The Impact of Social Interventions: Nonparametric Identification from Choice-Based Samples

Ricardo Paes de Barros**

Abstract
This paper presents conditions for the identification of the average effect of training constrained to the population of trainees. All of the identification conditions we consider are based on a variety of conditional independence assumptions. More specifically, the participation status and post-training latent earnings are assumed to be conditionally (mean) independent. The identification conditions differ according to the conditioning set used. Nowhere are any functional forms imposed. In other words, a nonparametric approach has been taken. Moreover, the conditions are robust to choice-based sampling.

Keywords: Identification, Average Treatment Effect, Conditional Independence Assumptions.

JEL Codes: C14, C21, C25.

*Submitted in May 2011. Revised in August 2011. I would like to thank Lars Hansen, James Heckman, Tom Mroz as well as my colleagues Marcelo Dabos and Stephen Cameron for very helpful comments. Naturally, the responsibility for all errors is my own.

**Under Secretary of Strategic Actions at the Brazilian Executive Office of the President (SAE/PR). E-mail: ricardo.barros@presidencia.gov.br

Brazilian Review of Econometrics
v. 30, n° 2, pp. 195–228 November 2010
1. Introduction

In this paper, we investigate alternative conditions which are sufficient to identify the impact of interventions using choice-based samples without evoking functional-form assumptions. Models for social interventions are natural examples of the dummy endogenous variable models proposed by Heckman (1975, 1978) and Lee (1978). Heckman and Macurdy (1986) present an excellent general discussion of dummy endogenous variable models as well as of their applications to the analysis of interventions. For a good introduction to the selection problems involved in the evaluation of social interventions, see Barnow et al. (1980). The specific identification question we consider has been already extensively investigated by Heckman and Robb (1985a,b). This paper extends some of their results. The focus is on nonparametric identification from choice-based samples of intervention effects which are stochastic as well as time-varying. Since we explicitly model the intervention effects as stochastic, our model closely resembles those studied by Lee and Trost (1978) and Lee (1978). Alternatively, we can classify the model we investigate as a switching model with known sample separation and outcomes observed in both regimes. The evaluation of social interventions also has a long tradition in statistics; for recent reviews of this literature, see Anderson et al. (1980) and Rubin (1984). In the statistical literature, the intervention effects are also usually explicitly treated as stochastic.

Mainly to simplify the exposition, most previous authors have studied the identification of these models under unnecessary restrictive assumptions. Examples are the normality assumption in Goldberger (1972) and the linearity of the regression functions in Heckman and Robb (1985a,b). A careful reading of these papers would clearly indicate that these assumptions are not generally required for identification. Since our goal is to look deeply into the identification mechanisms, we find preferable not to contaminate our analysis with unnecessary assumptions. In fact, the entire analysis will be conducted free of functional-form assumptions. A consequence, however, is a more abstract discussion. To partially compensate that, throughout the paper, we provide examples which should hopefully provide a link to the previous studies in the literature.

Consider a model in which, with some probability, an underlying stochastic process \( y(t) : -\infty < t < \infty \) may suffer an intervention at a given date \( T_0 \). Whenever an intervention occurs, the underlying process switches to a new stochastic process denoted by \( v(t) : T_0 < t < \infty \). For any \( t > T_0 \), the impact of the intervention is then defined as \( v(t) - y(t) \) and denoted by \( z(t) \). Thus, in general, the impact itself follows a stochastic process. Let \( d \) be a dichotomous random variable which indicates whether an intervention occurred. Then define

---

1. See also Maddala (1983, ch. 9) and Judd and Kenny (1981).
The Impact of Social Interventions

\[ w(t) = \begin{cases} 
  y(t) & \text{when } t < T_0, \\
  y(t) + dz(t) & \text{when } t > T_0
\end{cases} \]  

(1)

Whether an intervention has occurred is assumed known.

Moreover, we assume that, for a period of time \( I \), \( w(t) \) can also be observed. However, no direct information on \( z(t) : T_0 < t < \infty \) is available. \( y(t) \) can only be observed for those periods it coincides with \( w(t) \). It follows from (1) that \( w(t) = y(t) \) if \( t \leq T_0 \) or \( t > T_0 \) and no intervention has occurred.

We have a concrete application in mind. For us \( y(t) : -\infty < t < \infty \) is a worker’s earnings profile, the intervention is a training program, and \( z(t) : T_0 < t < \infty \) represents the program’s impact on earnings. Henceforth, in order to simplify the exposition, we refer to \( w(t) : -\infty < t < \infty \) as (total) earnings, \( y(t) : -\infty < t < \infty \) as latent earnings, and \( d \) as participation status.

Our ultimate goal is to estimate the distribution of \( z(t) : t > T_0, t \in I \). This paper, however, only considers the identification question. Nevertheless, as we show, these conditions are very indicative of how consistent estimators can actually be constructed. Section 2 is devoted to describing the concept of identification we are going to use.

Surprisingly, this does not depend upon the value of the outcome in both regimes, \( y(t) \) and \( v(t) = y(t) + z(t) \). For instance, whether workers take training or not depends, in general, on their latent earnings as well as on the impact of training. Thus, in general, \( d \) will depend upon \( y(t) : -\infty < t < \infty \) and \( z(t) : T_0 < t < \infty \). Usually, econometricians, as opposed to statisticians, explicitly model this relationship. For example, in Lee’s (1978) model of unionism \( d = l \) if and only if \( z(T_0) > C \); in Heckman and Robb’s (1985, p. 179) model for the impact of a training program \( d = l \) if and only if \((\frac{1}{1-r})z(T_0) > C\), where \( C \) is a measure of the cost of the respective intervention and \( r \) is the interest rate. As far as identification is concerned, however, explicit models for the participation decision, like these ones, are of little help. To emphasize this point, we study identification without modeling the relationship between \( d \) and the potential outcomes in the two regimes. This approach should clarify to what extent previous studies have been contaminated by unnecessary assumptions.

Most empirical evaluations of training programs\(^4\) have used samples stratified with respect to the participation status. In general, trainees are overrepresented

in these samples. This special kind of stratified sample is known as a choice-based sample. The effect of this particular sampling scheme on estimation has recently received large attention in the econometric literature.\(^5\) Traditionally, in order to solve the problem, the intervention-occurrence probability, \(P[d = 1]\), is assumed known. However, for a large number of applications, this is not a realistic assumption. For example, in the previously referenced literature on evaluation of training programs, the proportion of trainees in the overall population is rarely known. In fact, none of these studies has actually ever used this information. Accordingly, we opted to limit the search for identification conditions which are robust to choice-based sampling, i.e., these that do not depend on the knowledge of the probability with which an intervention may occur, \(P[d = 1]\).

Actually, the samples used in the evaluation literature have additional inconveniences. They could be appropriately referred to as enriched-contaminated samples, since they are formed by a random sample in which the information on participation status, \(d\), is missing, enriched by a sample of trainees. The enrichment device makes the sample choice-based. The fact that \(d\) is unobserved in the random sample, i.e., the contamination raises two questions. First, it precludes \(P[d = 1]\) from being estimated from the random segment of the sample. Secondly, it introduces measurement error if, as is usually done, we set \(d = 0\) for all individuals in the random sample. When \(P[d = 1]\) is small, a situation of important practical relevance, any bias due to measurement error of \(d\) should be insignificant. The entire analysis in this paper does not require \(P[d = 1]\) to be known, however, it does assume that there is no measurement error associated with \(d\). Hence, the results we derive can only be applied as an approximation to the analysis of enriched-contaminated samples when \(P[d = 1]\) is small.

Identification depends crucially on the available information. Stratified samples impose severe restrictions on the econometrician’s information. In such cases, it is impossible to recover the distribution of the variables used in the stratification from the sample. The sample provides mainly information about the conditional distribution of the variables that are not used in the stratification, compared to those which actually are. Section 3 describes the information econometricians are assumed to be endowed with throughout the paper.

In Section 4, we show that conditions for identification of first moments are much weaker than those sufficient to identify higher moments or other features of the probability distribution of the impact of training on earnings. Surely, this is an artifact of the expectation operator’s linearity property as well as of the way we define the training impact. As we show, only the average training effect can be identified when arbitrary dependence between latent earnings and the training impact is permitted. To identify other features of the distribution, these variables

---

are required to be conditionally independent. We suspect that in most applications this is not a realistic assumption. Hence, for a large variety of applications, only the first moments of \([z(t) : t > T_0, t \in I]\) are identifiable. Accordingly, we focus our attention on conditions for identification of average training effects.

As emphasized by Heckman and Robb (1985a, p. 174), Rosenbaum and Rubin (1983), and discussed in detail in Section 4, in the estimation of average training effects we also have an interesting dichotomy: the average effect RESTRICTED to the population of trainees requires much weaker assumptions to be identified than the average effect on the ENTIRE population. This fact should be expected, since the only information we might have on training comes from those individuals who actually took training. This question of how the average effect for the entire population can be identified is equivalent to the classical problem of how to obtain the unconditional mean from a censored sample. Without regressors or some functional-form assumptions, it is virtually impossible to recover the unconditional mean. Accordingly, we devote most of our analysis to deriving identification conditions for the mean effect of training restricted to the population of trainees. The analysis of the extra conditions sufficient to identify the unconstrained average is delayed up to Section 9.

Thus, the main objective of this paper is to present conditions for the identification of the AVERAGE effect of training CONSTRAINED to the population of trainees. All of the identification conditions we consider are based on a variety of conditional independence assumptions. More specifically, they assume that the participation status and post-training latent earnings are conditionally (mean) independent.\(^6\) The identification conditions differ according to the conditioning set they use. Nowhere is any functional form imposed. In other words, a nonparametric approach has been taken. Moreover, the conditions are robust to choice-based sampling.

In Section 5, we explore the binary nature of the participation status and prove several special and interesting relationships among the concepts of independence, mean independence, and uncorrelation. For example:

(i) if \(d\) is mean independent of \(y\), then \(d\) and \(y\) are independent;

(ii) if \(d\) and \(y\) are uncorrelated, then \(y\) is mean independent of \(d\). These special properties are then used to state the identification conditions in a form which is more intuitive and has a direct economic interpretation based on the participation decision process.

Naturally, it is the endogeneity of \(d\) that is the essence of the identification problem. Unfortunately, however, in this case, the instrumental variable, which is the traditional econometric technique to cope with endogeneity, is inappropriate for three reasons:

\(^6\)\(y\) is said to be mean independent of \(X\) when the expectation of \(Y\) exists, is finite, and \(E[y|X] = E[y]\).
(i) the stochastic nature of the intervention impact implies that we have to work with a random coefficient model. The task of finding valid instruments in this case is much harder than usual. Here, an instrument is also required to be uncorrelated with the random coefficient, which turns out to be the training impact in this particular case.

(ii) The model is totally nonparametric. The use of instrumental variables in this context has not received much attention and consequently has not been developed yet.

(iii) Finally, and perhaps more importantly, we have a choice-based sample where \( P[d = l] \) is unknown.

Heckman and Robb (1985a, p. 198) show that in the absence of information on \( P[d = l] \) even the existence of valid instruments is not sufficient for identification. To solve this simultaneity problem, we impose constraints on the information available to the agent who decides whether an intervention is going to occur. The same device has been used repeatedly in the econometric literature. Zellner et al. (1966) identify production functions from a cross-section assuming that risk-neutral profit-maximizing entrepreneurs are unable to predict the idiosyncratic shocks that will shift their production functions in the next period. Goldberger (1972) shows that the impact of an educational program may become unidentified if the program administrator selects children based on their ability, which is assumed to be unknown to the econometrician. However, he also shows that if program administrators cannot observe the children’s ability and are forced to base their decisions on some pre-test scores, which can be observed by the econometrician, then the impact of the program becomes identifiable. In this paper, we show that as far as the econometrician can predict future LATENT earnings as well as the decision maker at the decision time, then the impact of training is identified.

The identification conditions vary according to what information is available to the econometrician. As the econometrician’s information becomes richer, the identification requirements become weaker. The availability of panel data is essential to assure that the conditions for identification are of practical interest. In Section 4, the econometrician’s information set includes only a set of basic individual characteristics. In Section 6, on the control function method, pre-training earnings are added to this basic information set. Alternatively, in Section 7 about the eliminant method, some time-invariant latent variables are added. This class of models includes the fixed-effect model as well as Hause’s (1977) random growth model as special cases. In Section 8, we design a procedure which nests all previous ones. The procedure illustrates the great flexibility as well as the limitations of the methodology we develop. The section can be summarized as follows. Suppose the latent earnings process follows a linear random coefficient model, i.e.,

\[ \text{See Example 7 on page 16.} \]
where $b$ is a $k \times 1$ vector of time-invariant random coefficients. If the information available to the decision maker at $T_0$ consists of $[b, [u(t) : T_0 - m < t < T_0], [x(t) : t \in I], v]$, where $v$ is an unobserved component which is assumed to be independent of $u(t)$ for all $t > T_0$, then the average impact of training is identified as long as the econometrician has access to panel information about, at least, $k + m$ periods of pre-training earnings. However, if $v$ is correlated with $u(t)$ for some $t > T_0$ then, in general, the average training effect is not identified anymore. The general idea behind this methodology is to obtain identification by putting bounds on how much additional information the decision maker can have at the decision time vis-à-vis the econometrician.

2. Identification: Definition

Let $H$ denote the econometrician’s information set, i.e., the set of random variables he can observe. The best an econometrician can do then is to learn the probability distribution of $H$. Accordingly, we define a parameter as identifiable when it can be expressed as a KNOWN function of the distribution of $H$. Thus, once the econometrician learns the distribution of $H$, he automatically has sufficient information to determine the true value of any IDENTIFIABLE parameter. It is worthwhile noticing that the knowledge of such function, relating the distribution of $H$ to the parameter of interest, is a direct reflection of the econometrician’s prior information.

Example 1: Let $a = E[g(H, U)]$ be the parameter of interest. $U$ is assumed to be a random variable which cannot be observed by the econometrician. Without prior information $a$ is not identified. The knowledge of $g$ is not sufficient for identification. However, if $g$ and the CONDITIONAL distribution of $U$ given $H$ are known then $a$ is identified. Moreover, if $g$ is known and additively separable, then the knowledge of the MARGINAL distribution of $U$ suffices for the identification of $a$. Finally, if $g(H, U) = f(H) + U$, then $a$ is identified whenever $f$ and the mean of $U$ are known.

This definition of identification is precisely the one used by Elbers and Riddet (1982), Heckman and Singer (1984) and Manski (1985, 1986). However, in a paper closely related to this one, Heckman and Robb (1985a) take a different approach. They define a parameter as identifiable if and only if it can be consistently estimated.

The concept of identification we use permits us to separate out the identification question from the statistical problem of constructing a consistent estimator. This does not only simplify the analysis, but it also helps to separate the assumptions needed for identification from those used to prove consistency. Since we have little
to say about consistency itself, this approach seems to give a much clearer account of the identification question.

As already mentioned, most of the analysis will only consider conditions for identification of the average training effect on earnings. That is the case in Sections 6 to 9 as well as in part of Section 4. We show that once the identification assumptions are met, the average effect can be estimated by regression analysis. In this case, regularity conditions sufficient for consistency are well studied; they can be found in Jennrich (1969)\(^8\) and Stone (1977)\(^9\) for the parametric and non-parametric cases, respectively. In Section 4, however, we also study conditions for identification of the entire distribution. In this case, the estimation procedure our analysis suggests would involve solving an integral equation whereas regularity conditions for consistency deserve further research.

3. The Econometrician’s Information Set

Let \( E = (S, A, P) \) be our underlying probability space. All random variables are assumed to be defined in this space and have first moments which exist and are finite. Let \( I \) denote the long interval of the \( m \) period the econometrician can observe. We assume that his information set, \( H \), is made of three components: the stochastic process \( [w(t) : t \text{ in } I] \) which was defined in (1), the dichotomous random variable \( d \) which indicates whether an intervention occurred, and a random vector \( X \) which can be thought of as a set of observed characteristics.

Some clarifying remarks about the nature of the random vector \( X \) are in order: in the entire analysis, we will never need to make any assumption about \( X \). So, some components of \( X \) can be exogenous while others are endogenous. Different components may stand for different characteristics or for values of the same characteristic in different time periods. These time periods may refer to the past, the present or the future. In other words, in \( X \), we store all extra information the econometrician has besides \( d \) and \( [w(t) : t \text{ in } I] \). In several circumstances, however, it will be convenient to be more specific and let \( X \) denote a \( km \)-vector formed by all \( m \) observations for a \( k \)-vector \( x(t) \) of observed time-varying characteristics. In this case, for future reference, we write

\[
X = [x(t) : t \text{ in } I]
\]  

As we show in the next section, the key to identification is to make the following three underlying variables conditionally independent: latent earnings, program’s impact, and participation status. We plan to reduce and eventually eliminate the dependence between these random variables by including \( X \) in the conditioning set. Since that is the goal, nothing says we have to include all observed characteristics in

\(^8\)See also Amemiya (1983).
the conditioning set. Accordingly, whenever necessary, we redefine \( X \) excluding those observed characteristics the econometrician would not choose to include in his conditioning set.

In summary, we assume that the econometrician’s information is given by \( H = \left[ \{w(t) : t \in I, d, X\} \right] \). So, one possibility would be to endow the econometrician with the knowledge of the joint distribution of those random variables in \( H \). Since that is what he could recover, at least asymptotically, if increasing random samples were available.

We object to this assumption because, as we mentioned before, the econometrician rarely has access to random samples. In general, the available samples are stratified with respect to \( d \). Stratification on \( X \) is also common, particularly in the training evaluation literature in which matched samples are widely used. When a stratified sample is available, without extra assumptions, only conditional distributions are obtainable.

Due to these facts, \( H \) is redefined to exclude \( d \) and \( X \). Then the econometrician is endowed only with the distribution of \( H = \{w(t) : t \in I\} \) conditional on \( d \) and \( X \). Although that is our basic definition of the econometrician’s information, on several occasions, we consider what else could be identified if the marginal distribution of \( X \) or its distribution condition on \( d \) were known. Everywhere, however, the marginal distribution of \( d \) is assumed to be unknown, contrary to the traditional approach to choice-based samples.

4. Identification: General Results

The ultimate object of our analysis is the probability distribution of \( \{z(t) : t > T_0, t \in I\} \). However, in this paper, we confine ourselves to study only the probability distribution of \( z(r) \) for a fixed \( r \). For simplicity, we adjust our clock in such a way that \( T_0 < 0 \) and \( r = 0 \). In addition, we refer to \( w(0), y(0) \) and \( z(0) \) as \( w, y \) and \( z \), respectively.

Let \( L(w|d, X) \) denote the probability distribution law of \( w \) conditional on \( d \) and \( X \). As discussed in the previous section, \( L(w|d, X) \) is assumed to be known by the econometrician. Our task is to find sufficient conditions to identify the distribution of \( z \) conditional on \( X, L(z|X) \). In other words, we want to find a known function, \( \theta \), such that \( L(z|X) = G(L(w|d, X)) \). As we would expect, without strong prior knowledge about the joint distribution of \( (y, z, d, X) \) such a function would not exist.

We begin with the following benchmark set of assumptions: suppose that within groups which are homogeneous with respect to the observed characteristics, interventions occur at random and the magnitude of their impact is independent of the prevailing earnings level. In this case, \( L(z|X) \) could be identified. This example turns out to be extremely helpful to illustrate the basic structure of the

---

10 See example 8 on page 16.
identification problem. More formally, assume that for a suitable choice of $X$.

**Assumption 1:** $y$, $z$ and $d$ are mutually independent conditional on $X$.

Notice that even though A1 holds for a given choice of $X$, it may fail when we include OR exclude some variables. In particular, including all observed characteristics in $X$ is not necessarily the most appropriate choice. In this section, if we wish, we could also include (past) values for total earnings in the conditioning set. The entire analysis would still apply. Since that is the main topic of Section 6, and to simplify the exposition, we keep conditioning only on $X$. A1 has three basic implications:

\[
\begin{align*}
L(y|d, X) &= L(y|X), \quad (3) \\
L(y + z|d, X) &= L(y + z|X), \quad (4) \\
L(y + z|X) &= L(y|X)(+L(z|X)) \quad (5)
\end{align*}
\]

where (+) denotes convolution. Now, by the definition of $w$ in (1)

\[
\begin{align*}
L(w|1, X) &= L(y + z|1, X), \\
L(w|0, X) &= L(y|0, X)
\end{align*}
\]

Moreover, by (3) and (4)

\[
\begin{align*}
L(y|0, X) &= L(y|X), \\
L(y + z|1, X) &= L(y + z|X)
\end{align*}
\]

Consequently,

\[
\begin{align*}
L(y|X) &= L(y|0, X) = L(w|0, X), \quad (7) \\
L(y + z|X) &= L(y + z|1, X) = L(w|1, X) \quad (8)
\end{align*}
\]

Therefore, (5) can be rewritten as

\[
L(w|l, X) = L(w|0, X) \oplus L(z|X). \quad (9)
\]

Applying the Fourier transform, $F$, to equation (9) we obtain

\[
F[L(w|1, X)] = F[L(w|0, X)] \cdot F[L(z|X)]
\]

Therefore, $L(z|x)$ can be expressed as a known function of $L(w|d, X)$ as follows

11See example 8 on page 16.
\[ L(z|X) = F^{-1} \left[ \frac{F[L(w|l, X)]}{F[L(w|0, X)]} \right] \] (10)

This result can be summarized as the following theorem:

**Theorem 1:** If A1 is satisfied, then \( L(z|X) \) is identified.

**Example 2:** Let \( U(0) \) denote the utility of the relevant decision-maker when the worker does not engage in training; likewise, let \( U(l) \) denote the corresponding utility when the worker does engage in training. Let \((X, v)\) be the decision maker’s information at the decision time. \( X \) is assumed to be observed by the econometrician whereas \( v \) is not. Thus, the worker engages in training, \( d = l \), if and only if \( E[U(l) - U(0)]X, v > 0 \). Let \( y = f(X) + u \) and \( z = g(X) + e \). Assume that \((v, u, e)\) are conditionally mutually independent given \( X \). Since A1 is satisfied by Theorem 1, \( L(z|X) \) is identified. Next, to illustrate how \( L(z|X) \) can actually be obtained from \( L(w|d, X) \), assume that \( u \) and \( e \) are independent of \( X \) and normally distributed with zero mean and variances \( a \) and \( b \). Thus, \( L(w|l, X) = N(f(X) + g(X); a + b) \) and \( L(w|0, X) = N(f(X); a) \). The corresponding characteristic functions then are

\[
F[L(w|1, X)](t) = EXP \left[ it(f(X) + g(X)) - (a + b)t^2/2 \right]
\]

and

\[
F[L(w|0, X)](t) = EXP \left[ itf(X) - at^2/2 \right]
\]

Thus, \( L(z|X) \) can be recovered by applying equation (10). More specifically,

\[
F[L(z|X)](t) = F[L(w|1, X)](t)/F[L(w|0, X)](t) = EXP \left[ itg(X) - bt^2/2 \right]
\]

and

\[
L(z|X) = N(g(X); b)
\]

Notice that whenever the marginal distribution of \( X \) is known and \( L(z|X) \) is identified, we can also identify \( L[z] \) by integrating out \( X \).

We now turn to the study of what can be identified when at least one of the conditions (3)-(5) fails. To start with, suppose (4) fails. More specifically, assume:

**Assumption 2:**

(i) \( y \) and \( d \) are independent conditional on \( X \),

(ii) \( y \) and \( z \) are independent conditional on \( X \) and \( d \).

Notice that A2 differs from A1 exactly to the extent that in A2 \( z \) and \( d \) are not required to be conditionally independent given \( X \). So, similarly to A1, A2 (i) and (ii) imply respectively that
\[ L(y|d, X) = L(y|X), \quad (3) \]

\[ L(y + z|d, X) = L(y|d, X) \oplus L(z|d, X) \quad (5^*) \]

As before, by the definition of \( w \) and (3)

\[ L(w|1, X) = L(y + z|1, X), \quad (6) \]

\[ L(w|0, X) = L(y|0, X) = L(y|X), \quad (7) \]

\[ L(y|1, X) = L(y|X) \]

So, we can rewrite (5*) for \( d = 1 \) as

\[ L(w|1, X) = L(w|0, X)L(z|1, X) \quad (9^*) \]

Moreover

\[ F[L(w|1, X)] = F[L(w|0, X)], F[L(z|1, X)] \]

So

\[ L(z|l, X) = F - 1[F[L(w|1, X)]/F[L(z|0, X)]] \quad (10^*) \]

Since \( L(w|1, X) \) and \( L(w|0, X) \) are assumed known, this equation implies the following result:

**Theorem 2:** If \( A_2 \) is satisfied, then \( L(z|l, X) \) is identified.

This theorem implies that if \( z \) and \( d \) are not conditionally independent given \( X \), then, without additional assumptions, we can only identify the distribution of the training impact on earnings, \( z \), for the population of trainees, \( d = l \). This should have been expected, since the only information we might have about \( z \) comes from the population that actually suffered the intervention, \( d = l \). Hence, in order to identify the distribution of \( z \) for the entire population we will always need extra assumptions. To assume that \( z \) and \( d \) are independent conditional on \( X \) is one extreme possibility.

Finally, notice that whenever the distribution of \( X \) conditional on \( d = l \) is known, we can identify \( L[z|l] \) from \( L[z|1, X] \) by integrating out \( X \).

**Example 3:** Consider a variant of Example 2 where \( v \) and \( e \) are permitted to be arbitrarily correlated. Since \( v \) and \( u \) as well as \( e \) and \( u \) are still pairwise conditionally independent given \( X \),

**Assumption 2:** is satisfied and by Theorem 2 \( L(z|X) \) is identified.
To illustrate how \( L(z|l, X) \) can actually be calculated from \( L(w|d, X) \), assume, as before, that \( u \) and \( e \) are independent of \( X \) and normally distributed with zero mean and variances \( a \) and \( b \).

Consequently,

\[
L(w|1, X) = N(f(X) + g(X); a + b)
\]

and

\[
L(w|0, X) = N(f(X); a)
\]

The corresponding characteristic functions then are

\[
F[L(w|l, X)](t) = \text{EXP}[it.f(X) + g(X)] - (a + b)t^2/2
\]

and

\[
F[L(w|0, X)](t) = \text{EXP}[it.f(X) - at^2/2]
\]

Similarly to Example 2, we can then recover \( L(z|1, X) \) from \( L(w|d, X) \) as follows:

\[
F[L(z|l, X)](t) = F[L(w|l, X)](t)/F[L(w|0, X)](t) = \text{EXP}[it.g(X) - bt^2/2]
\]

and consequently

\[
L(z|1, X) = N(g(X); b)
\]

Next, suppose (5) fails while (3) and (4) still hold, i.e.,

Assumption 3: \( (y, z) \) and \( d \) are independent conditional on \( X \).

In this case, both (7) and (8) still hold, i.e.

\[
L(y|X) = L(w|0, X), \quad (7)
\]

\[
L(y + z|X) = L(w|l, X) \quad (8)
\]

So, both \( L(y + z|X) \) and \( L(y|X) \) can be recovered from \( L(w|d, X) \).

Nevertheless, (9) fails and so \( L(z|X) \) is not identified. Fortunately, however, by the linearity of the expectation operator

\[
E[z|X] = E[y + z|X] - E[y|X]
\]

So, by (7) and (8)

\[
E[z|X] = E[w|1, X] - E[w|0, X] \quad (11)
\]
Therefore, we still can identify \( E[z|X] \).

**Theorem 3:** If A3 is satisfied, then \( E[z|X] \) is identified.

Thus, whenever \( z \) and \( y \) are not conditionally independent given \( X \) and no additional assumptions are made, only first moments are identified. Further, notice from equation (11) that if \( y \) and \( z \) are just MEAN independent of \( d \) conditional on \( X \) and satisfied, then even if the econometrician’s information is limited to \( E[w|d, X] \) instead of to \( L(w|d, X) \), then \( E[z|X] \) still can be identified.

**Example 4:** Let us consider why we cannot identify the variance of the training effect when only A3 holds. In this case, (7) and (8) are satisfied. So, in particular,

\[
\begin{align*}
\text{Var}[w|0, X] &= \text{Var}[y|X], \\
\text{Var}[w|l, X] &= \text{Var}[y|X] + 2.Cov[y, z|X] + \text{Var}[z|X]
\end{align*}
\]

Differencing yields

\[
\begin{align*}
\text{Var}[z|X] &= \text{Var}[w|l, X] - \text{Var}[w|0, X] - 2.Cov[y, z|X]
\end{align*}
\]

This equation resembles (11), except for the last term, which unfortunately is unknown. Hence, a sufficient condition for \( \text{Var}[z|X] \) to be identified is \( Cov[y, z|X] = 0 \), which is indeed implied by the conditional independence between \( y \) and \( z \) given \( X \), equation (5).

**Example 5:** Consider the variant of Example 2 in which \( u \) and \( e \) are permitted to be arbitrarily correlated. Since \( v \) and \( u \) as well as \( v \) and \( e \) are still pairwise conditionally independent given \( X \), Assumption 3 is satisfied and by Theorem 3 \( E[z|X] \) is identified. Moreover,

\[
\begin{align*}
E[w|1, X] &= f(X) + g(X) \\
E[w|0, X] &= g(X)
\end{align*}
\]

So \( E[z|X] \) can be recovered from \( E[w|d, X] \) as follows,

\[
E[z|X] = E[w|l, X] - E[w|0, X] = g(X)
\]

Finally, if (4) and (5) fail, then without additional assumptions, we can only identify \( E[z|l, X] \), i.e.,

**Theorem 4:** If \( y \) and \( d \) are independent conditional on \( X \), then \( E[z|l, X] \) is identified.
Remark 1: If \( y \) is just MEAN independent of \( d \) conditional on \( X \), then \( E[z|l, X] \) is still identified.

Example 6: If in Example 5 \( v \) and \( e \) are permitted to be arbitrarily correlated, then only \( E[z|l, X] \) can be identified. In fact,
\[
E[z|l, X] = E[w|1, X] - E[w|0, X] = g(X)
\]
Whenever (3) fails, nothing of interest can be identified. In fact, the conditional independence of \( y \) and \( d \) is the crucial assumption. If it holds, we can identify \( E[z|l, X] \). If (4) also holds, i.e., \( z \) and \( d \) are also independent conditional on \( X \), then the average training effect for the entire population conditional on \( X \), \( E[z|X] \), is identified. Later, in Section 9, we derive weaker conditions which still suffice to identify \( E[z|X] \). Condition (5) \( (y \) and \( z \) are conditionally independent) is not necessary for the identification of first moments; it suffices to identify higher moments. We suspect that in most applications (5) will fail, for example, due to the presence of unobserved abilities which affect simultaneously \( z \) and \( y \). Accordingly, we constrain the subsequent analysis to the identification of conditional first moments, \( E[z|X] \) and \( E[z|l, X] \). The identification of \( E[z|l, X] \) is going to receive considerably more attention. We have two reasons to proceed this way. First, as emphasized by Heckman and Robb (1985a, p. 161), on several occasions \( E[z|l, X] \), instead of \( E[z|X] \), is the object of direct interest. Second, the assumptions necessary to obtain \( E[z|X] \) from \( E[z|l, X] \) are, to a large extent, unrelated to those necessary to identify \( E[z|1, X] \) itself. Hence, these assumptions can be studied separately.

So far, to a large extent, we have permitted the conditioning set, \( X \), to be arbitrary. That is an important peculiarity of this section. In particular, no specific use has been made of the longitudinal information on pre-training earnings. For example, if Al is satisfied for a specification of \( X \) which only includes contemporaneous values for the observed characteristics, then the training effect would be identified using just a post-training cross-section.

The economic theory proposes that if we want to choose a conditioning set to make \( d \) conditionally independent of \( y \), then a natural choice should be the information used by the agent who made the participation decision. Since pre-training earnings were certainly used by this decision maker, it would be wise to include this information in the conditioning set. This idea is formally implemented in a variety of ways in the following sections.

We end this section by considering the following two very important examples:

Example 7: Let \( D \) be the decision maker’s information set at the decision time. Then \( d = 1 \) if and only if \( E[U(l) - U(0)|D] > 0 \). Suppose that \( D \) includes \( X \), the characteristics the econometrician can observe. We want to show that if both the econometrician and the decision maker can predict future LATENT earnings
equally well, then \( E[z|1, X] \) is identified. Formally, we want to show that if
\[
E[y|D] = E[y|X]
\] (12)
then
\[
E[y|d, X] = E[y|x]
\] (13)

If this result is true, then from Remark 1 to Theorem 6 \( E[z|1, X] \) is identified.
To show that (13) holds, notice that since \( d \) and \( X \) are in \( D \) and \( X \) are in \( D \),
\[
E[y|D, d, x] = E[y|D]
\]
So, using equation (12) we obtain the even stronger result that
\[
E[y|D, d, X] = E[y|X]
\]

**Example 8:** With this example we want to illustrate that it is not always adequate to include all characteristics the econometrician can observe in the conditioning set. Suppose just two characteristics can be observed by the econometrician, \( X_g \) and \( X_b \). Suppose further that the decision maker’s information set is made of \( X_g \) and \( v \). Should we include \( X_b \) in the conditioning set? The answer depends on what else is assumed. To simplify the exposition, assume that
\[
y = f(X_g, X_b) + u
\]
Consider three cases:

(i) \((u, v)\) is independent of \((X_g, X_b)\), but \(v\) is correlated with \(u\). In this case, the average impact of training is not identified for all possible choice of \(X\).

(ii) \(u\) is independent of \((v, X_g, X_b)\), but \(v\) is correlated with \(X_b\). Now, the average impact of training is identified if we include both \(X_b\) and \(X_g\) in the conditioning set; we lose identification if only \(X_g\) is included.

(iii) \(u\) is independent of \((v, X_g)\), \(X_b\) is independent of \((v, X_g)\), but \((u, X_b)\) is NOT independent of \((v, X_g)\). In this case, the average impact is identified if \(X_g\) is chosen as the conditioning set whereas it is not identified if \((X_g, X_b)\) is included in the conditioning set.

5. **Independence versus Mean Independence**

In this section, we prove some interesting relationships among independence, mean independence and uncorrelation which are unique to binary random variables. These properties are going to be used in subsequent sections to derive
identification conditions which are more intuitively appealing and have simpler economic interpretations. Consider the following four conditions:

(i) \( E[d,y|X] = E[d|X].E[y|X] \),
(ii) \( E[y|d,X] = E[y|X] \),
(iii) \( E[d|y,X] = E[d|X] \),
(iv) \( y \) and \( d \) are independent conditional on \( X \).

In general (ii) \( \Rightarrow \) (i), (iii) \( \Rightarrow \) (i), (iv) \( \Rightarrow \) (ii) and even more interestingly that (iii) \( \Rightarrow \) (iv) and hence that (iii) \( \Rightarrow \) (ii).

**Lemma 1:** If \( E[d,y|X] = E[d|X].E[y|X] \), then \( E[y|d,X] = E[y|X] \), with probability one.

**Proof:** Consider three cases. First, assume that \( P[d=1|X] = 1 \). In this case, with probability one \( E[y|d,X] = E[y|l,X] \).

Moreover

\[
E[y|X] = E[y|l,X].P[d=1|X] + E[y|0,X].P[d=0|X] = E[y|1,X]
\]

since \( P[d=1|X] = 1 \) and \( P[d=0|X] = 0 \). Therefore,

\[
E[y|d,X] = E[y|X]
\]

with probability one. Secondly, assume \( P[d=0|X] = 1 \). Then, with probability one

\[
E[y|d,X] = E[y|0,X]
\]

Moreover,

\[
E[y|X] = E[y|l,X].P[d=1|X] + E[y|0,X].P[d=0|X] = E[y|0,X]
\]

since now \( P[d=1|X]=0 \) and \( P[d=0|X]=1 \). Therefore,

\[
E[y|d,X] = E[y|X]
\]

with probability one. Consider now the third case, \( 0 < P[d=1|X] < 1 \).

Note that

\[
E[d,y|X] = E[y|l,X].P[d=1|X],
\]
and by hypothesis

\[ E[d|y, X] = E[d|X].E[y|X] \]

Hence

\[ E[y|d, X].P[d = l|X] = E[y|X].P[d = l|X] \]

Because \( P[d = 1|X] \neq 0 \)
\[ E[y|X] = E[y|X] \]

Moreover,

\[ E[y|X] = E[y|0, X].P[d = 0|X] + E[y|l, X].P[d = l|X] \]

Because \( P[d = 0|X] \neq 0 \)
\[ E[y|0, X] = E[y|X] \]

Therefore,
\[ E[y|d, X] = E[y|X] \]

**Lemma 2**: If \( E[d|y, X] = E[d|X] \), then \( y \) and \( d \) are independent conditional on \( X \).

**Proof**: First notice that \( d \) and \( y \) are conditionally independent if and only if for any Borel set \( A \) of appropriate dimension

\[ P[d = 1, y \in A|X] = P[d = 1|X].P[y \in A|X] \quad (14) \]

If \( P[y \in A|X] > 0 \), the statement is trivial. Consider then the case in which \( P[y \in A|X] \neq 0 \). By hypothesis,
\[ E[d|y \in A, X] = E[d|X] \]

But
\[ E[d|y \in A, X] = P[d = l|y \in A, X] \]

and
\[ E[d|X] = P[d = l|X] \]
So,

\[ P[d = l | y \in A, X] = P[d = l | X] \]

Since \( P[y \in A | X] \neq 0 \), equation (12) follows immediately.

Lemma 1 and the fact that (iii) \( \Rightarrow \) (i) or Lemma 2 plus the fact that (iv) \( \Rightarrow \) (ii) imply that (iii) \( \Rightarrow \) (ii), i.e.,

**Corollary 1:** If \( E[d | y, X] = E[d | X] \), then \( E[y | d, X] = E[y | X] \).

This corollary as well as Lemmas 1 and 2 are very helpful to make identification conditions easier to interpret. For example, instead of assuming

\[ L(y | d, X) = L(y | X) \]

we can equivalently require that

\[ E[d | y, X] = E[d | X] \] (3*)

which has a much simpler interpretation. Thus, \( y \) does not appear in the participation decision rule, i.e., it was not part of the decision maker’s information at \( T_0 \).

In the subsequent analysis what is really necessary are assumptions that look like (ii). Since in this context (i) and (ii) are equivalent, we could have stated all assumptions in the form of either (i) or (ii). However, we decide to sacrifice some generality and state all assumptions in the more restrictive (iii) format. We believe that from all four options (iii) is the one which has the most direct and clearest economic interpretation.

The findings from this section are summarized in the following diagram:

<table>
<thead>
<tr>
<th>Binary case</th>
<th>General case</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) ⇔ (ii)</td>
<td>(i)</td>
</tr>
<tr>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>↑</td>
<td>(ii) ( \uparrow ) (iii) ( \downarrow )</td>
</tr>
<tr>
<td>(iii) ⇔ (iv)</td>
<td>( \downarrow ) (iv) ( \uparrow )</td>
</tr>
</tbody>
</table>

6. **Control Function Method**

Once we realize that without restrictions on the dependence between \( y \) and \( d \) it is virtually impossible to identify \( E[z | d] \) from choice-based samples, a search for suitable restrictions begins. Although the appropriateness of a given assumption is ultimately an empirical question, from a theoretical point of view we would like to have assumptions that are at the same time as general as possible and also
intuitively appealing. So, we could expect them to be justified or predicted by some general economic models.

In this section, we develop a very intuitive identification condition. The conditioning set used in Section 4 is enhanced by the inclusion of information on the pre-training earnings history.

In Section 4, we study the case in which \( \text{y} \) and \( \text{d} \) are independent conditional on \( \text{X} \). Doing the analysis conditional on \( \text{X} \) should reduce the dependence between \( \text{y} \) and \( \text{d} \). However, we suspect there are other unobserved variables which may affect both \( \text{y} \) and \( \text{d} \), leading to some dependence even after conditioning on \( \text{X} \). These same unobservables are also expected to be correlated with pre-training earnings. Thus, we hope that expanding the conditioning set by including past values of \( \text{y} \) will purge \( \text{d} \) from most of its dependence on \( \text{y} \). This is an example of a control function procedure as proposed by Heckman and Robb (1985a, p. 224) and implemented by Dickinson et al. (1984).

Intuitively, we assume that two agents with the same observed pre-training earnings history, \([y(t) : t \leq T_0, t \in I]\), and the same set of observed characteristics, \(X\), will have the same probability of participating in the training program independently of their future (latent) earnings profile. Formally,

**Assumption 4:** \( E[d|X, all(y)] = E[d|X, past(y)] \),

where past \((y) = [y(t) : t \leq T_0, t \in I]\) and all \((y) = [y(t) : t \in I]\). Since \(y(t) = w(t)\) for \(t \leq T_0\), past \((y) = past(w)\) where analogously we define past \((w) = [w(t) : t \leq T_0, t \in I]\). Notice that although past \((y)\) can be completely observed; all \((y)\) cannot. Since for \(t > T_0\), \(y(t)\) is unobserved for those who participate in the program.

By Corollary 1 we know that A4 implies that \(y\) is mean independent of \(d\) conditional on \(X\) and past \((y)\), i.e.,

\[
E[y|d, X, past(y)] = E[y|X, past(y)]
\]  \(15\)

Next, we show that A4 is a sufficient condition for the identification of \(E(z|l, X)\).

**Theorem 5:** If A4 is satisfied and the conditional distribution of \(H = [w(t) : t \in I]\) given \(d\) and \(X\) is known, then \(E[z|l, X]\) is identified.

**Proof:** From the definition of \(w\), equation (1)

\[
E[w|1, X, past(y)] = E[y|1, X, past(y)] + E(z|l, X, past(y)),
\]  \(16\)

\[
E[w|0, X, past(y)] = E[y|0, X, past(y)]
\]  \(17\)

From A4 or \((15)\)
Thus, taking the difference between (16) and (17) yields
\[
E[z|1, X, \text{past}(y)] = E[w|1, X, \text{past}(y)] - E[w|0, X, \text{past}(y)]
\]
Using the fact that past \(y = \text{past}(w)\) we obtain
\[
E[z|1, X, \text{past}(w)] = E[w|1, X, \text{past}(w)] - E[w|0, X, \text{past}(w)]
\]

Notice that both \(E[w|l, X, \text{past}(w)]\) and \(E[w|0, X, \text{past}(w)]\) can be derived from the conditional distribution of \([w(t) : t \in I]\) given \(d\) and \(X\). Since this distribution is assumed to be known, we have proved that \(E[z|l, X]\) is identified. Moreover, we can use our knowledge of this distribution to integrate out past \((w)\) and obtain \(E[z|l, X]\).

**Example 9:** Consider a variant of Example 6. Enlarge the decision maker’s information set to include \(u(T_0 - 1)\) which we denote by \(u_0\) for short. In this case, \(d = l\) if and only if
\[
E[U(l) - U(0)|X, v, u_0] > 0
\]
As before, we assume that
\[
y(t) = f(x) + u(t)
\]
and
\[
z(t) = g(X) + e(t)
\]
Moreover, assume that \(v\) and \([u(t) : -\infty < t < \infty]\) are conditionally independent given \(X\). It is easy to verify that under these assumptions A4 is satisfied. Consequently, by Theorem 5 \(E[z|l, X]\) is identified as far as \(y(T_0 - l) = w(T_0 - l)\) can be observed by the econometrician.

Notice that whenever the probability distribution of \(X\) conditional on \(d\) is known, we can also identify \(E[z|l]\) by integrating out \(X\).

### 7. Eliminant Method

Despite its generality and intuitive appeal, A4 imposes severe restrictions on the dependence between \(y\) and \(d\). In this section, we show that the largely used fixed-effect model is, in general, incompatible with A4.

We generalize the fixed-effect model along the lines of Heckman and Robb (1985a, p. 227), Madansky (1964), and Pudney (1982). Their procedures are based on instrumental variables methods and, consequently, are not robust to
choice-based sampling (see Heckman and Robb, 1985a, p. 198). Accordingly, we strengthen their assumptions in order to obtain identification conditions which are indeed robust to choice-based sampling. We follow Pudney by denoting the procedure suggested by our identification analysis as the eliminant method. In the next section, we show how to merge this procedure with the control function method.

Intuitively, we can see the incompatibility between a fixed-effect model and Assumption 4 as follows: suppose \( y \) and \( d \) are correlated, but they are independent conditional on earnings for a pre-training year. So, A4 is satisfied. This somehow indicates that \( d \) is correlated with \( y \) just because \( d \) depends on pre-training earnings, and pre- and post-training earnings are correlated. So, by conditioning on pre-training earnings, we purge the spurious or indirect dependence between \( y \) and \( d \). Suppose now earnings are made of a transitory and a permanent component. The permanent component is common to earnings in all periods whereas the transitory component is period-specific. Suppose \( d \) is correlated just with the permanent component. Assumption 4 would fail. Future earnings will always be useful to separate out the permanent from the transitory component. Hence, even conditional on pre-training earnings, \( d \) and \( y \) would still be correlated. The dependence is via the permanent component via pre-training earnings themselves.

Thus, one case where A4 fails is when \( d \) depends on earnings just via persistent components. In this section, these lasting components are assumed to be permanent, i.e., they are time-invariant. It is not clear whether this analysis could be extended to allow for some kind of time variation in these components. Moreover, everywhere we assume that the number of permanent components is KNOWN and that they affect earnings in a linear and additive fashion.

Let \( U \) denote an \( n \)-random vector of unobserved time-invariant characteristics, i.e., some latent variables. We assume

**Assumption 5:** \( E[d|X,U, all(y)] = E[d|X,U] \).

Thus, instead of increasing the conditioning set by including pre-training earnings, as in the control function method, we increase the conditioning set by introducing a vector of latent variables, \( U \). By A5 and Corollary 1,

\[
E[y(t)|d,X,U] = E[y(t)|X,U] \text{ for all } t \text{ in } 1 \tag{18}
\]

Further, assume the regression function is linear in \( U \):

**Assumption 6:** \( E[y(t)|X,U] = b(X,t) + B(X,t) \). \( U \) where \( B \) is a known function.

The fact that \( B \) is assumed to be known is what makes this analysis different from the work of Heckman and Robb (1985a), Madansky (1964) and Pudney (1982). This extra assumption allows us to identify the average effect of training

---


216 Brazilian Review of Econometrics 30(2) November 2010
without having to rely upon instrumental variables methods. In A6, we assume that the regression function is linear in $U$, but that the coefficients can be both time-varying and arbitrarily dependent on $X$. Note that $B$ is a $1 \times n$ row vector of known functions.

