Alternative Methods For Evaluating the Impact of Interventions: An Overview

Excerpt from the *Journal of Econometrics*, 1985

James J. Heckman & Richard Robb Jr.

Econ 312, Spring 2021

2. Notation and a model of program participation

2.1. Earnings functions

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

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

- Adopting a linear specification, latent earnings as

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

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

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

$$Y_{it} = \begin{cases} X_{it}\beta + d_i\alpha + U_{it} & t > k \\ X_{it}\beta + U_{it} & t \leq k \end{cases} \qquad (1)$$

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

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

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

- Observed earnings may be written as

$$\begin{aligned} Y_{it} &= X_{it}\beta + d_i\alpha + U_{it} & t > k \\ Y_{it} &= X_{it}\beta + U_{it} & t \leq k, \end{aligned} \tag{2}$$

where β and α are parameters.

- Because of the covariance between d_i and U_{it},

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

- Equation (2) assumes that training has the same effect on everyone.
- We can also develop the analysis when α varies among individuals, as is assumed in many analyses of experimental and nonexperimental data (see Fisher, 1953).
- Throughout, we largely ignore effects of training which grow or decay over time.

2.2. Enrollment rules

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

$$IN_i = Z_i\gamma + V_i. \tag{3}$$

- In terms of this function,

  $$d_i = \begin{cases} 1 & \text{iff } IN_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

  The distribution function of V_i is denoted as $F(v_i) = \Pr(V_i < v_i)$.
- V_i is assumed to be independently and identically distributed across persons.
- Let $p = E[d_i] = \Pr[d_i = 1]$ and assume $1 > p > 0$.

- Assuming that V_i is distributed independently of Z_i (a requirement not needed for most of the estimators considered in this paper), we may write $\Pr(d_i = 1 \mid Z_i) = F(-Z_i\gamma)$, which is sometimes called the "propensity score" in statistics (see, e.g., Rosenbaum and Rubin, 1983).
- We show later a special subclass of econometric selection-correction estimators can be expressed as functions of the propensity score.

- The condition for the existence of selection bias

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

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

A Behavioral Model

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

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

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

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

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

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

- The risk-neutral wealth-maximizing decision rule is to enroll in the program if $PV_i(I) > PV_i(0)$ or, letting IN_i denote the index function in decision rule (3),

$$IN_i = PV_i(1) - PV_i(0) = E_{k-1}[S_i - Y_{ik} + \alpha/r], \tag{4}$$

so the decision to train is characterized by the rule

$$d_i = \begin{cases} 1 & \text{iff } E_{k-1}[S_i - Y_{ik} + \alpha/r] > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

- Let W_i be the determinant of the subsidy that the econometrician observes (with associated coefficient ϕ) and let τ_i be the part which he does not observe:

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

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

- Collecting terms,

$$d_i = \begin{cases} 1 & \text{iff } S_i - Y_{ik} + \alpha/r = W_i\phi + \alpha/r - X_{ik}\beta + \tau_i - U_{ik} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

- Then $(\tau_i - U_{ik}) = V_i$ in (3) and (W_i, X_{ik}) corresponds to Z_i in (3).
- Assuming that (W_i, X_{ik}) is distributed independently of V_i makes (6) a standard discrete choice model.
- This assumption is only required for some of the estimators discussed here.

- Suppose decision rule (6) determines enrollment.
- If the costs of program participation are independent of U_{it} for all t (so both W_i and τ_i are independent of U_{it}), then $E[U_{it}d_i] = 0$ only if the unobservables in period t are (mean) independent of the unobservables in period k or

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

Question: Prove this.

- Whether or not U_{it} and d_i are uncorrelated hinges on the serial dependence properties of U_{it}.

- If U_{it} is a moving average of order m so

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

where the $\varepsilon_{i,t-j}$ are iid, then for $t - k > m$, $E[U_{it} d_i] = 0$.

- On the other hand, if U_{it} follows a first-order autoregressive scheme, then $E[U_{it} \mid U_{ik}] \neq 0$ for all finite t and k.

- The enrollment decision rules derived in this subsection give context to the selection bias problem.
- The estimators discussed in this paper differ greatly in their dependence on particular features of these rules.
- Some estimators do not require that these decision rules be specified at all, while other estimators require a great deal of *a priori* specification of these rules.
- Given the inevitable controversy that surrounds specification of enrollment rules, there is always likely to be a preference by analysts for estimators that require little prior knowledge about the decision rule.
- But this often throws away valuable information and ignores the subjective evaluation implicit in $d_i = 1$.

Link to Section 3. Appendix

4. Cross-sectional procedures

- Standard cross-sectional procedures invoke unnecessarily strong assumptions.
- All that is required to identify α in a cross-section is access to a regressor in (3).
- In the absence of a regressor, assumptions about the marginal distribution of U_{it}, can produce consistent estimators of the training impact.

4.1. Without distributional assumptions a regressor is needed

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

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

  for $0 < \sum d_i < I$, where I is the number of observations.

- We retain the assumption that the data are generated by a random sampling scheme.

- If no regressors appear in (1) then $X_{it}\beta = \beta_t$, and

$$\begin{aligned} \text{plim } \bar{Y}_t^{(1)} &= \beta_t + \alpha + E[U_{it} \mid d_i = 1], \\ \text{plim } \bar{Y}_t^{(0)} &= \beta_t + E[U_{it} \mid d_i = 0]. \end{aligned}$$

- Thus

$$\text{plim}\left(\bar{Y}_t^{(1)} - \bar{Y}_t^{(0)}\right) = \alpha + E[U_{it} \mid d_i = 1]/(1-p),$$

since $pE[U_{it} \mid d_i = 1] + (1-p)E[U_{it} \mid d_i = 0] = 0.$

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

4.2. Overview of cross-sectional procedures which use regressors

- If, however, $E[U_{it} \mid d_i = 1, Z_i]$ is a non-constant function of Z_i, it is possible (with additional assumptions) to solve this identification problem.
- Securing identification in this fashion explicitly precludes a fully non-parametric strategy in which both the earnings function (1) and decision rule (3) are estimated in each (X_{it}, Z_i) stratum.
- For within each stratum, $E[U_{it} \mid d_i = 1, Z_i]$ is a constant function of Z_i and α is not identified from cross-section data.
- Restrictions across strata are required.

