published in M. Lechner, F.Pfeiffer (Eds.), Econometric Evaluation of Labour Market Policies, Heidelberg: Physica, 2001, 43-58

Equation Section 1Identification and estimation of causal effects of multiple treatments under the conditional independence assumption

Michael Lechner*, University of St. Gallen, Swiss Institute for International Economics and Applied Economic Research (SIAW)

This version: August 2000

Correspondence to:

Michael Lechner
Professor of Econometrics
Swiss Institute for International Economics and Applied Economic Research (SIAW)
Universität St. Gallen
Dufourstr. 48
CH-9000 St. Gallen, Switzerland
Email: Michael.Lechner@unisg.ch
WWW: www.siaw.unisg.ch/lechner

* Financial support from the Deutsche Forschungsgemeinschaft (DFG, 228/2-2) and the Swiss National Science Foundation (NFP 12-53735.18) is gratefully acknowledged. The paper has been presented at ESEM 1999 in Santiago de Compostela and in seminars at SOFI, Stockholm University, and IFAU, Uppsala. I thank participants for helpful discussions. Furthermore, I thank Martin Eichler, Markus Frölich, Ruth Miquel, Friedhelm Pfeiffer, and three anonymous referees for helpful comments and suggestions on a previous version of the paper. I am particularly grateful to Bruno Crepon for insisting that a result obtained in a previous version of the paper could be sharpened considerably. All remaining errors and omissions are my own.

Abstract. The assumption that the assignment to treatments is ignorable conditional on attributes plays an important role in the applied statistic and econometric evaluation literature (**C**onditional **I**ndependence **A**ssumption). This paper discusses identification using CIA when there are more than two types of mutually exclusive treatments. It turns out that low dimensional balancing scores, similar to the ones valid in the case of only two treatments, exist and can be used for identification of various causal effects. Therefore, a comparable reduction of the dimension of the estimation problem is achieved and the approach retains its basic simplicity. Furthermore, a sample reduction property is derived showing that in certain important cases it is possible to base the estimation on the specific subsample of participants. The paper also outlines a matching estimator suitable in that general framework.

Keywords: *Treatment effects, balancing score, propensity score, causal model, programme evaluation, matching.*

JEL classification*: C30, C40.*

1 Introduction

The prototypical model of the microeconometric evaluation literature is the following: An individual can choose between two states, like participation in a training programme or non-participation in such a programme. For the potential participant in such a programme an hypothetical outcome is defined for both states. This model is also termed the Roy (1951)-Rubin (1974) model of potential outcomes and causal effects.[1] Since its statistical content is most clearly spelled out in Rubin (1974), this model is called the Rubin model in the following. It clarifies that the individual causal treatment effect - defined as the difference of the two potential outcomes, for example - is not identified. Therefore, the lack of identification has to be overcome by plausible, generally untestable assumptions that depend on the problem analyzed and the data available. One such assumption is that treatment participation and treatment outcomes are independent conditional on a set of (observable) attributes. Subsequent papers by Rubin (1977) and Rosenbaum and Rubin (1983) show how this assumption could effectively be used for treatment evaluation. In many cases this identifying assumption is exploited via a matching estimator, for recent examples see Angrist (1998), Dehejia and Wahba (1998, 1999), Heckman, