Minimal and Maximal Just-Identified Assumptions in Nonparametric Selection Models*

Ricardo Paes de Barros**

Abstract

In this paper, we show minimal just-identified restrictions for a class of selection models. We deepen our knowledge about selection on observables in three directions: First, we show that the model proposed by Rosenbaum and Rubin is just-identified. Second, we show that although their assumptions are not minimal, the weak selection on observables (WSO) hypothesis we introduce does, in fact, generate a set of minimal just-identified assumptions. Finally, we show that the WSO hypothesis has an immediate and sensible economic interpretation. Specifically, we show that this hypothesis can be interpreted as bounds on the information available to the analyst. Basically, identification is achieved as long as the analyst's information set is rich enough so that he is able to predict migration decision at the decision time. As a consequence, the identification conditions become weaker as the information which is assumed to be available to the analyst becomes larger. This approach certainly has a large tradition in econometrics (Zellner et al., 1966). In the context of selection models, similar identification strategies have been followed by Goldberger (1972a) and Heckman and Robb (1985, Sec. 5.4.3).

Keywords: Selection Models, Identification.

JEL Codes: C14, C21, C31.

*Submitted in December 2011. Revised in December 2011. I am grateful to Luiz Araújo, Stephen Cameron, James Heckman, Marcelo Dabos, Bo Honore, Joseph Hotz, and Tom Mroz for very helpful discussions. This research was supported by National Science Foundation grant SES-87-10145 (February 1991).

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

1. Introduction

Consider the question of estimating how much income, on average, workers who choose to migrate earn compared with what they could have earned if they had decided not to migrate. Estimation problems of this kind are, 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 the impact of smoking, other habits, or alternative education programs on test scores. 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 always the same: What would be the effect on the average outcome of those individuals who were allocated to a given alternative if they were reassigned to another one?

Let y and v denote, in a migration study, earnings in the places of origin and destination, respectively. Moreover; let d be a migration indicator so the $d = 1$ if the agent has decided to migrate and $d = 0$ otherwise. The average gain from migration among those who decided to migrate is then given by $E[v - y \mid d = 1]$. To estimate $E[v - y \mid d = 1]$ we may, for example, collect data on earnings for n migrants, at the place of destination, and for n non-migrants, at the place of origin. The average earnings within each of these samples would provide unbiased estimators for $E[v \mid d = 1]$ and $E[y \mid d = 1]$, respectively. Attempts to use the differences between these sample means as an estimator of $E[v - y \mid d = 1]$ would end up incurring a bias of the following magnitude

$$\begin{aligned}[E[v \mid d = 1] - E[y \mid d = 0]] - E[v - y \quad \mid \quad d = 1] &= E[y \mid d = 1] - E[y \mid d = 0] \\ &= \frac{1}{p(1-p)} Cov(y, d),\end{aligned}$$

where $p \equiv P[d = 1]$.[1] This bias has been referred to in the literature as *"selection bias"* It is a direct consequence of having a heterogeneous population being assigned nonrandomly to alternatives. The selection bias just reflects the fact that,

[1]The last equality is interesting per se. It implies, at least as long as $0 < p < 1$, that y is *mean*-independent of d ($E[y \mid d] = E[y]$) if y and d are uncorrelated. Of course, this is true only because d is a binary random variable. A related interesting result is that d is *mean*-independent of y ($E[d \mid y] = E[d]$) if and only if d and y are actually independent. This follows from the fact that for binary random variables $E[d \mid y] = P[d = 1 \mid y]$.

due to nonrandom allocation, migrants and non-migrants would have different average earnings even if migration had no effect whatsoever on earnings. Since Roy (1951), this question has been intensively studied in economics.

The very existence of selection bias should remind us that we simply do not have enough information in our sample to recover $E[v - y \mid d = 1]$. Some kind of additional prior knowledge is required. Heckman and Robb (1985) extensively studied how identification can be obtained by imposing several alternative prior restrictions. In particular, if the regression functions $E[y \mid X]$ and $E[v - y \mid d = 1, X]$ are constrained to be linear, but $P[d = 1 \mid X]$ is nonlinear, then $E[v - y \mid d = 1]$ is identified. In this case, an instrumental variable procedure could be used to actually estimate $E[v - y \mid d = 1]$ $\sqrt{n}-$consistently. No exclusion restriction is required; nonlinear functions of X provide the required instruments.

Rosenbaum and Rubin (1983), on the other hand, have shown how $E[v-y \mid d = 1]$ can be estimated without having to impose restrictions on how the conditional distributions of y and d are related to the covariates. They obtain identification by requiring conditional independence between y and d. Their identification hypothesis is referred to in the econometric literature as the Selection on Observables Hypothesis. When this hypothesis is satisfied $E[v - y \mid d = 1]$ can be consistently estimated by a matching procedure or more generally by the class of $D-$estimators introduced in Barros (1987).

In this paper, selection models in which the dependence of the conditional distributions of y and d on the covariates is not constrained to follow any prespecified functional form are referred to as Nonparametric Selection Models. We show how minimal just-identified restrictions can be obtained in such class of selection models.

We deepen our knowledge about selection on observables in three directions: First, we show that the model proposed by Rosembaum and Rubin is just-identified. Second, we show that although their assumptions are not minimal, the weak selection on observables-(WSO) hypothesis we introduce does, in fact, generate a set of minimal just-identified assumptions. Finally, we show that the WSO hypothesis has an immediate and sensible economic interpretation. Specifically, we show that this hypothesis can be interpreted as bounds on the information available to the analyst. Basically, identification is achieved as long as the analyst's information set is rich enough so that he is able to predict migration decision, at the decision time. As a consequence, the identification conditions become weaker as the information which is assumed to be available to the analyst becomes larger. This approach certainly has a large tradition in econometrics (Zellner et al., 1966). In the context of selection models, similar identification strategies have been followed by Goldberger (1972b) and Heckman and Robb (1985, Sec. 5.4.3).

It is worthwhile at this point to discuss the importance of searching for minimal just-identified assumptions. Notice that whenever a model is actually estimated, a host of auxiliary assumptions are usually made in order to achieve empirical

tractability. Since our objective is to describe the fundamental channels through which the available information can be used to obtain identification, superimposing auxiliary assumptions would only blur the nature of the basic identification channels. Besides, auxiliary assumptions are in general method-oriented. Thus, imposing them always makes it more difficult to determine what the methods which rest on the same set of underlying identifying restrictions are. The importance of studying identification separately from estimation is also emphasized in Manski (1988). With respect to minimality, it seems to be true that minimal restrictions which are, at the same time, easy to motivate are difficult to encounter. In general, well-motivated identifying restrictions are not minimal. However, by extracting a minimal set from such restrictions we can not only better understand how identification is obtained, but also usually generate alternative sets of restrictions which are also well-motivated and sufficient for identification.

This paper is organized as follows: Section 2 describes the basic elements of the model. Section 3 illustrates the general model in the context of a migration study. Section 4 formally defines the identification concepts we use. Section 5 studies the Selection on Observables Hypothesis.

2. Model, Notation, and Objective

Consider a universe of subjects Ω. Assume there are two states subjects Ω can occupy: 0 and 1. Each subject $\omega \in \Omega$ has two potential outcomes, $y(\omega)$ and $v(\omega)$, depending on the state he/she occupies. $y(\omega)$ will be the outcome if ω is in state 0 and $v(\omega)$ will be the outcome if ω is in state 1. In addition, let the state subject ω is currently in be indicated by the dichotomous variable $d(\omega)$. The intervention consists in moving subjects from state 0 to state 1. Hence, the impact of the intervention on subject ω is defined as $v(\omega) - y(\omega)$.

Let $\mathcal{L}(v-y \mid d=1, X_t)$ denote the probability distribution of $v-y$ conditional on $d=1$ and X_t, where X_t is a vector of covariates pre-specified and observed by the analyst. The purpose of this paper is to investigate conditions sufficient to identify the mean of this conditional distribution, $E[v-y \mid d=1, X_t]$.

Identification depends crucially on the information available to the econometrician. We assume that besides being able to observe d and a set of covariates, X_e, the econometrician can observe y if $d=0$ and v if $d=1$. More precisely, we assume that the econometrician's information consists of

$$\mathcal{L}(y, X_e \mid d=0), \quad \mathcal{L}(v, X_e \mid d=1), \quad \text{and } \mathcal{L}(d)$$

where $\mathcal{L}(y, X_e \mid d=0)$ denotes the joint distribution of y and X_e given $d=0$; $\mathcal{L}(v, X_e \mid d=1)$ denotes the joint distribution of v and X_e given $d=1$; and $\mathcal{L}(d)$ denotes the distribution of d. Actually, all results in the paper still hold if $\mathcal{L}(d)$ is not part of the econometrician's information set. Knowledge about $\mathcal{L}(d)$ is likely to be missing in choice-based samples. Since our identification results do not depend on knowledge of $\mathcal{L}(d)$, they are robust to choice-based sampling.

3. Application: Wage Gain from Migration

In this application, y and v denote earnings in places of origin and destination, respectively. The intervention is a migration process. Hence, $d = 1$ if the agent has decided to migrate and $d = 0$ otherwise. Let X_t be a set of demographic characteristics. Then, the goal is to identify $E[v - y \mid d = 1, X_t]$ - the average gain from migration among those who decided to migrate, $d = 1$, by demographic group.

The identification conditions we investigate in this paper are based on restrictions upon the *"relevant"* information available to the agents involved in the migration decision. To simplify the analysis, we assume there exists a unique decision-maker. Accordingly, we model the migration decision as follows: Let $U(0)$ denote the utility of the decision-maker when the worker does not migrate, likewise, let $U(1)$ denote the corresponding utility when the worker does migrate. Let X_d be the decision-maker's *"relevant"* information at the decision time. Thus, we assume the worker migrates, $d = 1$, if and only if

$$E[U(1) - U(0) \mid X_d] > 0.$$

It is worthwhile, at this point, to discuss a caveat about the decision-maker's *"relevant"* information, X_d. Consider a case in which a random variable ψ belongs to the decision-maker's information set but is orthogonal to $U(1)-U(0)$ conditional on X_d. So that,

$$E[U(1) - U(0) \mid X_d, \psi] = E[U(1) - U(0) \mid X_d].$$

In this case, ψ does not belong to the decision-maker's *"relevant"* information set. In summary, X_d should only include only the information available to the decision-maker, at the decision time, which is actually "relevant" to predict the sign of $U(1) - U(0)$.[2]

4. Identification

4.1 General identification

Let M be the Model space. Objects in M are alternative "complete" descriptions of a model structure. In identification studies, the analysis centers on two attributes of a model. These attributes are defined by functions $g : M \longrightarrow T$ and $h : M \longrightarrow S$, where S is called the *Source* space and T is the *Target* space. For any specific model structure, m, $h(m) = \{s\} \subset S$ represents those characteristics

[2]Of course, in order to establish whether or not a variable is helpful in predicting the sign of $U(1) - U(0)$, one ought to be more specific about (i) who is the decision-maker and (ii) how the migration decision is going to affect his/her utility,

of the model which can be directly observed;[3] whereas $g(m) = \{t\} \subset T$ denotes those characteristics we would like to identify.

The identification question is to establish whether objects in space T can be uniquely recovered from objects in space S. By construction, objects in these two spaces are related via the correspondence, $f \equiv g \circ h^{-1}$, $f : S \longrightarrow T$. The identification problem arises, precisely, because, for some s in S, $f(s)$ may have more than one element in T.

Definition 1: g is underidentified from h when, for at least one s in S, $f(s)$ has more than one element in T.

An important extreme case of underidentification occurs when $f(s) = T$ for all s in S. In this case, we say that g is *completely unidentified* from h.

Definition 2: g is *completely unidentified* from h when, for all s in S, $f(s) = T$.

To achieve identification, we have to rely upon certain prior restrictions, R. The role of these restrictions is to modify the correspondence f to f^R, in such a way that, for any s in S, $f^R(s)$ does have *at most* one element in T.

There are at least three natural ways of introducing these restrictions which are actually equivalent. One possibility would be to assume that a second characteristic of the model, $i : M \longrightarrow F$, can also be observed and to define $f^R(s) = g(h^{-1}(s) \cap i^{-1}(f))$, where $f \in F$ is the value of the characteristic actually observed.

A second option would be to impose a set of restrictions, R, directly on the *Target* space. In this case, we would define $f^R(s) = f(s) \cap R$. Notice that this is equivalent to defining $f^R(s) = g(h^{-1}(s) \cap g^{-1}(T))$.

A third possibility, which we follow in this paper, is to have the restrictions be imposed on the *Model* space. Restrictions on M, as opposed to restrictions on T, are commonly easier to interpret and motivate. So, let $R \subseteq M$. Define that, for each s in S, $f^R(s)$ has at most one element in T, and that R forms a set of identifying restrictions. Formally,

Definition 3: $R \subseteq M$ forms a set of sufficient restrictions to identify g from h when, for any s in S, $f^R(s) \equiv g(h^{-1}(s) \cap R)$ contains *at most* one element in T.

If R forms a set of sufficient restrictions to identify g from h and $R^* \subset R$, then R^* also forms a set of sufficient restrictions to identify g from h. Whenever $R^* \subset R$, we say that R^* is stronger than R.

When R is too restrictive, it may be the case that, for some values of s, $f^R(s) = \varnothing$. That is, R may be incompatible with some values for s. In this case, R is said to form a set of overidentifying restrictions. Otherwise, if $f^R(s)$ contains exactly one element T for each s in S, then R is said to form a set of just-identifying restrictions.

[3]We will be working with correspondences as well as with functions. To keep a common notation, we use $h(m)$ as a short notation for $h(\{m\})$.

Definition 4: $R \subseteq M$ forms a set of just-identified restrictions to identify g from h when, for any s in S, $f^R(s) \equiv g(h^{-1}(s) \cap R)$ contains *one and only one* element in T.

Next, we introduce the concepts of minimal and maximal just-identified restrictions.

Definition 5: R forms a set of *minimal just-identified restrictions* whenever (i) R forms a set of just-identified restrictions and (ii) there exists no other set of identifying restrictions which strictly contains R, i.e., all other sets of identifying restrictions are stronger than R.

Definition 6: R forms a set of *maximal just-identified restrictions* whenever (i) R forms a set of just-identified restrictions and (ii) there exists no other set of just-identified restrictions which is strictly contained in R. In other words, R forms a set of *maximal just-identified restrictions* whenever R is the strongest set of just-identified restrictions.

An important property of a set of restrictions R is to which extent R actually imposes restrictions on the Source and Target spaces. More generally, let $i : M \longrightarrow F$ be a function defining a characteristic of the model structure. We then define

Definition 7: R is $i-$empty when $i(R) = i(M)$. In particular, when $h(R) = h(M)$, we say that R is Source-empty; whereas when $g(R) = g(M)$, we say that R is Target-empty.

4.2 Useful lemmas

Lemma 1: g is completely unidentified from h if and only if for all pair (m^1, m^2) in $M \times M$ there exists a model $m(m^1, m^2)$ in M such that $h(m(m^1, m^2)) = h(m^1)$ and $g(m(m^1, m^2)) = g(m^2)$.

Proof: Assume that $f(s) = T$ for all s in M.

Lemma 2: R forms a set of just-identified restrictions if and only if R forms a set of identifying restrictions and $h(R) = S$ (Source-empty).

Proof: Note that $f^R(s) = \varnothing$ if and only if $h^{-1}(s) \cap R = \varnothing$, since, by assumption, $g(m) \neq \varnothing$ for all m in M. Moreover, $h^{-1}(s) \cap R = \varnothing$ for all s in S if and only if $h(R) = S$. Consequently, a set of identifying restrictions forms a set of just-identified restrictions if and only if $h(R) = S$.

Corollary 1: R being Source-empty is a necessary condition for R to form a set of

just-identified restrictions.

Lemma 3: R is Source-empty if and only if for all m in M there exists an $m^*(m)$ in R such that $h(m) \equiv h(m^*(m))$.

To establish identification using Definition 1 directly would require that h^{-1} be specified. In general, it is too tedious to specify h^{-1}. The following lemma permits us to obtain identification results without having to specify h^{-1}.

Lemma 4: If there exists a correspondence $H : S \longrightarrow T$ such that, for all $m \in R$, $g(m) = H(h(m))$ then R forms a set of sufficient restrictions to identify g from h.