- If $E[U_{it} \mid d_i = 1, Z_i]$ is a non-constant function of Z_i it is possible to exploit this information in a variety of ways depending on what else is assumed about the model.
- Here we simply sketch alternative strategies.

a. Suppose Z_i or a subset of Z_i is exogenous with respect to U_{it}. Under conditions specified more fully below, the exogenous subset may be used to construct an instrumental variable for d_i in eq. (1), and α can be consistently estimated by instrumental variables methods. No distributional assumptions about U_{it} or V_i are required [Heckman (1978)].

b. Suppose that Z_i, is distributed independently of V_i, and the functional form of the distribution of V_i, is known, or can be consistently estimated. Under standard conditions, γ in (3) can be consistently estimated by conventional methods in discrete choice analysis. If Z_i, is distributed independently of U_{it}, $F(-Z_i\hat{\gamma})$ can be used as an instrument for d_i, in eq. (1) [Heckman (1978)].

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

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

γ and α can be consistently estimated using $F(-Z_i\hat{\gamma})$ in place of $F(-Z_i\gamma)$ in the preceding equation [Heckman (1976,1978)] or else the preceding equation can be estimated by non-linear least squares, estimating β, α and γ jointly (given the functional form of F).

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

$$\begin{aligned} E[Y_{it} \mid d_i, Z_i] = & X_{it}\beta + d_i\alpha + d_i E[U_{it} \mid d_i = 1, Z_i] \\ & + (1 - d_i)E[U_{it} \mid d_i = 0, Z_i]. \end{aligned} \tag{7}$$

One way to acquire information about the functional form of $E[U_{it} \mid d_i = 1, Z_i]$ is to assume knowledge of the functional form of the joint distribution of (U_{it}, V_i) (e.g., that it is bivariate normal), but this is not required. Note further that this procedure does not require that Z_i, be distributed independently of V_i in (3) [Barnow, Cain and Goldberger (1980)].

e. Instead of (d), it is possible to use a two-stage estimation procedure if the joint density of (U_{it}, V_i) is assumed known up to a finite set of parameters. In stage one $E[U_{it} \mid d_i = 1, Z_i]$ and $E[U_{it} \mid d_i = 0, Z_i]$ are determined up to some unknown parameters by conventional discrete choice analysis. Then regression (7) is run using estimated E values in place of population E values on the right-hand side of the equation.

f. Under the assumptions of (e), use maximum likelihood to consistently estimate α ([Heckman (1978)]. Note that a separate value of α may be estimated for each cross-section so that depending on the number of crosssections it is possible to estimate growth and decay effects in training (e.g., α_t can be estimated for each cross-section).

- Conventional selection bias approaches (d)-(f) as well as (b)-(c) rely on strong distributional assumptions but in fact these are not required.
- Given that a regressor appears in decision rule (3), if it is uncorrelated with U_{it}, the regressor is an instrumental variable for d_i.
- It is not necessary to invoke strong distributional assumptions, but if they are invoked, Z_i need not be uncorrelated with U_{it}.
- In practice, however, Z_i and U_{it} are usually assumed to be independent.
- We next discuss the instrumental variables procedure in greater detail.

4.3. The instrumental variable estimator

- This estimator is the least demanding in the a priori conditions that must be satisfied for its use.

- It requires the following assumptions:

There is at least one variable in Z_i, Z_i^e, with a non-zero γ coefficient in (3), such that for some known transformation of Z_i^e, $g(Z_i^e)$, $E[U_{it} g(Z_i^e)] = 0$.

(8a)

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

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

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

(8b)

- With these assumptions, the IV estimator,

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

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

- It is important to notice how weak these conditions are.
- The functional form of the distribution of V_i need not be known.
- Z_i need not be distributed independently of V_i.
- Moreover, $g(Z_i^e)$ may be a non-linear function of variabies appearing in X_{it} as long as (8) is satisfied.

- The instrumental variable, $g(Z_i^e)$ may also be a lagged value of time-varying variables appearing in X_{it} provided the analyst has access to longitudinal data.
- The rank condition (8b) will generally be satisfied in this case as long as X_{it} exhibits serial dependence.
- Thus longitudinal data (on exogenous characteristics) may provide a source of instrumental variables.

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

- If no regressor appears in decision rule (3) the estimators presented so far in this section cannot be used to estimate α consistently unless additional restrictions are imposed.
- Heckman (1978) demonstrates that if (U_{it}, V_i) are jointly normally distributed, α is identified even if there is no regressor in enrollment rule (3).
- His conditions are overly strong.

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

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

$$E[U_{it}^3] = 0. \tag{9a}$$

$$E[U_{it}^5] = 0. \tag{9b}$$

$$\{U_{it}, V_i\} \text{ is iid.} \tag{9c}$$

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

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

$$(1/I_t)\sum_{i=1}^{I_t}[(Y_{it} - \bar{Y}) - \hat{\alpha}(d_i - \bar{d})]^3 \tag{10a}$$

and

$$(1/I_t)\sum_{i=1}^{I_t}[(Y_{it} - \bar{Y}) - \hat{\alpha}(d_i - \bar{d})]^5 \tag{10b}$$

as close to zero as possible in a suitably chosen metric where, as before, the overbar denotes sample mean.

