings function,

$$\tilde{K}_{it} = Y_{it'} - \beta_{t'} = U_{it'}, \qquad t' < k \tag{5.12}$$

and $\psi = \beta_{t'}$. For i.i.d. U_{it} with finite second moments

$$E(U_{it} - U_{it'})U_{it'} \neq 0$$

Least squares applied to

$$Y_{it} = \beta_t + d_i\alpha + \tilde{K}_{it} + (U_{it} - \tilde{K}_{it}) \tag{5.13}$$

is inconsistent. However, utilizing the prior information that the coefficient on $\tilde{K}_{it}$ is unity, we may rewrite (5.13) as a first-difference equation

$$Y_{it} - Y_{it'} = \beta_t - \beta_{t'} + d_i\alpha + U_{it} - U_{it'}$$

As a consequence of (A-3) least squares consistently estimates α. In this example, condition (c) of definition (D-1) is not required to identify α and so its failure is of no consequence.

In general this is not so and a condition like (c) is required to secure identification of α. The partial K function estimator replaces condition (c) of definition (D-1) with the requirement that there exists a vector of valid instrumental variables, $\mathbf{Z}_i^e$. (See, e.g., White, 1984, for one statement of such conditions.)

5.5.1. An example of a $\tilde{K}$ function. We depart from the format of previous sections and merely sketch an example of a $\tilde{K}$ function drawing freely from the work of Madansky (1964), Chamberlain (1977), and Pudney (1982). The reader is referred to those papers for a precise statement of conditions under which the estimator is consistent. Here we simply state the intuitive idea underlying one $\tilde{K}$ function.

The estimator assumes a factor structure for U_{it} so that

$$U_{it} = \pi_{1t}\phi_{1i} + \pi_{2t}\phi_{2i} + \cdots + \pi_{Nt}\phi_{Ni} + v_{it} \quad \text{for all} \quad t \tag{5.14}$$

where

(i) The v_{it} are mean zero mutually independent random variables with $\text{Var}(v_{it}) < \infty$ and they are distributed independently of the ϕ_{ij} for all i and j. $\text{Var}(\phi_{ij}) < \infty$ for all i and j, but the ϕ_{ij} may be freely correlated. The π_{jt} are nonzero bounded constants.

(ii) v_{it} is distributed independently of S_i in perfect-foresight decision rule (2.2) or in imperfect-foresight decision rule (2.7).

(iii) There are at least $2N$ years of earnings data in periods before t and at least one year in a period after t.

Note that if $N = 1$ and $\pi_{1t} = \pi$, the permanent–transitory model (5.1) is produced. If $\pi_{1t} = \pi_1$ and $\pi_{2t} = \pi_2 t$, (5.14) produces a special case of the Lillard–Weiss (1979) permanent component random growth model.

Array any N distinct observations on individual i for periods other than t into a matrix equation system

$$\begin{pmatrix} Y_{ij} \\ \cdots \\ Y_{ij'} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{ij} & d_i l_j \\ \cdots & \cdots \\ \mathbf{X}_{ij'} & d_i l_{j'} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \cdots \\ \alpha \end{pmatrix} + \begin{pmatrix} \pi_{1j} \cdots \pi_{Nj} \\ \cdots \quad \cdots \\ \pi_{1j'} \quad \pi_{Nj'} \end{pmatrix} \begin{pmatrix} \phi_{1i} \\ \cdots \\ \phi_{Ni} \end{pmatrix} + \begin{pmatrix} v_{ij} \\ \cdots \\ v_{ij'} \end{pmatrix}$$

where $l_j = 1$ if $j > k$, $l_j = 0$ otherwise. Define

$$\mathbf{Y}_1 = \begin{pmatrix} Y_{ij} \\ \cdots \\ Y_{ij'} \end{pmatrix}, \qquad \mathbf{X}_1 = \begin{pmatrix} \mathbf{X}_{ij} \\ \cdots \\ \mathbf{X}_{ij'} \end{pmatrix}, \qquad d_i \mathbf{l}_1 = \begin{pmatrix} d_i l_j \\ \cdots \\ d_i l_{j'} \end{pmatrix}, \qquad \mathbf{l}_1 = \begin{pmatrix} l_j \\ \cdots \\ l_{j'} \end{pmatrix}$$

$$\boldsymbol{\pi}_1 = \begin{pmatrix} \pi_{1j} & \pi_{Nj} \\ \cdots & \cdots \\ \pi_{1j'} & \pi_{Nj'} \end{pmatrix}, \qquad \mathbf{v}_{i1} = \begin{pmatrix} v_{ij} \\ \cdots \\ v_{ij'} \end{pmatrix}, \qquad \boldsymbol{\phi}_i = \begin{pmatrix} \phi_{1i} \\ \cdots \\ \phi_{Ni} \end{pmatrix}$$

In more compact notation, the equation system is

$$\mathbf{Y}_1 = [\mathbf{X}_1, d_i \mathbf{l}_1] \begin{bmatrix} \boldsymbol{\beta} \\ \alpha \end{bmatrix} + \boldsymbol{\pi}_1 \boldsymbol{\phi}_i + \mathbf{v}_{i1} \tag{5.15}$$

Assume that

(iv) $\boldsymbol{\pi}_i$ is of full rank.