Proof: Let $s \in S$. If $h^{-1}(s) \cap R = \varnothing$ then $f^R(s) = \varnothing$. If $h^{-1}(s) \cap R \neq \varnothing$ then all $m \in h^{-1}(s) \cap R$ satisfies

$$g(m) = H(h(m)) = H(s).$$

The first equality holds because $m \in h^{-1}(s)$. Hence, $H(s) = g(h^{-1}(s) \cap R) \equiv f^R(s)$. Let $\overline{m}$ be an element in $h^{-1}(s) \cap R$. Notice that $g(\overline{m}) = H(s) = f^R(s)$. Since g is a function, $f^R(s)$ has exactly one element in T.

Lemma 5: If R forms a set of sufficient restrictions to identify g from h, then for all $m \in R$, $g(m) = f^R(h(m))$.

Next, let $T_1 = g_1(M)$ and $T_2 = g_2(M)$ be two Target spaces. The following lemma is then very useful to obtain sufficient identification conditions for g_2 once it has already been proved that g_1 is identified.

Lemma 6: If R forms a set of sufficient restrictions to identify g_1 from h and there exists a correspondence $G : T_1 \longrightarrow T_2$ such that, for all $m \in R$, $g_2(m) = G(g_1(m))$, then R is also sufficient to identify g_2 from h.

Proof: Take any $m \in R$. Then $g_2(m) = G(g_1(m))$.

By *Lemma 5*, for any $m \in R$

$$g_1(m) = f_1^R(h(m)).$$

Hence, for any $m \in R$

$$g_2(m) = G \circ f_1^R(h(m)).$$

The result follows from *Lemma 4*.

In summary, to investigate identification, one needs first to specify four entities:

(i) A model Space, M;

(ii) a set of restriction R, $m \in R$;

(iii) a set of observed characteristics defined by the function $h : M \longrightarrow S$; and

(iv) a set of characteristics one wants to identify, which are defined via the function $g : M \longrightarrow T$.

5. Selection On Observables

In this section, we investigate the selection on observables hypothesis considered by Rosenbaum and Rubin (1983). Goldberger (1972b,c), Cain (1975), Barnow et al. (1980) and Heckman and Robb (1985, Sec. 3.5), also investigated this model, but in a parametric context. In parametric or even semiparametric contexts, the selection on observables hypothesis is, in fact, unnecessarily strong. However, as we are going to show, in a nonparametric context the selection on observables hypothesis generates a set of restrictions which is minimal and just-identified. Furthermore, we demonstrate that this hypothesis has a sensible economic interpretation.

5.1 Preliminaries

Let $(\Omega, \mathcal{F})$ be the underlying measurable space. This space is constructed as follows:

$$\Omega = \mathbb{R} \times \mathbb{R} \times \mathbb{X}$$

and

$$\mathcal{F} = \Re \otimes \Re \otimes \aleph$$

where $\Re$ denotes the Borel $\sigma-$algebra in $\mathbb{R}$, the real line, and $(\aleph, \mathbb{X})$ denotes a suitable measurable space where all covariates lie. In the definition of $\mathcal{F}$, the symbol $\otimes$ denotes the $\sigma-$algebra generated by all measurable rectangles. Let $\mathbb{P}$ denote the set of all probability measures on $(\Omega, \mathcal{F})$, i.e., the set of all countably-additive functions from $\mathcal{F}$ onto $[0, 1]$.

Let y, v, and X be random variables (vectors) defined by coordinate functions: $y : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \Re)$, $v : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \Re)$, $X : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \Re)$. Hence, for all $\omega \in \Omega$

$$\omega = (y(\omega), v(\omega), X(\omega)).$$

Each probability measure on $(\Omega, \mathcal{F})$ defines a probability distribution for (y, v, X) and conversely each probability distribution for (y, v, X) defines a probability measure on $(\Omega, \mathcal{F})$. Notice that the measurable space $(\aleph, \mathbb{X})$ is left unspecified. By a suitable choice of $(\aleph, \mathbb{X})$, one can impose restrictions on the set of possible distributions for the covariates. For example, if all covariates were known to be categorical, one could specify $\mathbb{X}$ as a set with a countable or even finite number of elements. On the other hand, the explicit choice of $(\mathbb{R} \times \mathbb{R}, \Re \times \Re)$

as the measurable space for (y, v) implies that all distributions for (y, v) are possible (see Shiryayev [1984, pg. 1581]). In other words, this construction procedure creates a model space in which nothing is pre-assumed about the joint distribution of (y, v).

Let $(\mathbb{X}_e, \aleph_e)$ and $(\mathbb{X}_d, \aleph_d)$ be two measurable spaces and $\mathbb{E} : (\mathbb{X}, \aleph) \longrightarrow (\mathbb{X}_e, \aleph_e)$ and $\mathbb{D} : (\mathbb{X}, \aleph) \longrightarrow (\mathbb{X}_d, \aleph_d)$ be two measurable functions. $X_e = \mathbb{E}(X)$ are the covariates observed by the econometrician. $X_d = \mathbb{D}(X)$ are the covariates observed and relevant to the decision-maker, in the sense that $\sigma(d) \subset \sigma(X_d)$, i.e., there exists a set $\overline{A} \in \aleph_d$ such that

$$d(\omega) = 1 \text{ if } X_d(\omega) \in \overline{A}$$

and

$$d(\omega) = 0 \text{ if } X_d(\omega) \notin \overline{A}.$$

Next, we construct two other sets of covariates: X_c - the set of covariates that will be actually used in the econometric estimations - and X_t - the set of covariates that the analyst wants the estimation results to be conditioned on. To introduce these covariates, let $(\mathbb{X}_c, \aleph_c)$ and $(\mathbb{X}_t, \aleph_t)$ be two measurable spaces and $\mathbb{C} : (\mathbb{X}_e, \aleph_e) \longrightarrow (\mathbb{X}_c, \aleph_c)$ and $\mathbb{T} : (\mathbb{X}_e, \aleph_e) \longrightarrow (\mathbb{X}_t, \aleph_t)$ be two measurable functions. Then define $X_c = \mathbb{C}(X_e)$ and $X_t = \mathbb{T}(X_e)$. The construction ensures that $\sigma(X_c) \subset \sigma(X_e)$ and $\sigma(X_t) \subset \sigma(X_e)$, X_c and X_t belong to the econometrician's information set.

Let Z be a random variable (vector) defined in $(\Omega, \mathcal{F})$ and let $\sigma(Z)$ be the $\sigma-$algebra generated by Z. Then $\sigma(Z) \subseteq \mathcal{F}$. For each $\wp \in \mathbb{P}$, $\mathcal{L}(Z)(\wp)$ denotes the restriction of $\wp$ to $\sigma(Z)$, i.e., $\mathcal{L}(Z)(\wp)$ is a probability measure on $(\Omega, \sigma(Z))$ with

$$\mathcal{L}(Z)(\wp) = \wp(A)$$

for all $A \in \sigma(Z)$. We refer to $\mathcal{L}(Z)(\wp)$ as the probability distribution of Z induced by $\wp$.

Each probability measure $\wp \in \mathbb{P}$ induces a probability distribution for $(y, v, X, X_e, X_d, X_c, X_t, d)$. Moreover, notice that, since

$$\sigma(X_e, X_d, X_c, X_t, d) \subseteq \sigma(X),$$

for each $\wp \in \mathbb{P}$, $\mathcal{L}(X_e, X_d, X_c, X_t, d)(\wp)$ can be obtained by restricting $\mathcal{L}(Z)(\wp)$ to $\sigma(X_e, X_d, X_c, X_t, d)$.

5.2 Model space

We could use $\mathbb{P}$ as our Model space. However, there is a subset of elements of $\mathbb{P}$ which is pathological, in the sense that there is no hope for the identification procedures which we are going to investigate to work if members are allowed in this subset. This occurs in most identification studies. For example, that is the

case of models subject to multicollinearity in the context of linear projection.[4] To achieve identification, such pathological cases must be excluded. However, as in the multicollinearity example, such exclusions usually impose restrictions on the *Source* space. In such circumstances, no set of identifying restrictions can exclude the pathological cases and be just-identified at the same time. To overcome this difficulty, we have to exclude these pathological cases from the *Model* space. Restrictions used to exclude cases from the *Model* space are referred to as preliminary restrictions.

We begin with the following two preliminary restrictions on $\mathbb{P}$.

(R1) $E|y|(\wp) < \infty and E|v|(\wp) < \infty$

(R2) $0 < \wp[d = 1] < 1$

Remark 1: Since$|v - y| \leq |v| + |y|$, R1 implies that $E|v - y|(\wp) < \infty$.

For any probability measure $\wp$ on (Ω, F) which satisfies R2, define, for all $A \in F$,

$$\wp_0[A] \equiv \frac{\wp[A \cap \{d = 0\}]}{\wp[d = 0]}$$

and

$$\wp_1[A] \equiv \frac{\wp[A \cap \{d = 1\}]}{\wp[d = 1]}$$

Whenever $\wp$ satisfies R2, $\wp_0$ and $\wp_1$ are two well-defined probability measures on (Ω, F). Notice that for all $A \in F$, $\wp[A]$ can be obtained from $\wp_0[A]$, $\wp_1[A]$, and $\wp[d = 1]$ via $\wp[A] = \wp_0[A] \cdot [1 - \wp[d = 1]] + \wp_1[A] \cdot \wp[d = 1]$.

Remark 2: R1, R2 imply that $E|y|(\wp_0) < \infty$ and $E|v|(\wp_1) < \infty$; with equivalent results for v and $y - v$.[5]

Let Z be a random variable (vector) in (Ω, F). $\wp$, $\wp_0$, and $\wp_1$ induce three probability distributions for $Z : L(Z)(\wp), L(Z)(\wp_0), L(Z)(\wp_1)$. Since for any $A \in$

[4] If the linear projection coefficients are the target, then in cases of multicollinearity g will not be a well-defined function.

[5] Note that

$$\frac{d\wp_0}{d\wp} = \frac{I[d = 0]}{\wp[d = 0]}$$

and

$$\int |y| \cdot I[d = 0] \cdot \, d\wp \leq \int |y| \cdot d\wp \leq \infty.$$

Hence,

$$\wp[d = 0] \cdot \int |y| \cdot d\wp_0 = \int |y| \cdot I[d = 0] \cdot \, d\wp < \infty$$

$\sigma(Z)$

$$L(Z)(A, \wp_0) = \wp_0(A) = \frac{\wp(A \cap \{d = 0\})}{\wp[d = 0]}$$

and

$$L(Z)(A, \wp_1) = \wp_1(A) = \frac{\wp(A \cap \{d = 1\})}{\wp[d = 1]}$$

We alternatively refer to $L(Z)(\wp_0)$ and $L(Z)(\wp_1)$, respectively, as $L(Z|d = 0)(\wp)$ and $L(Z|d = 1)(\wp)$.

Remark 3: For any random variable (vector) Z in (Ω, F) and $\wp \in \mathbb{P}$ satisfying R1 and R2, *Remark 2* and the Radon-Nikodym theorem imply that $E[y|Z](\wp)$, $E[y|Z](\wp_1)$, and $E[y|Z](\wp_0)$ exist, are Z-measurable, integrable and are uniquely defined almost surely $L(Z)(\wp)$, $L(Z)(\wp_1)$, $L(Z)(\wp_0)$, and respectively. Equivalent results hold for v and $v - y$.

Let $\delta(\wp)$, $\delta_0(\wp)$, and $\delta_1(\wp)$ be the distributions for X_c induced respectively by $\wp$, $\wp_0$, and $\wp_1$, i.e., $\delta(\wp) \equiv L(X_c)(\wp)$, $\delta_0(\wp) \equiv L(X_c|d = 0)(\wp)$ and $\delta_1(\wp) \equiv L(X_c|d = 1)(\wp)$. We define the concept of regular vector of covariates with respect to $\wp$ as follows:

Definition 8: Whenever $\wp$ satifies R2, we say that X_c is a *regular* vector of covariates with respect to $\wp$ when $\delta_0(\wp)$ and $\delta_1(\wp)$ are absolutely continuous with respect to each other.

Next, we consider an immediate but nonetheless important implication of X_c being regular.

Lemma 7: If $\wp$ satisfies R2, then X_c is a vector of *regular* covariates with respect to $\wp$ if and only if $\delta(\wp)$, $\delta_0(\wp)$, and $\delta_1(\wp)$ are all absolutely continuous with respect to each other.

Proof : This result follows immediately from the fact that for all $\wp$ that satisfies R2, $\delta(\wp) = \delta_0(\wp) \cdot \wp[d = 0] + \delta_1(\wp) \cdot \wp[d = 1]$.

Remark 4: In the following subsections, some statements hold almost surely with respect to δ_0 whereas others hold almost surely with respect to δ_1. The relevance of *Lemma 4* is to ensure that if X_c is regular then all statements also hold almost surely with respect to δ.

The next lemma provides an alternative characterization for a regular vector of covariates. That is exactly the characterization used by Rosenbaum and Rubin (1983). *Lemma 9* describes an important case where X_c is not regular.

Lemma 8: If $\wp$ satisfies R2, then X_c is a vector of regular covariates with respect to $\wp$ if and only if $0 < \wp[d = 1|X_c] < 1$ almost surely with respect to $\delta(\wp)$.

Proof: Take any $A \in \sigma(X_c)$ such that $\wp[A] > 0$. If X_c is a regular vector of covariates, then by *Lemma 7*, $\wp_1[A] > 0$ and $\wp_0[A] > 0$. Notice that, whenever $\wp[A] > 0$ and R2 is satisfied, $\wp_1[A] > 0$ and $\wp_0[A] > 0$ if and only if

$$0 < \int_A \wp[d = 1|X_c](\wp) \cdot d\delta(\wp) = \wp[A \cap \{d = 1\}] < 1.$$

Lemma 9: If d is measurable with respect to $\sigma(X_c)$, then X_c is not a vector of regular covariance with respect to any $\wp \in \mathbb{P}$.

Proof: If d is measurable with respect to $\sigma(X_c)$, then, for all $\wp$ in $\mathbb{P}$, $E[d|X_c](\wp) = d$. But $E[d|X_c](\wp) = \wp[d = 1|X_c]$. So, $\wp[d = 1|X_c] = d \in \{0, 1\}$. It follows from *Lemma 8* that X_c is not a regular variable of covariates with respect to $\wp$.

Consider our migration example. If the decision-maker's *"relevant"* information set is contained in X_c, i.e., $\sigma(X_d) \subset \sigma(X_c)$, then d would be measurable $\sigma(X_c)$ and consequently X_c would not be regular. Since, by construction, $\sigma(X_c) \subseteq \sigma(X_e), \sigma(X_d) \subseteq \sigma(X_c)$ implies that $\sigma(X_d) \subseteq \sigma(X_e)$. Therefore, in this case, the econometrician's information set is larger than the decision-maker's information set. Identification cannot be obtained because the econometrician knows too much! One way to avoid this case is to assume that X_d includes at least a random variable ε which is independent of X_c.

Whenever X_c is not regular, some sort of extrapolation would be required. In a nonparametric environment, there is simply not enough information to solve problems of this sort. Consequently, we eliminate these cases by constraining the *Model* space to those $\wp \in \mathbb{P}$ relative to which X_c is regular.

(R3) X_c is regular with respect to $\wp$.

For each $\wp \in \mathbb{P}$ that satisfies R1, R2, and R3, define $i_0(\wp) \equiv E[y|d = 0, Xc](\wp)$ and $i_1(\wp) \equiv E[v|d = 1, Xc](\wp)$. We will require the following cross-integrability condition:

(R4) $\int |i_0(\wp)| \cdot d\delta_1(\wp) < \infty$ and $\int |i_1(\wp)| \cdot d\delta_0(\wp) < \infty$.

Remark 5: Notice that $i_0(\wp)$ is, in principle, uniquely defined only almost surely $\delta_0(\wp)$. The fact that $\wp$ satisfies R3, however, ensures that $\delta_0(\wp)$ and $\delta_1(\wp)$ are absolutely continuous with respect to each other. Hence, $i_0(\wp)$ is, in fact,uniquely defined almost surely $\delta_1(\wp)$ and the integration performed in R4 is well-defined. Also note that R1 and R2 imply that $\int |i_0(\wp)| \cdot d\delta_0(\wp) < \infty$ and $\int |i_1(\wp)| \cdot d\delta_1(\wp) < \infty$. That is why we refer to R4 as the cross-integrability condition.

Let $M = \{\wp \in \mathbb{P} : \wp \text{ satisfies } R1, R2, R3, R4\}$.