- In our earlier paper, we establish the existence of a unique consistent root that sets (10a) and (10b) to zero in large samples.

4.5. Selection on Observables

- In the special case in which

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

selection is said to occur on the observables.

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

- In the notation of Dawid (1979) as used by Rosenbaum and Rubin (1983),

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

  i.e., given Z_i, d_i is strongly ignorable.

- In a random coefficient model the required condition is

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

- The strategy for consistent estimation presented in 4.2 must be modified; in particular, methods (a)-(c) are inappropriate.
- However, method (d) still applies and simplifies because

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

  so that we obtain in place of equation (8)

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

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

- Method (e) of Section 4.2 no longer directly applies.
- Except in unusual circumstances (e.g., a single element of Z_i), there is no relationship between any of the parameters of $E(U_{it} \mid Z_i)$ and the propensity score $\Pr(d_i = 1 \mid Z_i)$, so that conventional two-stage estimators generated from discrete choice theory do not produce useful information. Method (f) produces a consistent estimator provided that an explicit probabilistic relationship between U_{it} and Z_i is postulated.

4.6. Summary

- Conventional cross-section practice invokes numerous extraneous assumptions to secure identification of α.
- These overidentifying restrictions are rarely tested, although they are testable.
- Strong distributional assumptions are not required to estimate α.

- Assumptions about the distributions of unobservables are rarely justified by an appeal to behavioral theory.
- Assumptions about the presence of regressors in enrollment equations and assumptions about stochastic dependence relationships among U_{it}, Z_i, and d_i are sometimes justified by behavioral theory.

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

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

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

- To see why this claim is true, suppose that no regressors appear in the earnings function.
- (Comment: If regressors appear in the earnings function, the following procedure can be used. Rewrite (1) as $Y_{it} = \beta_t + X_{it}\pi + d_i\alpha + U_{it}$. It is possible to estimate π from pre-program data. Replace Y_{it} by $Y_{it} - X_{it}\hat{\pi}$ and the analysis in the text goes through. Note that we are assuming that no X_{it} variables become non-constant after period k.)

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

$$\begin{aligned} \text{plim}\, \overline{Y}_t &= \text{plim} \sum Y_{it}/I_t \\ &= E\left[\beta_t + \alpha d_i + U_{it}\right] = \beta_t + \alpha p, t > k \\ \text{plim}\, \bar{Y}_{t'} &= \text{plim} \sum Y_{it'}/I_{t'} \\ &= E\left[\beta_{t'} + U_{it'}\right] = \beta_{t'}, t' < k. \end{aligned}$$

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

$$\text{plim}\left(\overline{Y}_t - \overline{Y}_{t'}\right)/\hat{p} = \alpha,$$

where $\hat{p}$ is a consistent estimator of $p = E\left[d_i\right]$.

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

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

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

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

$$\begin{aligned} \beta_t &= \pi_0 + \pi_1 t && \text{for } t > k, \\ \alpha_t &= \phi_0(\phi_1)^{t-k} && \text{for } t > k. \end{aligned}$$

- In this case, π_0, π_1, ϕ_0, ϕ_1 are identified from the means of four cross-sections, so long as at least one of these means comes from a pre-program period.

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

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

6.1. The fixed effects method

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

$$E\left[U_{it} - U_{it'} \mid d_i, X_{it} - X_{it'}\right] = 0 \qquad \text{for all } t, t', \qquad t > k > t'. \tag{11}$$

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

$$E\left[Y_{it} - Y_{it'} \mid d_i, X_{it} - X_{it'}\right] = \left(X_{it} - X_{it'}\right)\beta + d_i\alpha, \qquad t > k > t'.$$

- Suppose that (11) holds and the analyst has access to one year of preprogram and one year of post-program earnings.
- Regressing the difference between post-program earnings in any year and earnings in any pre-program year on the change in regressors between those years and a dummy variable for training status produces a consistent estimator of α.

- Some decision rules and error processes for earnings produce (11).
- For example, consider a certainty environment in which the earnings residual has a permanent-transitory structure:

$$U_{it} = \phi_i + \varepsilon_{it}, \tag{12}$$

  where ε_{it} is a mean zero random variable independent of all other values of ε_{it}, and is distributed independently of ϕ_i, a mean zero person-specific time-invariant random variable.
- Assuming that S_i, in decision rule (6) is distributed independently of all ε_{it} except possibly for ε_{ik}, then (11) will be satisfied.
- With two periods of data (in t and t', $t > k > t'$) α is just identified. With more periods of panel data, the model is overidentified and hence condition (12) is subject to test.

- Eq. (11) may also be satisfied in an environment of uncertainty.
- Suppose eq. (12) governs the error structure in (1) and

$$E_{k-1}\left[\varepsilon_{ik}\right] = 0,$$

and

$$E_{k-1}\left[\phi_i\right] = \phi_i,$$

- Agents cannot forecast innovations in their earnings, but they know their own permanent component.
- Provided that S_i, is distributed independently of all ε_{it}, except possible for ε_{ik}, this model also produces (11).

- We investigate the plausibility of (11) with respect to more general decision rules and error processes in section 8.

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

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

$$U_{it} = \rho U_{i,t-1} + \nu_{it}, \tag{13}$$

where $E[\nu_{it}] = 0$ and the ν_{it} are mutually independently (not necessarily identically) distributed random variables with $p \neq 1$.

- Substitution using (1) and (13) to solve for $U_{it'}$ yields

$$\begin{aligned} Y_{it} = {} & \left[X_{it} - X_{it'}\rho^{t-t'}\right]\beta + \left(1 - \rho^{t-t'}\right)d_i\alpha + \rho^{t-t'}Y_{it'} \\ & + \left\{\sum_{j=0}^{t-(t'^{+1})} \rho^j \nu_{i,t-j}\right\}, t > t' > k. \end{aligned} \tag{14}$$

- Assume further that the perfect foresight rule (6) determines enrollment, and the ν_{ij} are distributed independently of S_i and X_{ik} in (6).