Then (5.15) may be solved for $\boldsymbol{\phi}_i$ to obtain

$$\boldsymbol{\phi}_i = \boldsymbol{\pi}_1^{-1} \mathbf{Y}_1 - \boldsymbol{\pi}_1^{-1} [\mathbf{X}_1, d_i \mathbf{l}_1] \begin{bmatrix} \boldsymbol{\beta} \\ \alpha \end{bmatrix} - \boldsymbol{\pi}_1^{-1} \mathbf{v}_{i1}$$

Then this expression for $\boldsymbol{\phi}_i$ can be substituted for $\boldsymbol{\phi}_i$ in equation (1.1) with error structure (5.14) in the following way:

$$Y_{it} = \mathbf{X}_{it} \boldsymbol{\beta} + d_i \alpha + (\pi_{1t}, \ldots, \pi_{Nt}) \boldsymbol{\phi}_i + v_{it}$$

$$Y_{it} = \mathbf{X}_{it} \boldsymbol{\beta} + d_i \alpha + (\pi_{1t}, \ldots, \pi_{Nt}) \times \left[\boldsymbol{\pi}_1^{-1} \mathbf{Y}_1 - \boldsymbol{\pi}_1^{-1} (\mathbf{X}_1, d_i \mathbf{l}_1) \begin{pmatrix} \boldsymbol{\beta} \\ \alpha \end{pmatrix} - \boldsymbol{\pi}_1^{-1} \mathbf{v}_{i1} \right] + v_{it}$$

Collecting terms, we get

$$Y_{it} = [\mathbf{X}_{it} - (\pi_{1t}, \ldots, \pi_{Nt}) \boldsymbol{\pi}_1^{-1} \mathbf{X}_1] \boldsymbol{\beta} + d_i \alpha (1 - (\pi_{1t}, \ldots, \pi_{Nt}) \boldsymbol{\pi}_1^{-1} \mathbf{l}_1) + (\pi_{1t}, \ldots, \pi_{Nt}) \boldsymbol{\pi}_1^{-1} \mathbf{Y}_1 + [v_{it} - (\pi_{1t}, \ldots, \pi_{Nt}) \boldsymbol{\pi}_1^{-1} \mathbf{v}_{i1}] \tag{5.16}$$

By virtue of (5.15), $\mathbf{Y}_1$ and the composite error in (5.16) are correlated so that least squares applied to (5.16) is an inconsistent estimator [condition (D-1c) is violated].

Use the remaining N values of the Y_{im}, where $m \neq t$ and the Y_{im} are not elements of $\mathbf{Y}_1$, as instrumental variables for $\mathbf{Y}_1$ in (5.16). They are valid instruments because v_{im} is uncorrelated with the composite error in (5.16) but each Y_{im} is correlated with elements of $\mathbf{Y}_1$ through their common dependence on $\boldsymbol{\phi}_i$.

Provided that standard rank conditions for IV estimators hold, it is possible to consistently estimate $(\pi_{1t}, \ldots, \pi_{Nt}) \boldsymbol{\pi}_1^{-1}$, and hence $\boldsymbol{\beta}$ and α, provided that the coefficient on α in (5.16) does not vanish $[1 - (\pi_{1t}, \ldots, \pi_{Nt}) \boldsymbol{\pi}_1^{-1} \mathbf{l}_1 \neq 0]$. Chamberlain (1977) presents interesting examples of such $\tilde{K}$ functions where the IV rank conditions fail for subtle reasons.

5.5.2. The repeated cross-section version. The methods of Section 4.1 can be used to consistently estimate α from repeated cross-section data when (5.14) characterizes the earnings disturbance. We have not produced a repeated cross-section version of the $\tilde{K}$ estimator that exploits knowledge of the training status of persons and that does not require extra assumptions beyond those given above.

5.5.3. Robustness to choice-based sampling and contamination bias. After transformation, the $\tilde{K}$ function estimator is an IV estimator. Accordingly, our discussion of the modifications required in the cross-section IV estimator to account for choice-based sampling and contamination bias applies to the $\tilde{K}$ function.

5.6 *U_{it} is covariance-stationary*

The final procedure considered in this section 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 the following:

(i) U_{it} is covariance-stationary and so

$$E(U_{it} U_{it-j}) = E(U_{it'} U_{it'-j}) = \sigma_j \quad \text{for} \quad j \geq 0$$

and all t, t'.

(ii) Access to at least two observations on preprogram earnings t', $t' - j$, and one on postprogram earnings t so that $t - t' = j$.

(iii) (A-6) describes the process generating the earnings data for each cross section.

(iv) $pE(U_{it'} | d_i = 1) \neq 0$.

Notably absent from the assumptions is any statement about the appropriate enrollment rule or about the stochastic relationship between U_{it} and the cost of enrollment S_i.

From the preprogram earnings data, it is possible to estimate $\boldsymbol{\beta}$ under the conditions given in Section 4 and adjust the Y_{it} back to the expression

$$\begin{aligned} \tilde{Y}_{it} &= \beta_t + d_i\alpha + U_{it}, \qquad t > k \\ \tilde{Y}_{it'} &= \beta_{t'} \qquad\quad + U_{it'}, \qquad t' < k \end{aligned} \tag{5.17}$$

where β_t and $\beta_{t'}$ are period-specific shifters and $\tilde{Y}_{it}$ is Y_{it} with the effect of the regressors removed.

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

$$m_1 = \frac{\sum(\tilde{Y}_{it'} - \bar{\tilde{Y}}_{t'})(\tilde{Y}_{i,t'-j} - \bar{\tilde{Y}}_{t'-j})}{I}$$

$$\text{plim } m_1 = \sigma_j$$

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 $\tilde{Y}_{it}$ and $\tilde{Y}_{it'}$:

$$m_2 = \frac{\sum(\tilde{Y}_{it} - \bar{\tilde{Y}}_{t})(\tilde{Y}_{it'} - \bar{\tilde{Y}}_{t'})}{I}$$

which has the probability limit

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

From the sample covariance between d_i and $\tilde{Y}_{it'}$

$$m_3 = \frac{\sum(\tilde{Y}_{it'} - \bar{\tilde{Y}}_{t'})d_i}{I}$$

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

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

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

This estimator is not to be confused with another apparently similar one. The alternative estimator notes that least squares applied to estimate α in

$$\tilde{Y}_{it} = \beta_t + d_i\alpha + U_{it}$$

from a random sample of postprogram year t earnings has the property that

$$\text{plim } \hat{\alpha} = \alpha + \frac{1}{1-p}E(U_{it}|d_i = 1)$$

where $\hat{\alpha}$ is the least squares estimator. If a dummy is entered in a random sample of preprogram earnings, the least squares estimator of γ in

$$\tilde{Y}_{it'} = \beta_{t'} + \gamma d_i + U_{it'}, \qquad t' < k$$

converges to

$$\text{plim } \tilde{\gamma} = \frac{1}{1-p}E(U_{it'}|d_i = 1)$$

If U_{it} is, for example, pivotally symmetric with respect to U_{ik}, $t - k = k - t'$, and the joint density of (S_i, U_{it}) is the same as the joint density of $(S_i, U_{it'})$, then

$$\text{plim } (\hat{\alpha} - \tilde{\gamma}) = \alpha$$

The estimator proposed in this section requires only that U_{it} be covariance-stationary (something not required for the alternative estimator) and does not require that any position be taken with regard to the stochastic dependence between the U_{it} and S_i. Note that the variance of U_{it} is *not* required to be identical in all periods.

5.6.1. Repeated cross-section version. For simplicity of exposition we assume that there are no regressors in the earnings function. If regressors are present, we assume that the analyst can adjust Y_{it} to $\tilde{Y}_{it}$ in equation (5.17).

Before presenting the estimator, it is helpful to record the following facts:

$$\text{Var}(Y_{it}) = \alpha^2(1-p)p + 2\alpha E(U_{it}|d_i = 1)p + \sigma_u^2, \qquad t > k \tag{5.18a}$$

$$\text{Var}(Y_{it'}) = \qquad\qquad\qquad\qquad \sigma_u^2 \qquad t' < k \tag{5.18b}$$

$$\text{Cov}(Y_{it}, d_i) = \alpha p(1-p) + pE(U_{it}|d_i = 1) \tag{5.18c}$$

Note that $E(U_{it}^2) = E(U_{it'}^2)$ by virtue of assumption (i) given in Section 5.6. Then

$$\hat{\alpha} = [(p(1-p)]^{-1}\left\{\frac{\sum(Y_{it} - \bar{Y}_t)d_i}{I_t} - \sqrt{\left(\frac{\sum(Y_{it} - \bar{Y}_t)d_i}{I_t}\right)^2 - p(1-p)\left(\frac{\sum(Y_{it} - \bar{Y}_t)^2}{I_t} - \frac{\sum(Y_{it'} - \bar{Y}_{t'})^2}{I_{t'}}\right)}\right\} \tag{5.19}$$

is consistent for α.

This expression arises by substracting (5.18b) from (5.18a). Then use (5.18c) to get an expression for $E(U_{it}|d_i = 1)$, which can be substituted into the expression for the difference between (5.18a) and (5.18b). Replacing

population moments by sample counterparts produces a quadratic equation in α, with the negative root given by (5.19). The positive root is inconsistent for α.

5.6.2. Robustness to choice-based sampling and contamination bias. If the available data are a choice-based sample and the population p is known or can be consistently estimated, the data can be reweighted to form a consistent estimator.

Letting d_i be the value of the variable indicating training status

$$\bar{\tilde{Y}}_t^* = p\frac{\sum d_i \tilde{Y}_{it}}{\sum d_i} + (1-p)\frac{\sum (1-d_i)\tilde{Y}_{it}}{\sum (1-d_i)}$$

and the reweighted moments are

$$m_1^* = p\frac{\sum d_i (\tilde{Y}_{it'} - \bar{\tilde{Y}}_{t'}^*)(\tilde{Y}_{i,t-j} - \bar{\tilde{Y}}_{t-j}^*)}{\sum d_i} + (1-p)\frac{\sum (1-d_i)(\tilde{Y}_{it} - \bar{\tilde{Y}}_t^*)(\bar{\tilde{Y}}_{i,t-j} - \bar{\tilde{Y}}_{t-j}^*)}{\sum (1-d_i)}$$

$$m_2^* = p\frac{\sum d_i(\tilde{Y}_{it'} - \bar{\tilde{Y}}_{t'}^*)(\tilde{Y}_{it} - \bar{\tilde{Y}}_t^*)}{\sum d_i} + (1-p)\frac{\sum (1-d_i)(\tilde{Y}_{it'} - \bar{\tilde{Y}}_{t'}^*)(\tilde{Y}_{it} - \bar{\tilde{Y}}_t^*)}{\sum (1-d_i)}$$

$$m_3^* = p\frac{\sum d_i(\tilde{Y}_{it'} - \bar{\tilde{Y}}_{t'}^*)}{\sum d_i}$$

and so

$$\operatorname{plim}\frac{(m_2^* - m_1^*)}{m_3^*} = \alpha \qquad \text{if } pE(U_{it'}|d_i = 1) \neq 0$$

If regressors appear in the equations, $\boldsymbol{\beta}$ can be consistently estimated from a regression on preprogram earnings using data weighted by ω_i introduced in Section 3.8, and the $\tilde{Y}_{it}$, $\tilde{Y}_{it'}$ and $\tilde{Y}_{i,t'-j}$ can be formed from Y_{it}, $Y_{it'}$, and $Y_{i,t'-j}$, respectively, using the procedure described in Section 4.

The covariance-stationary procedure can also be modified to solve the problems raised by contamination in the control group. Sample (iii) data can be used to consistently estimate m_1 and m_2.

Let $\bar{\tilde{Y}}_{t'}^{(\mathrm{i})}$ be the sample mean of $\tilde{Y}_{it'}$ in sample (i) data and let $\bar{\tilde{Y}}_{t'}^{(\mathrm{iii})}$ be the sample mean of $\tilde{Y}_{it'}$ in sample (iii) data. Then

$$\operatorname{plim}\left[p\left(\bar{\tilde{Y}}_{t'}^{(\mathrm{i})} - \bar{\tilde{Y}}_{t'}^{(\mathrm{iii})}\right)\right] = pE(U_{it'}|d_i = 1)$$

and

$$\operatorname{plim}\left[\frac{m_2 - m_1}{p\left(\bar{\tilde{Y}}_{t'}^{(\mathrm{i})} - \bar{\tilde{Y}}_{t'}^{(\mathrm{iii})}\right)}\right] = \alpha$$

No postprogram data from sample (i) are required.

6 Summary and conclusions

This chapter has presented alternative methods for estimating the impact of training on earnings when nonrandom selection characterizes the enrollment of persons into training. The analysis of this problem serves as a prototype for the analysis of such closely related problems as estimating the impact of schooling, unionism, migration, and job turnover on earnings in the presence of a selection rule determining participation in those activities.

One contribution of this chapter has been to 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 as the impact of training on the earnings of the trained, that is, the component of the increment in posttraining earnings attributable to training including the effect of enrollment rules on selecting people into training who have a more or less typical response to training. The two notions come to the same thing only when training has the same impact on everyone – that is, when a fixed (and not a random) coefficient earnings model describes the data, or when assignment to training is random. 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.

The first notion accords with the more commonly utilized definition of a "structural" impact of training on earnings and is a special case of a more general notion that defines the parameter of interest to be the impact of training on the earnings of the trained in the future if the future selection rule for trainees differs from previous selection rules.

This chapter presents a variety of new estimators for both versions of the parameter of interest. By considering the assumptions required to use the new estimators and more conventional ones to address the same problem using longitudinal, repeated cross-section, and cross-section data, we have explored the benefits of panel data and repeated cross-section data. We have also investigated the plausibility of assumptions required to justify various 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 (i.e., contamination bias) is pervasive in many available control samples, we have examined the robustness of all of the estimators discussed in this chapter to choice-based sampling and contamination bias.

We have reached the following main conclusions:

1. Unless distributional assumptions are invoked, methods based solely on cross-section data require at least one regressor in the decision rule analogous to the shifter variables required to secure identification of demand and supply functions from market data. Longitudinal and repeated cross-section methods do not require a regressor. This is a major benefit of access to multiple cross-section or longitudinal data.

2. However, given a regressor in the decision rule, if some function of it satisfies the exogeneity condition presented in Section 3.2 and if additional mild rank conditions are satisfied, the regressor can be used as an instrumental variable for the training status dummy in a fixed coefficient cross-section earnings equation. No further distributional or functional form assumptions are required in order to identify the structural impact of training on earnings, although such assumptions have frequently been invoked in the recent literature. Such assumptions *are* required for consistent estimation of random coefficient earnings functions produced from Roy (1951) choice models. The instrumental variable estimator does not consistently estimate the random coefficient earnings function in the presence of selection bias.

Because the IV estimator requires such weak assumptions, it produces consistent estimators of program impact for a wide variety of models of the enrollment decision provided that the earnings equation is of the fixed coefficient type. Virtually all of the other cross-section estimators require prior knowledge of the distribution of unobservables or of the functional form of the conditional mean of the unobservable in the earnings equation. The only exception to this rule is the two-stage method of Section 3.3 in which the choice probability can sometimes be estimated nonparametrically. The analysis of the decision rules of Section 2 provides little guidance on the choice of such distributions and functional forms. (However, the Barnow–Cain–Goldberger procedure that assumes selection on the basis of observables makes assumptions that appear to be implausible in the light of the prototypical enrollment rules of Section 2.)

3. Without invoking distributional assumptions, it is not possible to identify the population mean response to training in a random coefficient model of earnings unless a regressor appears in the enrollment rule or unless an individual does not know his own response to training at the time

enrollment decisions are made and the population mean forecast error is known to the econometrician. Accordingly, one benefit of longitudinal and repeated cross-section estimators that accrues in fixed coefficient earnings models vanishes in a random coefficient earnings model.

4. Provided that the population proportion of trainees is known or can be consistently estimated, all of the estimators presented in this chapter can be adapted to consistently estimate the structural parameters of interest from contaminated samples. Some require no modification at all (e.g., the direct nonlinear regression estimator of Section 3.3 or the repeated cross-section estimator of Section 4).

5. Provided that the training status of individuals is known, all of the estimators can be adapted to produce consistent estimators of program impact from choice-based samples. The control function estimators [in the sense of definition (D-2) in Section 1.5] require no modification at all. We have presented examples of cross-section, repeated cross-section, and longitudinal control function estimators.

6. Provided that the environment is time-homogeneous or sufficiently regular in the sense of definition (D-3) in Section 4, and provided that simple random samples are available, the repeated cross-section estimator which does not require that the training status of persons be known (but does require that the population proportion of trainees be known or consistently estimable) is robust to completely general specifications of the enrollment rule. In particular, no regressor need appear in the enrollment rule. No special assumptions need be invoked with respect to the stochastic dependence relationships among variables in the earnings and enrollment equations in order for this estimator to produce consistent estimators of program impact. Multiple-decision rules may characterize enrollment. The estimator is robust to contamination bias provided that the population proportion of trainees is known. For these reasons, this repeated cross-section estimator is quite attractive.

However, the estimator is inconsistent when applied to choice-based samples and to environments characterized by general forms of time inhomogeneity. To account for the problems induced by choice-based sampling and general forms of time inhomogeneity it is necessary to know the training status of persons.

7. The benefits from longitudinal data have been overstated in the recent literature. Many consistent estimators thought to be uniquely longitudinal in nature can in fact be implemented using repeated cross-section data on unrelated persons. In particular, the widely used fixed effect estimator can be employed to secure consistent estimators from repeated cross-section data. However, to use repeated cross-section data to implement the fixed effect estimator requires that the training status of persons be

known in preprogram cross sections. Longitudinal data with preprogram earnings observations will have this information while repeated cross-section data may not.

8. The longitudinal estimators secure identification of the impact of training on earnings by making assumptions about the time series processes of the unobservables and observables in the enrollment and earnings equations. The only exception to this rule is the covariance-stationary estimator of Section 5.6, which requires covariance stationarity for the unobservable in the earnings equation but requires no explicit specification of the enrollment rule. While some of these assumptions may be tested (e.g., assumptions about the time series process of the unobservables in the earnings equation), others cannot, in general, be tested (e.g., assumptions about independence between the unobservables in the enrollment and earnings equations).

Longitudinal estimators require different assumptions than cross-sectional estimators, and it is not obvious which sets of assumptions are more plausible. As noted by Chamberlain (1984), longitudinal data can sometimes be used to test cross-sectional identifying assumptions. But this is not always so. The sets of identifying assumptions for the two types of estimators are not necessarily nested (e.g., the existence of an IV estimator for the cross-section estimator and covariance stationarity of the earnings equation for the longitudinal estimator).

9. The specification of the earnings equation and the enrollment rule that justify the widely used fixed effect or first-difference estimator are very special and are not of general interest. Contrary to recent claims, fixed effect methods offer no panacea for selection and simultaneity problems that arise in estimating the impacts on earnings of training, unionism, job turnover, and migration.

10. The cross-section estimators do not require preprogram earnings data, nor do most of the longitudinal estimators. Two exceptions are the fixed effect longitudinal estimator and the covariance-stationary estimator of Section 5.6.

11. Virtually 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. 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 that justify use of the fixed effect estimator.

12. The estimators differ in their robustness to aging and decay effects (i.e., time subscripted α). The cross-section estimators are robust in the sense that they identify the value of α_t appropriate to a particular cross section. The repeated cross-section estimator with the training status of persons unknown requires a strengthened definition of a sufficiently regular

environment if the α values change over time [see definition (D-3) in Section 4]. The longitudinal estimators require straightforward modifications to account for aging and decay effects.

13. The covariance-stationary estimator of Section 5.6 and the repeated cross-section estimator of Section 4 can be applied without modification to models with multiple enrollment rules. The IV estimator of Section 3.2 does not require an explicit statement of the enrollment rule – just a list of valid instruments. All of the other estimators must be modified if there are multiple selection rules.

Throughout this chapter we have deliberately avoided discussing the efficiency of alternative estimators. Many of the estimators presented here invoke different assumptions about the true model generating the data. An efficiency comparison for such estimators is meaningless because the assumptions required to justify one estimator do not justify another. Only by postulating a common assumption set bigger than what is required to justify any single estimator is it possible to compare such estimators. The required common set depends on the pair of estimators being considered. The outcome of such efficiency comparisons will hinge on specific values assumed by parameters and the value of making such comparisons is not obvious.

Even if a common set of assumptions about the underlying model were 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 economic analyses.[30] In contrast, the *available* cross-section samples often have hundreds of 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 chapter 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.

Nonlinear (longitudinal and cross-section) estimators that are computationally more demanding are often implemented on only a small fraction of the available data in order to economize on computer costs. The relevant

comparison of the efficiency of apparently inefficient mean estimators recognizes that in most applications mean estimators use all the available data, whereas more sophisticated estimators use only a fraction of the data.

When we began writing this chapter we intended it to be a paean to longitudinal data; that it is not so is because the analysis presented here does not warrant such a conclusion. More data are preferred to less given zero cost, and different types of data are always useful. In certain cases, longitudinal data can be used to test assumptions maintained in cross-section work.

But a key conclusion of our analysis is that the benefits of longitudinal data have been overstated in the recent econometric literature 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 often can 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 the robust cross-section estimators.

ACKNOWLEDGMENTS

This research was supported by grants from the Department of Labor (DOL 20-17 82-20), National Science Foundation Grants SOC77-27136 and SES-8107963 and NIH-1-R01-HD16846-01 to the Economics Research Center/NORC at the University of Chicago. The first draft was written while Heckman was a fellow at the Center for Advanced Studies in the Behavioral Sciences in the year 1978 while supported, in part, by a fellowship from the J. S. Guggenheim Foundation. The first draft was read at a Social Science Research Council conference at Mt. Kisco, New York, October 1978. The second draft was read at a conference on Panel Data at London School of Economics in London, June, 1982. We are grateful for comments received at seminars at North Carolina State, M.I.T., Yale, Penn, Texas, and Michigan. We have benefited from comments received from John Abowd, Burt Barnow, Joe Hotz, William Kruskal, Robert Michael, Tom Stoker, Robert Tamura, Grace Tsiang, Jim Walker, and Adonis Yatchew. Ricardo Barros made especially helpful comments.

NOTES

1 Duncan, Juster, and Morgan (1982) claim that panel data are cheaper to collect per observation than cross-section data, but they do not present detailed evidence of their claim. They also do not consider the problem of nonrandom attrition that plagues panel data but not cross-section data.

2 Data on residual variances do not aid in securing identification unless the variance is known a priori. The sum of the squared normalized residuals converges to an expression that combines $\text{Var}(U_{it})$ with $\text{Cov}(U_{it}, d_i)$. Without prior information there is no information on α from higher moments.

3 Thus we assume that in large samples the sample frequency of $(\mathbf{X}, \mathbf{Z}, U)$ given d converges to the population density $f(\mathbf{X}, \mathbf{Z}, U \mid d)$.

4 Because in that case $E(\hat{\alpha}_{RC}) = \beta_t - \beta_{t'} + (p_t^* - p_{t'}^*)\alpha$. If $\beta_t = \beta_{t'}$ and the proportion of trainees is known for t' and t, then α is identified from $\bar{Y}_t$ and $\bar{Y}_{t'}$ by dividing $\bar{Y}_t - \bar{Y}_{t'}$ by $p_t^* - p_{t'}^*$ assuming $p_t^* \neq p_{t'}^*$.

5 If the α differ among time periods, the longitudinal estimator requires only that the values of α be identical in two successive time periods, whereas the repeated cross-section estimators require that it be identical in three successive time periods.

6 Longitudinal data aid in identifying $\text{Var}(\varepsilon_i \mid d_i = 1, \mathbf{Z}_i)$. With sufficient time series structure on U_{it} or the pre- and posttraining data on earnings, it is possible to use the squared least squares residuals to identify this parameter if, e.g., U_{it} and ε_i are independent, using standard components of variance models. See, e.g., Judge et al. (1980) for a description of such models.

7 In this case, the variance in ε_i can be estimated by standard variance components methods. See, e.g., Judge et al. (1980, chap. 8). Assuming normally distributed ε_i, τ_i, and U_{it}, this model is identical to that presented in Heckman's (1978) dummy endogenous variable model.