5.3 Source and target spaces

To investigate identification, we must specify functions h and g which determine which objects are observed and for which objects we want to seek identification. The specification of function h depends on the information available to the econometrician. Consistent with our earlier discussion in Section 2, let

$$h(\wp) \equiv (L(v, X_e|d = 1)(\wp), L(y, X_e|d = 0)(\wp), L(d)(\wp)),$$

for each $\wp \in M$. The fact that all $\wp \in M$ satisfy R2 ensures that the above expressions are well-defined. The Source space S is implicitly defined by $S = h(M)$. Since, for any random variable (vector) Z, to know $L(Z)(\wp)$ means to know $\wp(A)$ for all $A \in \sigma(Z)$. Our choice for h is equivalent to assuming that the econometrician's information is restricted to

(i) $\wp(A|d = 1)$ for all $A \in (v, X_e)$,
(ii) $\wp(A|d = 0)$ for all $A \in (y, X_e)$,
(iii) $\wp(d = 1)$.

The specification of function g depends on the object we want to identify. In this paper, we seek identification conditions for objects defined via

$$g(\wp) \equiv E[v - y|d = 1, X_t](\wp),$$

for each $\wp \in M$. Since each $\wp \in M$ satisfies R1 and R2, by *Remark 3* $g(\wp)$ exists and is uniquely defined as $L(X_t)(\wp 1)$. Hence, for g to actually be a function with domain M, we have to use the convention of treating random variables as equal if they differ only on sets with zero probability.

More specifically, let Z be a random variable (vector) in (Ω, F) and $A(Z)$ be the set of all random variables in (Ω, F) which are measurable with respect to $\sigma(Z)$. For each $\wp \in M$ define $\sim \wp$ by: for any $(f_1, f_2) \in A(Z)^2$, $f_1 \sim \wp f_2$ whenever $\wp(f_1 = f_2) = 1$. Note that $\sim \wp$ defines an equivalence relation in $A(Z)$. Let $Q(Z, \wp)$ be the quotient space consisting of the equivalent classes of $A(Z)$ under $\sim \wp$. Finally let

$$\Im(Z) = \underset{\wp \in M}{U} Q(Z, \wp)$$

Note that $g(\wp) \in Q(X_t, \wp_1)$ and $\wp_1 \in M$. Hence g is a function from M into $\Im = (X_t)$. Let the *Target* space be defined as the image of g, i.e., $T = g(M)$.

Remark 6: It is easy to show that if $f_1 \in Q(Z, \wp)$, $f^- = f_1 - f_2$, and $f^+ = f_1 + f_2$, then $f^- \in Q(Z, \wp)$and $f^+ \in Q(Z, \wp)$.

5.4 Underidentification

In any identification study, it is important first to demonstrate that the characteristic we want to identify is, based only on observed characteristics, underidentified before beginning to search for identification conditions.

In non-parametric selection models, a very strong result related to underidentification holds:

These models are completely underidentified, i.e., for all s in S, $f(s) = T$.

Theorem 1: Relative to M, $g \equiv L(v - y|d = 1, X_t)$ is completely unidentified from $h \equiv (L(v.X_e|d = 1), L(y, X_e|d = 0, L(d))$.

5.5 The Selection On Observables Hypothesiss

In this section, we introduce the strong and the weak selection on observables hypothesis. We also introduce a set of restrictions (R5) which, jointly with the strong selection on observables hypothesis, are proved in the next section to form a set of maximal just-identified restrictions. More importantly, we show that all these restrictions on the model space impose themselves no restrictions on the target space, i.e., they are all *target-empirically empty.*

Definition 9: $\wp \in M$ satisfies the strong selection on observables (SSO) hypothesis when conditional on $X_c, (y, v)$ and d are independent, i.e., for all $A \in \sigma(y, v)$ and $B \in \sigma(d)\wp[A|X_c]\cdot\ \wp[B|X_c]$.

Definition 10: $\wp \in M$ satisfies the weak selection on observables (WSO) hypothesis when conditional on X_c, y is mean independent of d, i.e., $E[y|d, X_c](\wp) = E[y|X_c](\wp)$.

The following alternative characterizations of these two hypotheses are going to be very useful in the following selection.

Lemma 10 : $\wp \in M$ satisfies the strong selection on observables (SSO) hypothesis if and only if for all $A \in \sigma(y, v)$, $\wp_0(A|X_c)$.

Lemma 11 : $\wp \in M$ satisfies the weak selection on observables (WSO) hypothesis if and only if for all $E[y|X_c](\wp_0) = E[y|X_c](\wp_1)$.

Lemma 12: If $\wp \in M$ satisfies the strong selection on observables (SSO) hypothesis, then for all $A \in \sigma(y)$, $\wp_0(A|X_c) = \wp_1(A|X_c)$. It follows from Theorem 34.5 in Billingsley (1979) that $E[y|X_c](\wp_0) = E[y|X_c](\wp_1)$. Therefore, by *Lemma 8*, $\wp$ also satisfies the WSO hypothesis.

(R5) (i) Conditional on $d = 1$ and X_c, y and (v, X) are independent, i.e., for any $A \in \Re$ and $B \in \Re \times \aleph$,
$L(y, v, X|X_c)(\wp_0)(A \times B) = L(v|X_c)(\wp_0)(A)\cdot L(y, X|X_c)(\wp_0)(B)$

(ii) Conditional on $d = 0$ and X_c, v and (y, X) are independent, i.e., for any $A \in \Re$ and $B \in \Re \times \aleph$,
$L(y, v, X|X_c)(\wp_0)(A \times B) = L(v|X_c)(\wp_0)(A)\cdot L(y, X|X_c)(\wp_0)(B)$.

Lemma 13: If $\wp \in M$ and $\wp$ satisfy R5, then y and v are independent conditional on X_c and d.

Theorem 2: Let R be the subset of M formed by all probability measures in M that satisfies the SSO hypothesis and R5, then $g(R) = g(M) \equiv T$ where $g = L(v - y|d = 1, X_c)$.

Corollary 2: Let R be the subset of M formed by all probability measures in M that satisfies the SSO hypothesis, then $g(R) = g(M) \equiv T$ where $g = L(v - y|d = 1, X_c)$.

Corollary 3: Let R be the subset of M formed by all probability measures in M that satisfies the WSO hypothesis, then $g(R) = g(M) \equiv T$ where $g = L(v - y|d = 1, X_c)$.

5.6 Identification results

In this section, we demonstrate that the strong selection on observables hypothesis generates a set of just-identified restrictions whereas the weak selection on observables hypothesis provides a set of minimal just-identified restrictions. We also show that the SSO hypothesis, jointly with R5, generates a set of maximal just-identified restrictions.

Remark 7 : It follows from Remarks 3 and 4 that, for all $\wp \in M$, $i_0(\wp)$, $f(\wp)$, and $g_c(\wp)$ are elements in $Q(X_c, \wp)$. By Remark 5, linear operations with these objects always generate elements in $Q(X_c, \wp)$.

Theorem 3: Relative to M, the weak selection on observables-WSO hypothesis is sufficient to identify $g_c \equiv E[v - y|\ d = 1, X_t]$ from $i \equiv (E[y|d = 0, X_c], E[v|d = 1, X_c])$.

Theorem 4: Relative to M if $\sigma(X_t) \subseteq \sigma(X_c)$, the weak selection on observables-WSO hypothesis is sufficient to identify $g_t \equiv E[v - y|d = 1, X_t]$ from $h \equiv (L(y, X_e|d = 0),\ L(v, X_e|d = 1))$.

Theorem 5: Relative to M if $\sigma(X_t) \subseteq \sigma(X_c)$, R4 and the strong selection on observables-SSO hypothesis form a set of *just-identified* restrictions to identify $g_t \equiv L(v - y|d = 1, X_t)$ from $h \equiv (L(y, X_e|d = 0)\ ,\ L(v, X_e|d = 1)\ ,\ L(d))$.

Corollary 4: Relative to M, if $\sigma(X_t) \subseteq \sigma(X_c)$, R4 and the strong selection on observables-SSO hypothesis form a set of *just-identified* restrictions to identify $g_t \equiv E[v - y|d = 1, X_t]$ from $h \equiv (L(y, X_e|d = 0), L(v, X_e|d = 1), L(d))$.

Theorem 6: Relative to M, if $X_d = (X_e, d)$, then R4 and the SSO hypothesis form a set of maximal just-identified restrictions to identify $g_c \equiv E[v - y|d = 1, X_c]$ from $h \equiv (L(y, X_e|d = 0)\ ,\ L(v, X_e|d = 1)\ ,\ L(d))$.