- As a consequence of these assumptions,

$$E\left[Y_{it} \mid X_{it}, X_{it'}, d_i, Y_{it'}\right] = \left(X_{it} - X_{it'}\rho^{t-t'}\right)\beta + \left(1 - \rho^{t-t'}\right)d_i\alpha + \rho^{t-t'}Y_{it'}, \tag{15}$$

  so that (linear or non-linear) least squares applied to (15) consistently estimates α as the number of observations becomes large.

- (The appropriate non-linear regression increases efficiency by imposing the cross-coefficient restrictions.)

- As is the case with the fixed effect estimator, increasing the length of the panel and keeping the same assumptions, the model becomes overidentified (and hence testable) for panels with more than two observations.

6.3. U_{it} is covariance-stationary

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

- Assume U_{it} is covariance stationary:

$$E\left[U_{it}U_{i,t-j}\right] = E\left[U_{it'}U_{i,t'-j}\right] = \sigma_j \text{ for } j \geq 0 \text{ for all } t, t', \quad (16a)$$

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

(16b)

$$pE\left[U_{it'} \mid d_i = 1\right] \neq 0.$$

(16c)

- We make no assumptions here about the appropriate enrollment rule or about the stochastic relationship between U_{it} and the cost of enrollment S_i.

- Let

$$
\begin{aligned}
Y_{it} &= \beta_t + d_i\alpha + U_{it}, && t > k, \\
Y_{it'} &= \beta_{t'} + U_{it'}, && t' < k,
\end{aligned}
$$

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

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

$$m_1 = \left(\sum \left(Y_{it'} - \overline{Y}_{t'} \right) \left(Y_{i,t'-j} - \overline{Y}_{t'-j} \right) \right) / I, \qquad \operatorname{plim} m_1 = \sigma_j.$$

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

$$m_2 = \left(\sum \left(Y_{it} - \overline{Y}_{t} \right) \left(Y_{i,t'} - \overline{Y}_{t'} \right) \right) / I,$$

$$\operatorname{plim} m_2 = \sigma_j + \alpha p E \left[U_{it'} \mid d_i = 1 \right], \qquad t > k > t'.$$

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

$$m_3 = \left(\sum \left(Y_{it'} - \overline{Y}_{t'} \right) d_i \right) / I,$$

$$\operatorname{plim} m_3 = pE \left[U_{it'} \mid d_i = 1 \right], \qquad t' < k.$$

- Combining this information and assuming $pE \left[U_{it'} \mid d_i \neq 0 \right]$ for $t' < k$,

$$\operatorname{plim} \hat{\alpha} = \operatorname{plim} \left(\left(m_2 - m_1 \right) / m_3 \right) = \alpha.$$

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

6.4 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

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

i.e. that predictions of future U_{it} are made solely on the basis of previous values of U_{it}.

- Past values of the exogenous variables are assumed to have no predictive value for U_{ik}.

- Assume further

  b the relevant earnings history goes back N periods before period k;

  c the enrollment decision is characterized by equation (4);

  d S_i and X_{ik} are known as of period $k-1$ when the enrollment decision is being made;

  e X_{it} is distributed independently of U_{ij} for all t and j; and

  f S_i is distributed independently of U_{ij} for all j.

- Defining

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

and

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

- Under these conditions α can be consistently estimated.
- Define

$$p = E(d_i),$$

and

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

- Rewrite (2) in the following way:

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

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

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

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

- Thus, it remains to show that

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

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

$$E\left(d_i \mid U_{it}, U_{i,t-1}, \ldots, U_{i,k-1}, \ldots, U_{i,k-N}\right)=E(d_i \mid U_{i,k-1}, \ldots, U_{i,k-N}) \tag{18}$$

Question: Prove this.

- This relationship is proved in our companion paper.
- Since only preprogram innovations determine participation and because U_{it} is distributed independently of X_{ik} and S_i in the decision rule of equation (4), the conditional mean of d_i does not depend on postprogram values of U_{it} given all preprogram values.
- Intuitively, the term $U_{it} - c(G(\psi_i) - p)$ is orthogonal to $G(\psi_i)$, the best predictor of d_i based on ψ_i; if $U_{it} - c(G(\psi_i) - p)$ were correlated with d_i, it would mean that U_{it} helped to predict d_i, contradicting condition (18).

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

$$
\begin{aligned}
E\left\{d_i\left[U_{it} - c(G(\psi_i) - p)\right]\right\} &= E\left[E\left\{d_i\left[U_{it} - c(G(\psi_i) - p)\right]\right\} \mid \psi \\
&= E\left\{\left[U_{it} - c(G(\psi_i) - p)\right] E(d_i \mid \psi_i \\
&= E\left\{\left[U_{it} - c(G(\psi_i) - p)\right] G(\psi_i)\right\} \\
&= 0
\end{aligned}
$$

as a consequence of the definition of c.

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

6. Repeated cross-section analogues of longitudinal procedures

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

- The previous section presented longitudinal estimators of α.
- In each case, however, α can actually be identified with repeated cross-section data.
- Here we establish this claim.