8 In a more general model agents might forecast ε_i using a richer information set. Denote the information set by $\mathcal{I}_i$ which may include lagged values of Y

$$E(\varepsilon_i \mid \mathcal{I}_i) = \Lambda(\mathcal{I}_i)$$

and the forecast error is

$$\rho_i = \varepsilon_i - \Lambda(\mathcal{I}_i)$$

where $\bar{\alpha} + \Lambda(\mathcal{I}_i)$ replaces $\bar{\alpha}$ everywhere in the text equations. This model does not have the same statistical structure as a random coefficients model under perfect certainty.

Earnings equation (1.14) becomes

$$Y_{it} = \beta_t + d_i\alpha + d_i\Lambda(\mathcal{I}_i) + (d_i\rho_i + U_{it}), \qquad t > k \tag{*}$$

where

$$E(\rho_i \mid d_i = 1, \mathcal{I}_i) = 0$$

Assuming ε_i is independent of U_{it}, and the functional form of $\Lambda(\mathcal{I}_i)$ is known equation (*) can be consistently estimated using, e.g., the methods in Heckman (1978).

9 This is the basis for a test between the two specifications of the assignment process conditional on error structure (2.8), which is itself testable using longitudinal data. However, the test is critically dependent on the assumption that the prospective trainee knows $\mathbf{X}_{ik}$ in period $k-1$ and that $\mathbf{X}_{ik}$ is not contained in the space spanned by $\mathbf{X}_{i,k-1}$ (so $\mathbf{X}_{i,k-1}$ cannot consist entirely of time-invariant variables that also appear in $\mathbf{X}_{ik}$).

10 For an analysis of normal multiple-selection rules see Catsiapsis and Robinson (1982) and Abowd and Farber (1982).

11 One example with multiple-selection rules involves no new analysis whatsoever. Suppose that IN_1 is $PV(1) - PV(0)$. Suppose also that administrators fill available openings by choosing at random among willing prospective participants. Now IN_2 is independent of X, V, and S. There is no loss of generality in assuming that IN_2 exists for individuals who do not present themselves for

training and that IN_2 is independent of IN_1. Then we can define

$$S_i^* = S_i \quad \text{if} \quad IN_{2i} > 0$$
$$S_i^* = -\infty \quad \text{if} \quad IN_{2i} < 0$$

Replacing S_i with S_i^* in equation (2.2), the previous analyses go through.

12 See Lee (1982) for some nonnormal models.

13 Note that χ need not necessarily be identified to identify α as long as (A-9e) is satisfied.

14 This is the basis for a test of the null hypothesis of no selection bias in the absence of knowledge of the functional form of the density function h. Provided (A-9e) is satisfied, it is possible to expand $E(U_{it}|d_i = 0, \mathbf{Z}_i)$ and $E(U_{it}|d_i = 1, \mathbf{Z}_i)$ in terms of polynomials in $\Pr(d_i = 0|\mathbf{Z}_i)$, exploiting the fact that $E(U_{it}|\mathbf{Z}_i = 0) = 0$. Under the null hypothesis of no selection bias, polynomials in $\Pr(d_i = 0|\mathbf{Z}_i)$ should not appear as statistically significant in a regression of Y_{it} on d_i, $\mathbf{X}_{it}$, and the polynomials. For details on this test see Heckman (1980).

Note further that if the null is rejected, it appears possible to use the results of Gallant (1981) to expand $E(U_{it}|d_i, \mathbf{Z}_i)$ in a Fourier expansion in terms of estimated $\Pr(d_i = 0|\mathbf{Z}_i)$ and to estimate the parameters of (3.4) nonparametrically.

15 Note that even though α is identified, ω need not be uniquely identified.

16 In independent work, Heckman and Neumann (1977) estimate a more general random coefficients model of this type with $\boldsymbol{\beta}$ random in addition to α_i. The only essential difference between the Lee (1978) model and the Heckman dummy endogenous variable model is that Heckman requires that $E(\varepsilon_i|d_i = 1) = 0$, whereas Lee does not. As noted in Section 2.3, in some environments of decision making under uncertainty it is plausible that $E(\varepsilon_i|d_i = 1) = 0$.

17 Note that in a simple random sample $E(\mathbf{X}_{it}'U_{it}|d_i)$ and the other conditional expectations are common for all i and the expressions in the text simplify to

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

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

18 If $g(Z_i^e)$ and $\mathbf{X}_{it}$ are not independent of $\mathbf{X}_{ik}$ and U_{it} is serially dependent, the terms inside the braces are not zero when perfect-foresight rule (2.4) characterizes the enrollment decision.

19 In a random sample these expressions simplify to

$$\operatorname*{plim}_{I_t \to \infty} \frac{\sum \mathbf{X}_{it}'d_i}{I_t} = pE(\mathbf{X}_{it}|d_i = 1)$$

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

20 *Proof:* In the reweighted choice-based sample,

$$E(d_i\phi_{it}|\mathbf{X}_{it}, \mathbf{Z}_i) = 1 - F(-\mathbf{Z}_i\gamma)$$

and

$$E(\phi_{it}U_{it}|\mathbf{X}_{it}, \mathbf{Z}_i) = E(U_{it}|d_i = 1, \mathbf{X}_{it}, \mathbf{Z}_i)\left(\frac{\Pr(d_i = 1|\mathbf{Z}_i)}{p^*(d_i = 0|\mathbf{Z}_i)}\right)p^*(d_i = 1|\mathbf{Z}_i)$$
$$+ E(U_{it}|d_i = 0, \mathbf{X}_{it}, \mathbf{Z}_i)\left(\frac{\Pr(d_i = 0|\mathbf{Z}_i)}{p^*(d_i = 0|\mathbf{Z}_i)}\right)p^*(d_i = 0|\mathbf{Z}_i)$$
$$= 0$$

by virtue of (A-8d).

21 The asymptotic theory is considerably simplified if the frequency is a member of a finite parameter family.

22 This assumption is not needed if p_t varies with t. See the discussion in Section 1.3.2.

23 We note parenthetically that in a random coefficient earnings model with regressors in the enrollment rule and with (5.1) as the error term in the earnings equation, the first-difference method does not consistently estimate $\bar{\alpha}$. Thus the Lee (1978) model of unionism is fundamentally different from the model implicit in Chamberlain's work (1982), although this difference has not been noticed by labor economists (see, e.g., Lewis, 1982). Heckman and Neumann (1977) and Lee (1978) estimate a different coefficient than does Chamberlain. Hence, comparisons of the impacts of union status from the two procedures are meaningless.