Theorem 7: Relative to M if $\sigma(X_t) \subseteq \sigma(X_c)$, the WSO hypothesis forms a set of minimal just-identified restrictions to identify $g_c \equiv E[v - y|d = 1, X_c]$ from $h \equiv (L(y, X_e|d = 0)\ ,\ L(v, X_e|d = 1)\ ,\ L(d))$.

5.7 Interpreting the selection on observables hypothesis

In our migration example, the SSO hypothesis implies that workers with identical X_c characteristics have identical migration probability, independently of their potential wages after migration, (y, v).Under this hypothesis, as Rosenbaum and Rubin (1983) have noticed, we can actually identify $E[v-y]$, the overall gain from migration among migrants and non-migrants. Next, consider the restriction

(R5) $\qquad E[y \mid X_d] = E[y \mid X_c]$ with $X_d \supset X_c$.

This restriction states that although the decision-maker has more information than the econometrician, his extra information is useless to predict post-migration latent wages, y. Note that R5 is stronger than the WSO hypothesis, since $\sigma\ (X_d) \supseteq \sigma\ (d, X_c) \supseteq \sigma\ (X_c)$. Nevertheless, due to its finer conditioning set, R5 is not implied by either WSO hypothesis or the SSO hypothesis. In summary,

$$SSO \Rightarrow WSO \Leftarrow R5$$

In conclusion, this section has shown that if X_c is regular, then, in order to identify $E[x \mid d = 1, X_c]$ conditions like R5, the SSO hypothesis, and the WSO hypothesis are sufficient. R5 and the SSO hypothesis are the easier conditions to interpret, while theWSO hypothesis is the minimal condition.

References

Barnow, B. S., Cain, G. C., & Goldberger, A. S. (1980). Issues in the analysis of selectivity bias. In Stromsdorfer, W. E. & Farkas, G., editors, *Evaluation Studies Review Annual*, volume 5, pages 42–59. Sage Publications, Beverly Hills and London.

Barros, R. (1987). *Two essays on the nonparametric estimation of economic models with selectivity using choice-based samples.* PhD thesis, University of Chicago.

Cain, G. G. (1975). Regressions and selection models to improve nonexperimental comparison. In Bennett, C. A. & Lumsdaine, A. A., editors, *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*, pages 297–317. Wiley, New York.

Goldberger, A. S. (1972a). Selection bias in evaluations treatment effects: Some formal illustrations. Discussion Paper 120-71, Institute for Research on Poverty, University of Wisconsin.

Goldberger, A. S. (1972b). Selection bias in evaluations treatment effects: Some formal illustrations. Discussion Paper 120-71, Institute for Research on Poverty, University of Wisconsin.

Goldberger, A. S. (1972c). Selection bias in evaluations treatment effects: The case of interactions. Discussion Paper 129-72, Institute for Research on Poverty, University of Wisconsin.

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

Manski, C. F. (1988). *Identification for Binary Response Models*. JASA.

Rosenbaum, S. E. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Zellner, A., Kmenta, J., & Dreze, J. (1966). Specification and estimation ofcobb-douglas production function models. *Econometrica*, Oct:784–795.

Appendix: Some Proofs

Proof of Theorem 1: By Lemma 1, we only need to show that for all pair $(\wp^1, \wp^2)$ in $M \times M$ there exists a model $\wp(\wp^1, \wp^2)$ in M such that $h(\wp(\wp^1, \wp^2)) = h(\wp^1)$ and $g(\wp(\wp^1, \wp^2)) = h(\wp^2)$. To describe our candidate for $\wp(\wp^1, \wp^2)$, let $\psi(\wp) \equiv L(v|d = 1, Xt)(\wp)$, $\gamma(\wp) \equiv L(v, X|d = 1, Xt)(\wp)$, and $\xi_1(\wp) \equiv L(Xt|d = 1)$. Furthermore, let $*$ denote convolution. For any $A \in \Re$, $B \in \Re$, $C \in \aleph$, $D \in \aleph_c$, define $\wp_{_0}(\wp^1, \wp^2)$, $\wp_1(\wp^1, \wp^2)$ and $\wp(\wp^1, \wp^2)$ as follows:

$$\wp_{_0}(\wp^1, \wp^2)(A \times\ B \times\ C)$$

$$\wp_1(\wp^1, \wp^2)(A \times\ B \times\ C \cap D) = \int_D \{(\psi(\wp^1) * g(\wp^2)(A).\gamma(B \times C, \wp^1)\}.d\xi_1(\wp^1)$$

and

$$\wp(\wp^1, \wp^2)(A \times\ B \times\ C \cap \{d = 0\}) = \wp_0(\wp^1, \wp^2)(A \times\ B \times\ C) \cdot\ \wp^1(d = 0)$$

$$\wp(\wp^1, \wp^2)(A \times\ B \times\ C \cap \{d = 1\}) = \wp_{_1}(\wp^1, \wp^2)(A \times\ B \times\ C) \cdot\ \wp^1(d = 1)$$

Hence,

$$\wp(\wp^1, \wp^2)(A \times B \times C) = \wp_0(\wp^1, \wp^2)(A \times B \times C).\wp^1(d = 0) + \wp_1(\wp^1, \wp^2)(A \times B \times C)\wp^1(d = 1)$$

Proof of Lemma 10: Let $A \in \sigma(y, v)$ and $B \in \sigma(d)$. Note that $\sigma(d) = \{\emptyset, \Omega, \{d = 0\}, \{d = 1\}\}$.

Therefore, $\wp[A \cap B|X_c] = \wp[A|X_c].\wp[B|X_c]$ for all $A \in \sigma(y, v)$ and $B \in \sigma(d)$ if and only if for all $A \in \sigma(y, v)$

$$\wp[A \cap \{d = 1\}|X_c] = \wp[A|X_c].\wp[\{d = 1\}|X_c] \tag{.1}$$

First we are going to prove that (.1) implies that for all $A \in \sigma(y, v)$

$$\wp_0(A|X_c) = \wp_1(A|X_c) = \wp(A|X_c) \tag{.2}$$

To prove that $\wp(A|X_c) = \wp_1(A|X_c)$, it is enough to show that for all $D \in \sigma(X_c)$

$$\int_D \wp(A|X_c) \cdot\ d\wp_1 = \wp_1(A \cap D)$$

Since for all $D \in \sigma(X_c)$,

$$\int_D \frac{\wp(d=1|X_c)}{p(d=1)}\cdot d\wp = \frac{\wp(D\cap\{d=1\}|X_c)}{\wp(d=1)} = \wp_1(D)$$

It follows that

$$\begin{aligned}\int_D \wp(A|X_c)\cdot d\wp_1 &= \int_D \wp(A|X_c)\frac{\wp(d=1|X_c)}{\wp(d=1)}\cdot d\wp \\ &= \int_D \frac{\wp(A\cap\{d=1\}|X_c)}{\wp(d=1)}\, d\wp\end{aligned} \qquad (.3)$$

where the last equality is a consequence of (.1). Hence,

$$\int_D \wp(A|X_c)\cdot d\wp_1 = \frac{\wp(A\cap D\cap\{d=1\})}{\wp(d=1)} = \wp_1(A\cap D)$$

and we have proved that $\wp(A|X_c) = \wp_1(A|X_c)$. That $\wp(A|X_c) = \wp_0(A|X_c)$ can be proved similarly.