6.1. The fixed effect model

- As in section 5.1, assume that (12) holds so

$$E[U_{it}|d_i = 1] = E[U'_{it}|d_i = 1], E[U_{it}|d_i = 0] = E[U'_{it}|d_i = 0],$$

for all $t > k > t'$. Let $X_{it}\beta = \beta_t$ and define, in terms of the notation of section 3.1,

$$\hat{\alpha} = [\bar{Y}_t^{(1)} - \bar{Y}_t^{(0)}] - [\bar{Y}_{t'}^{(1)} - \bar{Y}_{t'}^{(0)}].$$

- Assuming random sampling, consistency of $\hat{\alpha}$ follows immediately from (11):

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

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

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

- The need for pre-program earnings to identify α is, however, only an artifact of the fixed effect assumption (12).
- Suppose instead that U_{it} follows a first-order autoregressive process given by (13) and that

  $$E[V_{it}|d_i] = 0, \qquad t > k, \tag{19}$$

  as in section 5.2.
- With three successive post-program cross-sections in which the identity of trainees is known, it is possible to identify α.

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

$$\begin{aligned} \text{plim } \bar{Y}_j^{(1)} &= \beta_j + \alpha + E[U_{ij} | d_i = 1], \\ \text{plim } \bar{Y}_j^{(0)} &= \beta_j + E[U_{ij} | d_i = 0], \end{aligned}$$

- From (19),

$$\begin{aligned} E[U_{i,t+1} | d_i = 1] &= \rho E[U_{it} | d_i = 1], \\ E[U_{i,t+1} | d_i = 0] &= \rho E[U_{it} | d_i = 0], \\ E[U_{i,t+2} | d_i = 1] &= \rho^2 E[U_{it} | d_i = 1], \\ E[U_{i,t+2} | d_i = 0] &= \rho^2 E[U_{it} | d_i = 0]. \end{aligned}$$

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

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

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

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

is consistent for α.

- For this model, the advantage of longitudinal data is clear.
- Only two time periods of longitudinal data are required to identify α, but three periods of repeated cross-section data are required to estimate the same parameter.
- However, if Y_{it} is subject to measurement error, the apparent advantages of longitudinal data become less clear.
- Repeated cross-section estimators are robust to mean zero measurement error in the variables.

- The longitudinal regression estimator discussed in section 6.2 does not identify α unless the analyst observes earnings without error.
- Given three years of longitudinal data and assuming that measurement error is serially uncorrelated, one could instrument (14) using earnings in the earliest year as an instrument.
- Thus one advantage of the longitudinal estimator disappears in the presence of measurement error.

6.3. Covariance stationarity

- For simplicity, suppress regressors in the earnings equation and let $X_{it}\beta = \beta_t$.
- Assume that conditions (16) are satisfied.
- Before presenting the repeated cross-section estimator, it is helpful to record the following facts:

$$var(Y_{it}) = \alpha^2(1-p)p + 2\alpha E[U_{it}|d_i = 1]p + \sigma_u^2, \;\; t > k, \tag{20a}$$

$$var(Y_{it}) = \sigma_u^2, \qquad t < k, \tag{20b}$$

$$cov(Y_{it}, d_i) = \alpha p(1-p) + pE[U_{it}|d_i = 1]. \tag{20c}$$

- Note that $E[U_{it}^2] = E[U_{it'}^2], t > k > t'$, by virtue of assumption (16a).
- Then

$$\hat{\alpha} = (p(1-p))^{-1} \left(\frac{\sum (Y_{it} - \bar{Y}_t) d_i}{I_t} \right. \tag{21}$$
$$\left. - \sqrt{\left(\frac{\sum (Y_{it} - \bar{Y}_t) d_i}{I_t} \right)^2 - p(1-p) \left(\frac{\sum (Y_{it} - \bar{Y}_t)^2}{I_t} - \frac{\sum (Y_{it'} - \bar{Y}_{t'})^2}{I_{t'}} \right)} \right)$$

is consistent for α.

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

- Notice that the estimators of sections 5.3 and 6.3 exploit different features of the covariance stationarity assumptions.
- The longitudinal procedure only requires that $E[U_{it}U_{i,t-j}] = E[U_{it'}U_{it'-j}]$ for $j > 0$; variances need not be equal across periods.
- The repeated cross-section analogue presented above only requires that $E[U_{it}U_{i,t-j}] = E[U_{it'}U_{i,t'-j}]$ for $j = 0$; covariances may differ among equispaced pairs of the U_{it}.

7. First difference methods

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

7.1. Models which justify condition (11)

- Whenever condition (11) holds, a can be estimated consistently from the difference regression method described in section 6.1.
- Section 6.1 presents a model which satisfies condition (11): the earnings residual has a permanent-transitory structure, decision rule (5) or (6) determines enrollment, and S_i is distributed independently of the transitory component of U_{it}.

- However, this model is rather special.
- It is very easy to produce plausible models that do not satisfy (11).
- For example, even if (12) characterizes U_{it}, if S_i in (6) does not have same joint (bivariate) distribution with respect to all ϵ_{it}, except for ϵ_{ik}, (11) may be violated.

- Even if S_i in (6) is distributed independently of U_{it} for all t, it is still not the case that (11) is satisfied in a general model.
- For example, suppose X_{it} is distributed independently of all U_{it} and let

  $$U_{it} = \rho U_{i,t-l} + V_{it},$$

  where V_{it} is a mean-zero, iid random variable and $|\rho| < 1$.
- If $\rho \neq 0$ and the perfect foresight decision rule characterizes enrollment, (11) is not satisfied for $t > k > t'$ because

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

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

- In that case

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

so $E[U_{it} - U_{it'}|d_i = 1] = 0$ only for t, t' such that $k - t' = t - k$.

- Thus (11) is not satisfied for all $t > k > t'$.

- For more general specifications of U_{it} and stochastic dependence between S_i and U_{it}, (11) will not be satisfied.

7.2. More general first difference estimators

- Instead of (11), assume that

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

- Two new ideas are embodied in this assumption.
- In place of the assumption that $U_{it} - U_{it'}$ be conditionally independent of $X_{it} - X_{it'}$ and d_i, we only require uncorrelatedness.