There are two ways to motivate this model for earnings. First, we can see these permanent components as time-invariant unobserved characteristics and make assumptions about the regression function in order to satisfy the linearity and additivity requirements. As a second approach, we can interpret the permanent components as random coefficients in a linear model.

Suppose $X$ is made of a set of characteristics $x(t)$ observed during a period of time $I$, i.e., we define $X$ as in (2)

$$X = [x(t) : t \in I]$$

Two largely used models satisfy A6. First we have

(i) $n = 2$, $b(X, t) = A.x(t), B(X, t) = [1, t]$ where $A$ is a $1 \times k$ row vector of unknown constants. This is the Hause’s (1977) model with fixed-effect and random growth. In the training literature, this model has been studied by Ashenfelter and Card (1985). In this case, the regression function can be simplified as

$$E[y(t)|X, U] = A.x(t) + U_1 + t.U_2$$

where $U_1$ and $U_2$ are the two components of $U$. The second model can be specified as follows

(ii) $n = k$, $b(X, t) = x(t)^T A, B(X, t) = x(t)^T$ where $A$ is a $k \times 1$ vector of unknown parameters. This is the linear random coefficient model extensively used in econometrics (see, for example, Judge et al., 1980, ch. 8, 9). Here, the regression function is simplified as

$$E[y(t)|X, U] = x(t)^T.[A + U]$$

Notice that A6 is satisfied.

Next, suppose the available pre-training longitudinal information is long enough:

**Assumption 7**: There are more than $n$ years of available pre-training data.

Therefore, there exists a subset of $I$ formed by exactly $n$ pre-training years. Let $J$ denote such a subset of $I$. Vertically stack $y(t)$ and $b(X, t)$ for all $t$ in $J$ to form the following $n \times 1$ vectors:

$$Y[: J] = [y(t) : t \text{ in } J],$$

$$bb[: J] = [b(X, t) : t \text{ in } J],$$
then vertically stack the \( n \)-row vectors \( B(X,t) \) to form the following \( n \times n \) matrix

\[
BB[:J] = [B(X,t) : t \in J]
\] (20)

Finally, suppose

**Assumption 8:** \( B : BB[:J] \) is non-singular almost surely.

In Hause’s model, this is automatically satisfied; in the linear random coefficient model, A8 is satisfied as long as at most one regressor is not time-varying, and a standard rank condition is satisfied. From (19), (20) and A6 we obtain

\[
\] (21)

where the fact that \( U \) is time-invariant plays a major rule. By (20) and AB we know that \( BB[:J] \) is a full rank square matrix. Hence, we can solve out for \( U \) in (21) to obtain,

\[
U = Inv(BB[:J]).(E[y[:J]|X,U] - bb[:J])
\]

In the subsequent analysis, we essentially ELIMINATE \( U \) by using this expression.

Let \( v(t) = y(t) - B(X,t).Inv(BB[:J]).Y[:J] \). We can now prove the following basic lemma:

**Lemma 3:** If A5 to A8 are satisfied, then \( E[v(t)|d,X,U] = E[v(t)|X] \) for all \( t \).

**Proof:** Initially, notice that from (18) and (19) and the definition of \( v(t) \)

\[
E[v(t)|d,X,U] = E[v(t)|X,U]
\] (22)

Further,

\[
\]

By A6 and (21)

\[
\]

\[
= b(X,t) - B(X,t).Inv(BB[:J]).bb[:J]
\] (23)

Since the right-hand side of (23) does not depend on \( U \), we have proved that

\[
E[v(t)|X,U] = E[v(t)|X]
\] (24)

218
Finally, (22) and (24) together imply the desired result.

Next, let \( r(t) \) be defined, similarly to \( v(t) \), as:

\[
r(t) = w(t) - B(X, t).\text{Inv}(BB[:J]).W[:J]
\]

where \( W[:J] = [w(t) : t \in J] \) is an \( n \times 1 \) vector. Note that, due to the definitions of \( w \) and \( J, Y[:J] = W[:J] \).

Finally, based on Lemma 3, we can prove the following identification result:

**Theorem 6:** If A5 to A8 are satisfied and the conditional distribution of \( [w(t) : t \in I] \) given \( d \) and \( X \) is known, then \( E[z|l, X] \) is identified.

**Proof:** By the definition of \( v \) and \( r \), and the fact that \( Y[:J] = W[:J] \)

\[
E[r|l, X] = E[v|l, X] + E[z|l, X], \quad \text{(14)}
\]

\[
E[r|0, X] = E[v|0, X] \quad \text{(15)}
\]

By Lemma 3

\[
E[v|d, X] = E[v|X]
\]

Therefore,

\[
E[z|l, X] = E[r|l, X] - E[r|0, X] \quad \text{(25)}
\]

A6 is known and, by hypothesis, \( E[w(t)|d, X] \) is known for all \( t \in I \). Therefore, by the definition of \( r \), the right-hand side of (25) is known. This proves that \( E[z|d = l, X] \) is identified.

When the distribution of \( X \) conditional on \( d \) is known, \( E[z|l] \) can be identified by integrating out \( X \) from \( E[z|d = l, X] \).

**8. Control Function/Eliminant Method**

In this section, we consider a model for explaining the dependence between participation status and post-training earnings. In Section 7, we assumed that participation and post-training earnings were correlated only via their common dependence or pre-training earnings. Conditional on pre-training earnings, they were independent. In Section 8, we assumed that participation was related to earnings only via some unobserved permanent components. Conditional on these components, participation and pre- as well as post-training earnings were assumed to be independent. Under either set of assumptions, we showed that average training effects could be identified. We also show that, to a large extent, these two models are incompatible.
In this section, we consider a more general model that embeds those models analyzed previously. We permit participation status to depend directly on both pre-training earnings and unobserved permanent components. However, the following constraint is imposed on the dependence between participation and pre-training earnings: their direct dependence is limited to occur only for a few years immediately prior to the beginning of the training program. Notice that, despite this assumption, participation and pre-training earnings for all other years will still be indirectly correlated via the permanent components.

More precisely, suppose the observable pre-training period can be decomposed into two parts: $J$ and $K$. $K$ denotes the set of all time periods when participation and earnings are supposed to be directly correlated. $J$, as in the previous section, denotes an exactly $n$-long pre-training period with the limitation in this section of being disjoint of $K$. Needless to say, these definitions of $J$ and $K$ implicitly impose constraints on the length of the pre-training longitudinal information required for identification. So, let $p, f$, and $n$ denote the length of the available pre-training panel, the number of periods of pre-training earnings directly correlated with participation, and the number of the permanent components of earnings, respectively. What we are implicitly requiring is $p > n + f$. This condition replaces A7 ($p > n$) from the previous section. Let $K$-past $(y) = \{y(t) : t \in K\}$. Then consider the following assumption:

**Assumption 9:** $E[d|X, U, \text{all } (y)] = E[d|X, U, K$-past$(y)]$.

Note that A9 is a proper generalization of A4 and A5, as desired.

From A9 and Corollary 1,

$$E[y(t)|d, R, U] = E[y(t)|R, U] \text{ for all } t \in I - K$$

where $R = [X, K$-past $(y)]$.

Next, we make the following assumption similar to A6 and A8.

**Assumption 10:**

(i) $E[y(t)|R, U] = b(R, t) + B(R, t)U$;

(ii) $B$ is a known $n$-row vector function; and

(iii) $BB[\cdot : J]$ is non-singular almost surely, where $BB[\cdot : J]$ is defined as before in equation (20).

Finally, substituting $X$ into $R$ and following exactly the same steps in Lemma 3 and Theorem 6, the following proposition can be proved.