Next, we are going to prove that $\wp_1(A|X_c) = \wp_0(A|X_c)$ implies (.1). As an intermediate step, we show that $\wp_1(A|X_c) = \wp_0(A|X_c)$ implies (.2). Notice that

$$\int_D \wp_1(A|X_c)\cdot d\wp_1 = \wp_1(A\cap D) = \frac{\wp(A\cap D\cap\{d=1\})}{\wp(d=1)}$$

$$\int_D \wp_0(A|X_c)\cdot d\wp_0 = \wp_0(A\cap D) = \frac{\wp(A\cap D\cap\{d=0\})}{\wp(d=0)}$$

Using the fact that $\wp_1(A|X_c) = \wp_0(A|X_c)$, we obtain

$$\wp(A\cap D) = \int_D \wp_1(A|X_c)(\alpha\cdot d\wp_1 + (1-\alpha)\cdot d\wp_0)$$

where $\alpha = \wp(d=1)$. Since $\alpha = \alpha\cdot \wp_1 + (1-\alpha)\cdot \wp_0$,

$$\wp(A\cap D) = \int_D \wp_1(A|X_c)\cdot d\wp$$

Hence, we have proved that $\wp(A|X_c) = \wp_1(A|X_c)$. Therefore,

$$\int_D \wp_1(A|X_c)\cdot d\wp_1 = \int_D \wp(A|X_c)\cdot d\wp_1$$

From (.3)

$$\int_D \wp(A|X_c)\cdot d\wp_1 = \int_D \wp(A|X_c).\frac{\wp(d=1|X_c)}{\wp(d=1)}.d\wp$$

Moreover,

$$\begin{aligned}\int_D \wp_1(A|X_c)\cdot\, d\wp_1 &= \wp_1(A\cap D) = \frac{\wp(A\cap D\cap\{d=1\})}{\wp(d=1)}\\ &= \int_D \frac{\wp(A\cap\{d=1\}|X_c)}{\wp(d=1)}\cdot\, d\wp\end{aligned}$$

Therefore $\wp[A\cap\{d=1\}|X_c] = \wp[A|X_c].\wp[\{d=1\}|X_c]$, as we would like to show.

Proof of Lemma 11: The WSO hypothesis implies that for any $D\in\sigma(X_c)$

$$\int_{D\cap\{d=1\}} y\cdot\, d\wp = \int_{D\cap\{d=1\}} E[y|X_c](\wp)\cdot\, d\wp = \int_D E[y|X_c](\wp)\cdot\, I[d=1]\cdot\, d\wp$$

Moreover,

$$\wp_1(D) = \int_D \frac{I[d=1]}{\wp(d=1)}.d\wp$$

Hence,

$$\int_D E[y|X_c](\wp)\cdot\, I[d=1]\cdot\, d\wp = \wp(d=1)\cdot\int_D E[y|X_c](\wp)\cdot\, d\wp_1$$

and

$$\int_{D\cap\{d=1\}} y\cdot\, d\wp = \int_D y\cdot\, I[d=1]\cdot\, d\wp = \wp(d=1)\cdot\int_D y\cdot\, d\wp_1$$

Hence, for all $D\in\sigma(X_c)$

$$\int_D y\cdot\, d\wp_1 = \int_D E[y|X_c](\wp)\cdot\, d\wp_1$$

which implies that $E[y|X_c](\wp) = E[y|X_c](\wp_1)$. Similarly, we can prove that $E[y|X_c](\wp) = E[y|X_c](\wp_0)$. Therefore, we have shown that the WSO hypothesis implies that $E[y|X_c](\wp_0) = E[y|X_c](\wp_1) = E[y|X_c](\wp)$.

The following notation will be useful to prove the remaining theorems:

Notation: For each $\wp\in\mathbb{P}$ define
$f(\wp)\equiv E[y|d=1,X_c](\wp)$,
$g_c(\wp)\equiv E[v-y|d=1,X_c](\wp)$,
$g_t(\wp)\equiv E[v-y|d=1,X_t](\wp)$,
$h(\wp)\equiv L(y,X_e)(\wp)$,

$$\eta(\wp) \equiv L(y, X)(\wp),$$
$$\lambda(\wp) \equiv L(y, X_c)(\wp),$$
$$\mu(\wp) \equiv L(X_c|d = 1, X_t)(\wp)$$

Remark 6: It follows from *Remark 3* and *4* that, for all $\wp \in M$, $i_0(\wp)$, $i_1(\wp)$, $f(\wp)$ and $g_c(\wp)$ are elements in $Q(X_c, \wp)$. By *Remark 5*, linear operations with these objects always generate elements in $Q(X_c, \wp)$.

Proof of Theorem 3: Take any $\wp \in M$, by the linearity of conditional expectations

$$E[v - y|d = 1, X_c](\wp) = E[v|d = 1, X_c](\wp) - E[y|d = 1, X_c](\wp)$$

or equivalently

(1) $g_c(\wp) = i_1(\wp) - f(\wp)$

Next, assume $\wp$ satisfies the WSO hypothesis. Hence,

$$E[y|d = 1, X_c](\wp) = E[y|d = 0, X_c](\wp)$$

or equivalently

(2) $f(\wp) = i_0(\wp)$

Define $H : i(M) - \rightarrow g_c(M)$ such that

$$H(i(\wp)) = H(i_0(\wp), i_1(\wp)) \equiv i_1(\wp) - i_0(\wp)$$

Since, for each $\wp \in M$, $i_0(\wp)$, and $i_1(\wp)$ are elements of $Q(X_c, \wp)$, by Remark 5 $H(h(\wp))$ is in $Q(X_c, \wp)$. Consequently, H is a function from $S = i(M)$ into $\Im = (X_c)$.

It follows from (.1) and (.2) that for all $\wp \in M$ that satisfies the WSO hypothesis

$$g_c(\wp) = i_1(\wp) - f(\wp) = i_1(\wp) - i_0(\wp)$$

Hence, for all $\wp \in M$ that satisfies the WSO hypothesis

$$g_c(\wp) = H(i(\wp))$$

and identification follows from *Lemma 2*.

Proof of Theorem 4: This theorem is proved in five steps:

(a) Without any restriction, i_0 can be identified from h_0.

By *remark 2*, for any $\wp \in M$, $E[|y|](\wp_0) < \infty$. Since $h_0(\wp)$ is the restriction of $\wp_0$ to $\sigma(y, X_e)$, $\sigma(y, X_e) \supseteq \sigma(y)$, and $\sigma(y, X_e) \supseteq \sigma(X_c)$, by lemma A1 there exists a conditional probability of y given X_c associated with $h_0(\wp)$, and

$$E[y|Xc](h_0(\wp)) = E[y|Xc](\wp_0) = E[y|d=0, X_c](\wp) \equiv i_0(\wp)$$

(b) Without any restriction, i_1 can be identified from h_1.

By *Remark 2*, for any $\wp \in M$, $E[|v|](\wp_1) < \infty$. Since $h_1(\wp)$ is the restriction of $\wp_1$ to $\sigma(v, X_e)$, $\sigma(v, X_e) \supseteq \sigma(v)$, and $\sigma(v, X_e) \supseteq \sigma(X_c)$, by Lemma A1, there exists a conditional probability of v given X_c associated with $h_1(\wp)$, and

$$E[v|X_c](h_1(\wp)) = E[v|X_c](\wp_1) = E[v|d=1, X_c](\wp) \equiv i_1(\wp).$$

(c) Without any restriction, μ can be identified from h_1.

Since, for all $\wp \in M$, $h_1(\wp)$ is the restriction of $\wp_1$ to $\sigma(v, X_e)$, $\sigma(v, X_e) \supseteq \sigma(X_c)$, and $\sigma(v, X_e) \supseteq \sigma(X_t)$, by *Lemma A2*, there exists a regular conditional distribution function of X_c given X_t associated with $h_1(\wp)$, and

$$L(X_c|X_t)(h_1(\wp)) = L(X_c|X_t)(\wp_1) = L(X_c|d=1, X_t)(\wp) \equiv \mu(\wp)$$

(d) If $\sigma(X_t) \subseteq \sigma(X_c)$, then g_t can be identified from (g_c, μ).

For each $\wp \in M$, $g_c(\wp)$ is an X_c-measurable function and integrable (see *Remark 3*). From Billingsley (1979, Theorem 34.5)

$$\int g_c(\wp) \cdot d\mu(\wp) = E[g_c(\wp)|X_t](\wp_1).$$

Since $\sigma(X_t) \subseteq \sigma(X_c)$, by the iterated expectation theorem (Billingsley (1979, Theorem 34.4))

$$E[g_c(\wp)|X_t](\wp_1) = g_t(\wp)$$

In other words, for all $\wp \in M$,

$$g_t(\wp) = \int g_c(\wp) \cdot d\mu(\wp).$$

(e) Summing up, For any $\wp \in M$, we have shown that

$$i_0(\wp) = E[y|X_c](h_0(\wp));$$

$$i_1(\wp) = E[v|X_c](h_1(\wp));$$

and

$$\mu(\wp) = L(X_c|X_t)(h_1(\wp)).$$

If $\wp$ satisfies the WSO hypothesis, *Lemma 7* shows that

$$g_c(\wp) = i_1(\wp) - i_0(\wp).$$

By (d), if $\sigma(X_t) \subseteq \sigma(X_c)$,

$$g_t(\wp) = \int g_c(\wp) \cdot d\mu(\wp).$$

Therefore, if $\sigma(X_t) \subseteq \sigma(X_c)$ and $\wp \in M$ and if the WSO hypothesis holds

$$g_t(\wp) = \int \{E[v|X_c](h_1(\wp)) - E[y|Xc](h_0(\wp))\} \cdot dL(X_c|X_t)(h_1(\wp)).$$

Identification then follows from *Lemma 2*.