- Also, rather than assume that $E[U_{it} - U_{it'} | d_i, X_{it} - X_{it'}] = 0$ for all $t > k > t'$, the correlation needs to be zero only for some $t > k > t'$.
- For the appropriate values of t and t', least squares applied to the differenced data consistently estimates α.

Example That Satisfies (22) but not (12)

U_{it} is covariance stationary, (23a)

U_{it} has a linear regression on U_{ik} for all t

$(i.e., E[U_{it}|U_{ik}] = \beta_{tk} U_{ik})$, (23b)

U_{it} is mutually independent of (X_{ik}, S_i) for all t, (23c)

α is common to all individuals (so the model is of the fixed coefficient form), (23d)

The environment is one of perfect foresight where decision rule (6) determines participation. (23e)

- Under these assumptions, condition (22) characterizes the data. **Prove.**

- To see this note that (23a) and (23b) imply there exists a δ such that

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

and

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

- Now observe that

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

- But, as a consequence of (23c),

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

  since $E[\omega_{it}] = 0$ and because (23c) guarantees that the mean of ω_{it} does not depend on X_{ik} and S_i.
- Similarly,

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

  and thus (22) holds.

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

7.3. Anomalous features of first difference estimators

- Nearly all of the estimators require a control group (i,e., a sample of non-trainees), The only exception is the fixed effect estimator in a time homogeneous environment.
- In this case, if condition (11) or (22) holds, if we let $X_{it}\beta = \beta_t$ to simplify the exposition, and if the environment is time homogeneous so $\beta_t = \beta_{t'}$ then

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

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

- Most of the procedures considered here can be implemented using only post-program data.
- The covariance stationary estimators of sections 5.3 and 6.3, certain repeated cross-section estimators and first difference methods constitute an exception to this rule.
- In this sense, these estimators are anomalous.

- Fixed effect estimators are also robust to departures from the random sampling assumption.
- For instance, suppose condition (11) or (22) is satisfied, but that the available data oversample or undersample trainees (i.e., the proportion of sample trainees does not converge to $p = E[d_i]$).
- Suppose further that the analyst does not know the true value of p.
- Nevertheless, a first difference regression continues to identify α.
- Most other procedures do not share this property.

8. Non-random sampling plans

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

- The data available for analyzing the impact of training on earnings are often non-random samples.
- Frequently they consist of pooled data from two sources:
  1. a sample of trainees selected from program records and
  2. a sample of non-trainees selected from some national sample.
- Typically, such samples overrepresent trainees relative to their proportion in the population.
- This creates the problem of choice based sampling analyzed by Manski and Lerman (1977) and Manski and McFadden (1981).

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

- Both of these sampling situations combine the following types of data:
  1. Earnings, earnings characteristics, and enrollment characteristics (Y_{it}, X_{it} and Z_i for a sample of trainees ($d_i = 1$),
  2. Earnings, earnings characteristics, and enrollment characteristics for a sample of non-trainees ($d_i = 0$),
  3. Earnings, earnings characteristics, and enrollment characteristics for a national 'control' sample of the population (e.g., CPS or Social Security Records) where the training status of persons is not known.

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

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

8.1. The IV estimator: Choice-based sampling

- If condition (8a) is strengthened to read

$$E[X'_{it}U_{it}|d_i] = 0, \qquad E[g(Z^e_i)U_{it}|d_i] = 0, \tag{24}$$

and (8b) is also met, the IV estimator is consistent for α in choice-based samples.

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

$$\begin{aligned}
&\begin{pmatrix} \frac{\sum X_{it}' X_{it}}{I_t} & \frac{\sum X_{it}' d_i}{I_t} \\ \\ \frac{\sum g(Z_i^e) X_{it}}{I_t} & \frac{\sum g(Z_i^e) d_i}{I_t} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} \frac{\sum X_{it}' Y_{it}}{I_t} \\ \\ \frac{\sum g(Z_i^e) Y_{it}}{I_t} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\sum X_{it}' X_{it}}{I_t} & \frac{\sum X_{it}' d_i}{I_t} \\ \\ \frac{\sum g(Z_i^e) X_{it}}{I_t} & \frac{\sum g(Z_i^e) d_i}{I_t} \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \begin{pmatrix} \frac{\sum X_{it}' U_{it}}{I_t} \\ \\ \frac{\sum g(Z_i^e) U_{it}}{I_t} \end{pmatrix} .
\end{aligned} \tag{25}$$

- Since (24) guarantees that

$$\operatorname*{plim}_{I_t \to \infty} \frac{\sum X_{it}' U_{it}}{I_t} = 0 \qquad \text{and} \qquad \operatorname*{plim}_{I_t \to \infty} \frac{\sum g(Z_i^e) U_{it}}{I_t} = 0, \quad (26)$$

and the rank condition (8b) holds, the IV estimator is consistent.

- In a choice based sample, let the probability that an individual has enrolled in training be p^*.
- Even if (8a) and (8b) are satisfied, there is no guarantee that condition (26) will be met without invoking (24).
- This is so because

$$\underset{I_t\to\infty}{\text{plim}} \frac{\sum X'_{it}U_{it}}{I_t} = E[X'_{it}U_{it}|d_i = 1]p^* + E[X'_{it}U_{it}|d_i = 0](1-p^*),$$

$$\underset{I_t\to\infty}{\text{plim}} \frac{\sum g(Z^e_i)U_{it}}{I_t} = E[g(Z^e_i)U_{it}|d_i = 1]p^* + E[g(Z^e_i)U_{it}|d_i = 0](1-p^*).$$

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

- In the case of random sampling, $p^* = \Pr[d_i = 1] = p$ and the above expressions are identically zero.
- They are also zero if (24) is satisfied.
- However, it is not necessary to invoke (24).
- Provided p is known, it is possible to reweight the data to secure consistent estimators under the assumptions of section 4.