24 The requirement that $|\rho| < 1$ is needed only to guarantee the validity of the forecasting rules used in Section 2 for infinite horizon problems.

25 Notice that even if S_i and $\mathbf{X}_{it}$ are distributed independently of U_{it} for all t, (A-15b) does not imply error structure (5.1). (A-15b) only implies that

$$E(U_{it}|d_i) = E(U_{it'}|d_i) = l(d_i)$$

so we may uniquely write $U_{it} = \psi_i + \varepsilon_{it}$, where $E(\psi_i|d_i) = l(d_i)$ and $E(\varepsilon_{it}|d_i) = 0$. Here ψ_i need not be independent of ε_{it}.

26 Thus writing (1.1) as $Y_{it} = \beta_t + \mathbf{X}_{it}\boldsymbol{\pi} + d_i\alpha + U_{it}$, it is possible to estimate $\boldsymbol{\pi}$ from sample (iii) preprogram data. (This assumes there are no time-invariant variables in $\mathbf{X}_{it}$. If there are such variables, they may be deleted from the regressor vector and $\boldsymbol{\pi}$ appropriately redefined without affecting the analysis.) Then from the mean of sample (i) it is possible to consistently estimate $\beta_t - \beta_{t'} + (\bar{\mathbf{X}}_t^{(i)} - \bar{\mathbf{X}}_{t'}^{(i)})\boldsymbol{\pi} + \alpha$, where $\bar{\mathbf{X}}_t^{(i)}$ is the population mean of $\mathbf{X}_{it}$ in period t for sample (i), and from the mean of sample (iii) differences it is possible to consistently estimate $\beta_t - \beta_{t'} + (\bar{\mathbf{X}}_{it} - \bar{\mathbf{X}}_{it'})\boldsymbol{\pi} + \alpha p$. Since $\boldsymbol{\pi}$ is known, the means can be adjusted for the effect of $\mathbf{X}$. The adjusted means can then be used in the procedure described in the text. Note that we are assuming that no $\mathbf{X}_{it}$ variables become nonconstant after period k.

27 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). Kagan, Linnik, and Rao (1973, p. 10) give necessary and sufficient conditions for linearity of the regression. The multivariate t density is just one example of a family of densities with linear regressions. There are many more.

28 From (iii), the joint density of U_{it}, U_{ik}, $\mathbf{X}_{ik}$, S_i may be written as $f(U_{it}, U_{ik})f(\mathbf{X}_{ik}, S_i)$. Substitute for U_{it} using $U_{it} = \delta U_{ik} + \omega_{it}$. Then $E(\omega_{it}|U_{ik}) =$

0 means that

$$\int \omega_{it} f(\delta U_{ik} + \omega_{it}, U_{ik})\, d\omega_{it} = 0. \qquad (*)$$

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

$$E(\omega_{it}|d_i = 1) = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{-\mathbf{X}_{ik}\boldsymbol{\beta}+S_i+(\alpha/r)}\int_{-\infty}^{\infty} \omega_{it} f(\delta U_{ik} + \omega_{it}, U_{ik}) f(\mathbf{X}_{ik}, S_i)\, d\omega_{it}\, dU_{ik}\, d\mathbf{X}_{ik}\, dS_i}{\Pr(d_i = 1)}$$

and the firstfold integral (from within) is identically zero because of (*).

29 *Proof:* Using (2.7) and (i) above, define $R(\mathbf{X}_{ik}, U_{i,k-1}, \ldots, U_{i,k-N}, S_i)$ so that

$$R(\mathbf{X}_{ik}, U_{i,k-1}, \ldots, U_{i,k-N}, S_i)$$

$$= 1 \quad \text{if} \quad \mathbf{X}_{ik}\boldsymbol{\beta} + E(U_{ik}|U_{i,k-1}, \ldots, U_{i,k-N}) > S_i + \frac{\alpha}{r}$$

$$= 0 \qquad \text{otherwise}$$

From (i), $R(\mathbf{X}_{ik}, U_{i,k-1}, \ldots, U_{i,k-N}, S_i) = d_i$. Thus

$$E(d_i|U_{it}, \ldots, U_{i,k-N})$$

$$= \underset{(S_i, \mathbf{X}_{ik})}{E}\left[R(\mathbf{X}_{ik}, U_{i,k-1}, \ldots, U_{i,k-N}, S_i)\middle|U_{it}, \ldots, U_{i,k-1}, \ldots, U_{i,k-N}\right]$$

$$= \underset{(S_i, \mathbf{X}_{ik})}{E}\left[R(\mathbf{X}_{ik}, U_{i,k-1}, \ldots, U_{i,k-N}, S_i)\middle|U_{i,k-1}, \ldots, U_{i,k-N}\right]$$

as a consequence of assumptions (v) and (vi). Similarly,

$$E(d_i|U_{i,k-1}, \ldots, U_{i,k-N}) = \underset{(S_i, \mathbf{X}_{ik})}{E}\left[R(\mathbf{X}_{ik}, U_{i,k-1}, \ldots, U_{i,k-N}, S_i)\middle|U_{i,k-1}, \ldots, U_{i,k-N}\right]$$

30 This is true for most analyses based on the Panel Survey of Income Dynamics, which began with observations on 5000 families. A typical empirical project based on this data set is for a single demographic group.

REFERENCES

Abowd, J., and Farber, H. (1982). "Jobs Queues and the Union Status of Workers." *Industrial and Labor Relations Review* 35, 354–67.

Aigner, D. (1973). "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1, 49–60.

Amemiya, T. (1981). "Qualitative Response Models: A Survey." *Journal of Economic Literature* 19, 1483–1536.

(1983). "A Comparison of the Amemiya GLS and the Lee–Maddala–Trost G2SLS in a Simultaneous Equations Tobit Model." *Journal of Econometrics* 23, 295–300.

Ashenfelter, O. (1978). "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60, 47–57.

Barnow, B. (1983). Personal discussions.

Barnow, B., Cain, G., and Goldberger, A. (1980). "Issues in the Analysis of Selectivity Bias." In *Evaluation Studies*, vol. 5, edited by E. Stromsdorfer and G. Farkas. San Franciso: Sage.

Bassi, L. (1983). "Estimating the Effect of Training Programs with Nonrandom Selection." Ph.D. dissertation, Princeton University.

Björklund, A., and Moffitt, R. (1983) "Estimation of Wage Gains and Welfare Gains from Self-Selection Models." Institute for Research on Poverty, University of Wisconsin.

Catsiapsis, B., and Robinson, C. (1982). "Sample Selection Bias with Multiple Selection Rules: An Application to Student Aid Grants." *Journal of Econometrics* 18, 351–68.

Chamberlain, G. (1977). "An Instrumental Variables Interpretation of Identification in Variance Components and MIMIC Models." In *Kinometrics: The Determinants of Socio-Economic Success within and between Families*, edited by P. Taubman. Amsterdam: North-Holland.

(1982). "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18, 1–46.

(1984). "The Analysis of Panel Data." In *Handbook of Econometrics*, vol. II, edited by Z. Griliches and M. Intriligator. Amsterdam: North-Holland.

Cochran, W. (1968). "Errors in Measurement in Statistics." *Technometrics* 10, 637–66.

Cosslett, S. (1981). "Maximum Likelihood Estimation for Choice-Based Samples." *Econometrica* 49, 1289–316.

(1983). "Distribution-Free Maximum Likelihood Estimators of the Binary Choice Model." *Econometrica* 51, 765–872.

Duncan, G. J., Juster, T., and Morgan, J. (1982). "The Role of Panel Studies in a World of Scarce Research Resources." Paper presented at SSRC Conference on Designing Research with Scarce Resources, Washington, D.C.

Durbin, J. (1954). "Errors in Variables." *Review of International Statistics Institute* 22, 23–32.

Gallant, A. R. (1981). "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form." *Journal of Econometrics* 15, 211–45.

Hausman, J. (1978). "Specification Tests in Econometrics." *Econometrica* 46, 1251–71.

Heckman, J. (1976). "Simultaneous Equations Models with Continuous and Discrete Endogenous Variables and Structural Shifts." In *Studies in Nonlinear Estimation*, edited by S. Goldfeld and R. Quandt. Cambridge, Mass.: Ballinger.

(1978). "Dummy Endogenous Variables in a Simultaneous Equations Systems." *Econometrica* 46, 931–61.

(1979). "Sample Selection Bias as a Specification Error." *Econometrica* 47, 153–61.

(1980). "Addendum to Sample Selection Bias as a Specification Error." In *Evaluation Studies*, vol. 5, edited by E. Stromsdorfer and G. Farkas. San Francisco: Sage.

Heckman, J., and Neumann, G. (1977). "Union Wage Differentials and the Decision to Join Unions." Typescript, University of Chicago.

Heckman, J., and Wolpin, K. (1976). "Does the Contract Compliance Program Work? An Analysis of Chicago Data." *Industrial and Labor Relations Review* 29(4), 554–64.
Judge, G., Griffiths, W., Hill, R., and Lee, T. (1980). *The Theory and Practice of Econometrics*. New York: Wiley.
Kagan, A., Linnik, T., and Rao, C. (1973). *Some Characterization Theorems in Mathematical Statistics*. New York: Wiley.
Lee, L. F. (1978). "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables." *International Economic Review* 19, 415–33.
(1982). "Some Approaches to the Correction of Selectivity Bias." *Review of Economic Studies* 49, 355–72.
Lewis, H. G. (1982). "Union Relative Wage Effects: A Survey." Typescript, Duke University.
Lillard, L., and Weiss, Y. (1979). "Components of Variation in Panel Earnings Data: American Scientists, 1960–70." *Econometrica* 47, 437–54.
MaCurdy, T. (1982). "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis." *Journal of Econometrics* 18(1), 83–114.
Madansky, A. (1964). "Instrumental Variables in Factor Analysis." *Psychometrika* 29, 105–18.
Manski, C., and Lerman, S. (1977). "The Estimation of Choice Probabilities from Choice-Based Samples." *Econometrica* 45, 1977–88.
Manski, C., and McFadden, D. (1981). "Alternative Estimators and Sample Designs for Discrete Choice Analysis." In *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. Manski and D. McFadden. Cambridge: MIT Press.
Mincer, J., and Jovanovic, B. (1981). "Labor Mobility and Wages." In *Studies in the Labor Market*, edited by S. Rosen. Chicago: University of Chicago Press.
Mundlak, Y. (1961). "Empirical Production Function Free of Management Bias." *Journal of Farm Economics* 43, 45–56.
(1978). "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46, 69–85.
Pudney, S. E. (1982). "Estimating Latent Variable Systems When Specification Is Uncertain: Generalized Component Analysis and the Eliminant Method." *Journal of the American Statistical Association* 77, 883–9.
Roy, A. (1951). "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3, 135–46.
Sargent, T. (1979). *Macroeconomic Theory*. New York: Academic Press.
Sisam, C. (1940). *College Algebra*. New York: Holt and Day.
White, H. (1980). "Nonlinear Regression on Cross Section Data." *Econometrica* 48, 721–46.
(1982). "Instrumental Variables Regressions with Independent Observations." *Econometrica* 50, 483–99.
(1984). *Asymptotic Theory for Econometricians*. Orlando, Fla.: Academic Press.
Willis, R., and Rosen, S. (1979). "Education and Self Selection." *Journal of Political Economy*, 87, S7–S36.
Wu, D. M. (1973). "Alternative Tests of Independence between Stochastic Regressors and Disturbances." *Econometrica* 41, 733–50.

(1983). "Tests of Causality, Predeterminedness, and Exogeneity." *International Economic Review* 24, 547–58.
Zellner, A., Kmenta, J., and Dreze, J. (1966). "Specification and Estimation of Cobb–Douglas Production Function Models." *Econometrica* 34, 784–95.