Proof of Theorem 6: Take any $\wp \in M$ satisfying R4 and the SSO hypothesis. If $\wp$ satisfies the SSO hypothesis, then $\wp$ satisfies the WSO hypothesis. Hence, by *Lemma 8*, R4 and the SSO hypothesis are sufficient to identify

$$g \equiv E[v - y|d = 1, X_t]$$

from

$$h \equiv (L(y, X_e|d = 0), L(v, X_e|d = 1), L(d)).$$

It remains to show that they form a set of just-identified restrictions. By Lemma 1, it is enough to show that for any $s \in S$ there exists a $\wp \in M$ satisfying R4 and the SSO hypothesis with $h(\wp) = s$. Since, by construction, $S = h(M)$, without any loss of generality, let

$$s = (L(y, X_e|d = 0)(\overline{\wp}), L(v, X_e|d = 1)(\overline{\wp}), L(d)(\overline{\wp}))$$

for some $\overline{\wp} \in M$. For any $A \in \Re$, $B \in \Re$,$C \in \aleph$, $D \in \aleph_c$ define $\widetilde{\wp_0}, \widetilde{\wp_1}$, and $\widetilde{\wp}$ as follows:

$$\widetilde{\wp_0}(A \times\ B \times\ C \cap D) = \int_D \{\lambda_1(B, \overline{\wp}) \cdot\ \eta_0(A \times\ C, \overline{\wp})\} \cdot\ d\delta_0(\overline{\wp}),$$

$$\widetilde{\wp_1}(A \times\ B \times\ C \cap D) = \int_D \{\lambda_0(B, \overline{\wp}) \cdot\ \eta_1(A \times\ C, \overline{\wp})\} \cdot\ d\delta_1(\overline{\wp}),$$

and

$$\widetilde{\wp}(A \times\ B \times\ C \cap \{d=0\}) = \widetilde{\wp_0}(A \times\ B \times\ C) \cdot\ \overline{\wp}(d=0)$$
$$\widetilde{\wp}(A \times\ B \times\ C \cap \{d=1\}) = \widetilde{\wp_1}(A \times\ B \times\ C) \cdot\ \overline{\wp}(d=1)$$

Hence,

$$\widetilde{\wp}(A \times\ B \times\ C) = \widetilde{\wp}_0(A \times\ B \times\ C) \cdot\ \overline{\wp}(d=0) + \widetilde{\wp}_1(A \times\ B \times\ C) \cdot\ \overline{\wp}(d=1)$$

Notice that (i) $A = \{(A \times B \times C) : A \in \Re, B \in \Re, \text{and } C \in \aleph\}$ is a semi-ring, (ii) $F = \sigma(A)$, and (iii) $\widetilde{\wp}$ is countably additive on A. Therefore $\widetilde{\wp}$ can be uniquely extended to F. Let $\wp^*$ denote the extension of $\widetilde{\wp}$ to F. Hence, $\wp^* \in \mathbb{P}$ and it can be easily shown that $\wp^*$ satisfies the following properties:

(P1)

$$L(d)(\wp^*) = L(d)(\overline{\wp}),$$

Hence $\wp^*$ satisfies R2 since $\overline{\wp} \in M$.

(P2)

$$\begin{aligned} L(y, X|d &= 0)(\wp^*) = L(y, X|d=0)(\overline{\wp}), \\ L(y, X|d &= 1)(\wp^*) = L(v, X|d=1)(\overline{\wp}) \end{aligned}$$

It follows from (P2) that

$$\begin{aligned} L(X_c|d &= 0)(\wp^*) = L(X_c|d=0)(\overline{\wp}), \\ L(X_c|d &= 1)(\wp^*) = L(X_c|d=1)(\overline{\wp}). \end{aligned}$$

Therefore, $\wp^*$ satisfies R3 since $\overline{\wp} \in M$.

(P3)

$$\begin{aligned} L(y, v, X|d &= 0, X_c)(\wp^*) = L(v|d=1, X_c)(\overline{\wp}) \cdot\ L(y, X|d=0, X_c)(\overline{\wp}), \\ L(y, v, X|d &= 1, X_c)(\wp^*) = L(y|d=0, X_c)(\overline{\wp}) \cdot\ L(v, X|d=1, X_c)(\overline{\wp}), \end{aligned}$$

It follows from P3 that

(1)

$$\begin{aligned} L(y, X|d &= 0, X_c)(\wp^*) = L(y, X|d=0, X_c)(\overline{\wp}) \\ L(v, X|d &= 1, X_c)(\wp^*) = L(v, X|d=1, X_c)(\overline{\wp}) \\ L(v|d &= 0, X_c)(\wp^*) = L(v|d=1, X_c)(\overline{\wp}) \end{aligned}$$

(2)

$$L(y|d=1,X_c)(\wp^*) = L(y|d=0,X_c)(\overline{\wp}),$$

Hence,

$$L(y,v,X|d=0,X_c)(\wp^*) = L(v|d=0,X_c)(\wp^*)\cdot\ L(y,X|d=0,X_c)(\wp^*)$$

and

$$L(y,v,X|d=1,X_c)(\wp^*) = L(y|d=1,X_c)(\wp^*)\cdot\ L(v,X|d=1,X_c)(\wp^*).$$

Hence, $\wp^*$ satisfies R4. From (1) and (2), it follows that

$$L(y|d=0,X_c)(\wp^*) = L(y|d=1,X_c)(\wp^*).$$

Hence, $\wp^*$ satisfies the SSO hypothesis. Moreover,

$$L(y,X|d=0)(\wp^*) = L(y,X|d=0)(\overline{\wp}),$$
$$L(v,X|d=1)(\wp^*) = L(v,X|d=1)(\overline{\wp}).$$

Hence, in particular,

$$L(y,X_c|d=0)(\wp^*) = L(y,X_c|d=0)(\overline{\wp}),$$
$$L(v,X_c|d=1)(\wp^*) = L(v,X_c|d=1)(\overline{\wp}).$$

Therefore, $h(\wp^*) = h(\overline{\wp}) = s$. In summary, we have shown that $\wp^* \in M$ satisfies R4 and the SSO hypothesis and $h(\wp^*) \equiv s$. So, by Lemma 1, R4 and the SSO hypothesis form a set of just-identified restrictions.

Proof of Theorem 8: Let

$$R^W = \{\wp \in M : \wp \text{ satisfies theWSO hypothesis}\},$$

and

$$R^S = \{\wp \in M : \wp \text{ satisfies the SSO hypothesis}\},$$

By *Lemma 8*, R^W is a set of identifying restriction. By *Lemma 9*, R^S is a set of just-identified restrictions. Since $R^W \supseteq R^S$, by *Lemma 1*, R^W is also a set of just-identified restrictions.

It remains to be shown that R^W is minimal. Let R^* be an alternative set of identifying restrictions such that $R^* \supseteq R^W$. Take any $\wp^* \in R^*$. Since R^W is just-identified, there exists a $\wp \in R^W$ such that

$$h(\wp) = h(\wp^*)$$

So, in particular,
(1) $E[y|d=0, X_c](\wp) = E[y|d=0, X_c](\wp^*)$,
(2) $E[v|d=1, X_c](\wp) = E[v|d=1, X_c](\wp^*)$.
Since $R^* \supseteq R^W$,

$$g_c(\wp) = g_c(\wp^*).$$

Hence,

$$\begin{aligned} &E[v|d=1, X_c](\wp) - E[y|d=0, X_c](\wp) \\ &= E[v|d=1, X_c](\wp^*) - E[y|d=0, X_c](\wp^*) - \{E[y|d=1, X_c](\wp^*) - E[v|d=0, X_c](\wp^*)\} \end{aligned}$$

Using (1) and (2) we obtain

$$E[y|d=1, X_c](\wp^*) = E[y|d=0, X_c](\wp^*).$$

Hence $\wp \in R^W$ and $R^* = R^W$, as we would like to show.