- Multiplying eq. (1) by the weight

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

  and applying IV to the transformed equation produces an estimator that satisfies (26).

- It is straightforward to check that weighting the sample at hand back to random sample proportions causes the IV method to consistently estimate α and β.

8.2. The IV estimator: Contamination bias

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

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

$$\frac{\sum X'_{it}X_{it}}{I_C}, \quad \frac{\sum g(Z_i^e)X_{it}}{I_C}, \quad \frac{\sum X'_{it}Y_{it}}{I_C}, \quad \frac{\sum g(Z_i^e)Y_{it}}{I_C}$$

  which converge to the desired population counterparts where I_C denotes the number of observations in sample (C).

- Missing is information on the cross-products

$$\frac{\sum X'_{it}d_i}{I_C}, \quad \frac{\sum g(Z_i^e)d_i}{I_C}.$$

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

$$\operatorname*{plim}_{I_C\to\infty} \frac{\sum X'_{it}d_i}{I_C} = pE[X'_{it}|d_i = 1],$$

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

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

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

respectively.

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

8.3. Repeated cross-section methods with known training status and choice-based sampling

- The repeated cross-section estimators discussed in section 4 are inconsistent when applied to choice-based samples unless additional conditions are assumed.
- For example, when the environment is time-homogeneous and (11) also holds, $(\bar{Y}_t - \bar{Y}_{t'})/p$ remains a consistent estimator of α in choice-based samples as long as the same proportion of trainees are sampled in periods t' and t.

- If a condition such as (11) is not met, it is necessary to know the identity of trainees in order to weight the sample back to the proportion of trainees that would be produced by a random sample in order to obtain consistent estimators.
- Hence the class of estimators that does not require knowledge of individual training status is not robust to choice-based sampling.

8.4. Control function estimators

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

K_{it} depends on variables $\ldots, Y_{i,t+1}, Y_{it}, Y_{i,t-1}, \ldots, X_{i,t+1}, X_{it}, X_{i,t-1}, \ldots, d_i$ and parameters ψ, and

$$E[U_{it} - K_{it} | d_i, X_{it}, K_{it}, \psi] = 0, \tag{27a}$$

$$\psi \text{ is identified.} \tag{27b}$$

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

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

- The purged disturbance $\{U_{it} - K_{it}\}$ is orthogonal to the right-hand-side variables in the new equation.
- Thus (possibly non-linear) regression applied to (28) consistently estimates the parameters (α, β, ψ).

- Moreover, (27) implies that $\{U_{it} - K_{it}\}$ is orthogonal to the right-hand-side variables conditional on d_i, X_{it} and K_{it}:

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

- Thus if type (A) and (B) data are combined in *any* proportion, least squares performed on (28) produces consistent estimates of (α, β, ψ) provided the number of trainees and non-trainees in the sample both approach infinity.
- The class of control function estimators which satisfy (27) can be implemented without modification in choice-based samples.

- We encountered a control function in section 6.
- For the model satisfying (13) and (19),

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

so $\psi = (\rho, \beta, \alpha)$.

- The sample selection bias methods (d)-(e) described in section 4.2 exploit the control function principle.
- Our longer paper gives further examples of control function estimators.

9. Conclusion

- This paper presents alternative methods for estimating the impact of training on earnings when non-random selection characterizes the enrollment of persons into training.
- We have explored the benefits of cross-section, repeated cross-section and longitudinal data for addressing this problem by considering the assumptions required to use a variety of new and conventional estimators given access to various commonly encountered types of data.

- We also investigate the plausibility of assumptions needed to justify econometric procedures when viewed in the light of prototypical decision rules determining enrollment into training.
- Because many of the available samples are choice-based samples and because the problem of measurement error in training status is pervasive in many available control samples, we examine the robustness of the estimators to choice-based sampling and contamination bias.

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

Appendix of Section 3

3. Random coefficients and the structural parameter of interest

- We identify two different definitions associated with the notion of a selection bias free estimate of the impact of training on earnings.
- The first notion defines the structural parameter of interest as the impact of training on earnings if people are randomly assigned to training programs.
- The second notion defines the structural parameter of interest in terms of the difference between the post-program earnings of the trained and what the earnings in post-program years for these same individuals would have been in the absence of training.

- The two notions come to the same thing only when training has an equal impact on everyone or else assignment to training is random and attention centers on estimating the mean response to training.
- The second notion is frequently the most useful one for forecasting future program impacts when the same enrollment rules that have been used in available samples characterize future enrollment.

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

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

  Q2 'How do the post-program mean earnings of the trained compare to what they would have been in the absence of training?'

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

- In the presence of non-random assignment and variation in the impact of training among persons, the two questions have different answers.
- Question 2 is the appropriate one to ask if interest centers on forecasting the change in the mean of the post-training earnings of trainees when the same selection rule pertains to past and future trainees.
- It is important to note that the answer to this question is all that is required to estimate the future program impact if future selection criteria are like past criteria.