**Theorem 7:** If A9-A10 are satisfied and the conditional distribution of $[w(t) : t$ in $I]$ given $d$ and $X$ is known, then $E(z|1, X]$ is identified.
Whenever the distribution of $X$ conditional on $d$ is known, $E[z|d]$ can also be identified by integrating out $X$ from $E[z|d = 1, X]$.

9. Conclusion

Consider a Roy’s (1951) type of economy in which heterogeneous workers can choose to engage in one of two alternative activities, “a” or “b”. Let income in these two activities be denoted by $Y_a$ and $Y_b$, respectively. Activity “b” is assumed to generate some externalities which are personally valued as $v$ by the worker. Later, for short, we refer to $v$ as the worker’s taste. The utility, $U$, then equals $U(Y_a, 0)$ or $U(Y_a, v)$ depending on which activity the worker engages in. Let $X$ denote human capital. Assume that income is related to human capital as follows: $Y_a = A(X, u)$ and $Y_b = B(X, U)$ where $u$ is a random component that accounts for person-specific productivity shocks. Finally, let $d$ be an indicator function which equals one if a worker is engaged in activity “a” and zero otherwise.

Consider then the question of estimating how much more income those workers who choose to engage in activity “a” earn, on average, compared with what they could have earned in activity ‘b’. In other words, we would like to estimate $E[Y_a - Y_b|d = 1]$. This estimation problem is, in fact, ubiquitous in empirical studies involving human populations. Examples are studies of wage differentials across occupations, sectors of the economy, or geographical regions; evaluations of the impact of unions or training programs on earnings; evaluations of smoking, other habits, or alternative medical procedures on death rates; evaluations of alternative education programs on test scores; and evaluations of the use of seat belts on death rates in car accidents. The common features in all of these studies are the following:

(i) an underlying heterogeneous population,
(ii) a finite set of alternatives, and
(iii) a well-defined and observable outcome.

The question is also always the same: What would be the effect on the average outcome of those individuals who have chosen a given alternative if they were reassigned to another one?

To estimate $E[Y_a - Y_b|d = t]$, we sample $2n$ workers: $n$ from those who choose activity “a” and $n$ from those who choose activity “b”. For each sampled worker, information is obtained on current activity, $d$, income, $Y_a$ or $Y_b$, and human capital, $X$. The average income within each of these samples provides unbiased estimators for $E[Y_a|d = 1]$ and $E[Y_b|d = 0]$, respectively. Therefore, if we try to use the difference between these sample means as an estimator for $E[Y_a - Y_b|d = 1]$, we will end up incurring a bias of the following magnitude

$$E[Y_a - Y_b|d = 1] - [E[Y_a|d = 1] - E[Y_b|d = 0]] = E[Y_b|d = 0] - E[Y_a|d = 1]$$
This bias has been referred to in the literature as “selection bias”. It is a direct consequence of having a heterogeneous population of workers being assigned non-randomly to alternative activities.

The very existence of selection bias should remind us that we simply do not have enough information in our sample to recover $E[Y_a - Y_b|d = 1]$. Some additional prior knowledge is required. This prior information can be of two sorts: (i) knowledge of the functional form of the distribution for some of the random variables in the model; (ii) some additional knowledge about the nature of the decision process used to assign workers to activities. We investigate procedures which are based on this second sort of prior information.

As an illustrative example, consider an expected utility maximizing model with the following two additional features:

(i) at the decision time the worker knows his human capital, $X$, and taste, $v$, but he does not know his productivity shock $u$;

(ii) the taste, $v$, and the productivity shock $u$, are independent.

Under these circumstances, a worker would decide to engage in activity “$a$”, that is, $d = 1$, if and only if

$$L(X, v) = E[U(Y_a, 0)|X, v] - E[U(Y_b, v)|X, v] > 0$$

As a consequence of (i) and (ii)

$$P[d = 1|X, Y_a] = P[L(X, v) > 0|X, A(X, u)]$$

$$= P[L(X, v) > 0|X] = P[d = 1|X]$$

Hence, $d$ and $Y_a$ are conditionally independent given $X$. So, in particular,

$$E[Y_a|X, d = 1] = E[Y_a|X]$$

(1)

Following a similar argument, we can also show that

$$E[Y_b|X, d = 0] = E[Y_b|X]$$

(2)

These equations are the basic ingredients of our study. Selection models in which such relationships hold are referred to as “selection on observables” (SOO) models. The defining characteristic of these models is that while the average income within each group becomes distorted by the selection process, the regression functions remain unchanged. This ought to be the case whenever the selection process depends only upon a set of observed variables and pure noise, like $X$ and $v$ in our example.

The SOO model is clearly a very special kind of selection model. In the context of our example, if either the productivity shock, $u$, were observed at the decision time.
time or \( u \) and \( v \) were correlated, then equations (1) and (2) would not be satisfied and, consequently, the model could not be appropriately classified as an SOO model.

So far, the econometric literature has paid very little attention to this class of selection models. This lack of interest probably comes from the fact that when some traditional, but not necessarily adequate, assumptions are evoked, the SOO model appears as unnecessary simplism. However, once such traditional assumptions are removed, the SOO model assumes paramount significance if it does not become the only alternative available. In this sense, the SOO model is a complementary tool which should be more systematically studied and used by econometricians.

Selection models are traditionally based on the following assumptions:

(a) \( L(X, v) = X.c + V \) for some vector of constants \( c \);
(b) \( v \) and \( X \) are independent;
(c) \( A(X, u) = X.b + u \) for some vector of constants \( b \);
(d) \( E[u|X] = 0 \);
(e) \( B(X, u) - A(X, u) = a \) for some constant \( a \);

Moreover, in order to obtain identification or to simplify the statistical procedures, the following additional assumptions are frequently evoked:

(g) some components of \( b \) are assumed to be zero;
(h) \( E[u|v] = m(v) \) is assumed to be a known function; typically it is assumed to be a linear function;
(i) the distribution of \( v \) is assumed to belong to a known parametric family.

It is extremely important to notice that NONE of these assumptions, (a)-(i), is required when an SOO model is adopted. Moreover, in the case of choice-based samples where \( E[d] \) is unknown, this kind of model is in fact the ONLY FEASIBLE alternative if an investigator does not want to impose functional form or distribution assumptions on the data.

Moreover, the SOO model is more flexible than one may expect. Consider the following example in our Roy’s kind of economy. Assume now that although the productivity shock remains unknown, it follows a Markov process. So, a rational worker would condition his current activity choice on the realization of his productivity shock in the previous period, \( u_{t-1} \). So, now \( d = l \) if and only if

\[
L(X, v, u_1) = E[U(Y_a, 0)|X, v, u_1] - E[U(Y_b, v)|X, v, u_1] > 0
\]
Following similar arguments used before, we can show that

\[ E[Y_a|X, y - 1, d = 1] = E[Y_a|X, y - 1] \]

Therefore, as long as we have longitudinal information on earnings, this case can be treated as a selection on observables – \((X, y - 1)\) model, even though the selection process seems to depend on an unobservable, \(u_{-1}\). Likewise, the traditional transitory-permanent component model can also be formulated as an SOO model, at least as long as the worker is assumed to know his permanent component. If he is learning about this component, however, then the adequate selection model surely lies outside the family of SOO models.

A deep look at identification in SOO models is the object of this article. We demonstrate the flexibility of SOO models by deriving a variety of sufficient conditions for identification. In particular, we show that when adequate use is made of longitudinal information, SOO models become surprisingly flexible. Moreover, these models are able to deliver identification without any subsidiary functional-form assumption, even if the econometrician has access only to choice-based samples and \(E[d]\) is unknown. This extremely useful property does not seem to be shared by any other kind of selection model. The identification conditions we derive essentially put boundaries on the difference between the information available to the decision maker at the decision time and the information available to the econometrician. More specifically, we demonstrate that if they agree on their predictions about future outcomes, then the model would be identified.

References