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

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

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

$$Y_{it} = X_{it}\beta + d_i\bar{\alpha}_i + \{U_{it} + d\varepsilon_i\}. \tag{29}$$

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

- Let
$$Y_{1it} = X_{it}\beta_1 + U_{1it}$$
be the wage of individual i in sector 1 in period t.
- Let
$$Y_{0it} = X_{it}\beta_0 + U_{0it}$$
be the wage of individual i in sector 0.
- Letting $d_i = 1$ if a person is in sector 1 and letting $d_i = 0$ otherwise, we may write the observed wage as

$$\begin{aligned} Y_{it} &= d_i Y_{1it} + (1 - d_i) Y_{0it} \\ &= X_{it}\beta_0 + E[X_{it} \mid d_i = 1](\beta_1 - \beta_0) d_i \\ &\qquad + \left[(X_{it} - E[X_{it} \mid d_i = 1])(\beta_1 - \beta_0) + U_{1it} - U_{0it}\right] d_i + U_{0it}. \end{aligned}$$

- Letting

$$\begin{aligned}
\bar{\alpha} &= E[X_{it} \mid d_i = 1](\beta_1 - \beta_0), \\
\varepsilon_i &= (X_{it} - E[X_{it} \mid d_i = 1])(\beta_1 - \beta_0) + U_{1it} - U_{0it} \\
\beta_0 &= \beta, \\
U_{0it} &= U_{it},
\end{aligned}$$

  produces eq. (29).
- In this model there is a fundamental non-identification result when no regressors appear in the decision rule (3).
- Without a regressor in (3) and in the absence of any further distributional assumptions it is not possible to identify $\bar{\alpha}$ unless $E[\varepsilon_i \mid d_i = 1, Z_i] = 0$ or some other known constant.

- To see this, note that

$$
\begin{aligned}
E\left[Y_{it} \mid d_i = 1, Z_i, X_{it}\right] &= X_{it}\beta + \bar{\alpha} + E[\varepsilon_i \mid d_i = 1, Z_i, X_{it}] \\
&\qquad + E[U_{it} \mid d_i = 1, Z_i, X_{it}], \\
E\left[Y_{it} \mid d_i = 0, Z_i, X_{it}\right] &= X_{it}\beta + E[U_{it} \mid d_i = 0, Z_i, X_{it}].
\end{aligned}
$$

- Unless $E[\varepsilon_i \mid d_i = 1, Z_i, X_{it}]$ is known, without invoking distributional assumptions it is impossible to decompose $\bar{\alpha} + E[\varepsilon_i \mid d_i = 1, Z_i, X_{it}]$ into its constituent components unless there is independent variation in $E[\varepsilon_i \mid d_i = 1, Z_i, X_{it}]$ across observations [i.e., a regressor appears in (3)].
- Without a regressor, $E[\varepsilon_i \mid d_i = 1, Z_i, X_{it}]$ is a constant which cannot be distinguished from $\bar{\alpha}$.

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

$$Y_{it} = X_{it}\beta + d_i\alpha^* + \{U_{it} + d_i(\varepsilon_i - E[\varepsilon_i \mid d_i = 1])\}, \qquad (30)$$

where

$$\alpha^* = \bar{\alpha} + E[\varepsilon_i \mid d_i = 1],$$

and content ourselves with the estimation of α^*.

- If everywhere we replace α with α^*, the fixed coefficient analysis of eq. (1) applies to (30).

- The parameter α^* answers Q2.
- It addresses the question of determining the effect of training on the people selected as trainees.
- This parameter is useful in making forecasts when the same selection rule operates in the future as has operated in the past.
- In the presence of non-random selection into training it does not answer QI.
- Indeed, without regressors in decision rule (3) this question cannot be answered unless specific distributional assumptions are invoked.

- Random assignment of persons to training does not usually represent a relevant policy option.
- For this reason, we will focus attention on question two.
- Hence, if the training impact varies among individuals, we will seek to estimate α^* in (30).
- Since eq. (30) may be reparametrized in the form of eq. (1) we work exclusively with the fixed coefficient earnings function.
- Heckman & Smith (1997) gives precise statements of conditions under which $\overline{\alpha}$ is identified in a random coefficient model.

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

- It is also important to note that any definition of the structural treatment coefficient is conditioned on the stability of the environment in which the program is operating.
- In the context of a training program, a tenfold expansion of training activity may affect the labor market for the trained and raise the cost of the training activity (and hence the content of programs).
- For either $\bar{\alpha}$ or α^* to be interesting parameters, it must be assumed that such effects are not present in the transition from the sample period to the future.

- If they are present, it is necessary to estimate how the change in the environment will affect these parameters.
- In this note, we abstract from these issues, as well as other possible sources of interdependence among outcomes.
- The resolution of these additional problems would require stronger assumptions than we have invoked here.

- Before concluding this section, it is important to not that there is a certain asymmetry in our analysis which, while natural in the context of models for the evaluation of the impact of training on earnings, may not be as natural in other contexts.
- In the context of a training program (and in the context of the analysis of schooling decisions), it is natural to reason in terms of a latent earnings function Y_{it}^* which exists in the absence of schooling or training options.
- "U_{it}" can be interpreted as latent ability or as skill useful in both trained and untrained occupations.

- Because of the natural temporal ordering of events, pretraining earnings is a natural concept and α_i is the markup (in dollar units) of skills due to participation in training.
- Note that nothing in this formulation restricts agents to have one or just two skills.
- Training can uncover or produce a new skill or enhance a single common skill.
- Parameter α^* is the gross return to training of the trained before the direct costs of training are subtracted.

- In other contexts there is no natural temporal ordering of choices.
- In such cases the concept of α^* must be refined since there is no natural reference state.
- Corresponding to a definition of the gross gain using one state as a benchmark, there is a definition of gross gain using the other state as a benchmark.

- In the context of the Roy model [discussed following equation (6)], it is appropriate for an analysis of economic returns to outcomes to compute a gross gain for those who select sector 1 which compares *their* average earnings in sector 1 with what they would have earned on average in sector 0 and to compute a gross gain for those who select sector 0 which compares their average earnings in sector 0 with what they would have earned on average in sector 1.

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

$$\begin{aligned} Y_{1i} &= \mu_1 + U_{1i} \\ Y_{0i} &= \mu_0 + U_{0i}. \end{aligned}$$

- In this notation

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

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

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

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

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

- Both coefficients compare the average earnings in the outcome state and the average earnings in the alternative state for those who are in the outcome state.

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

Return to main text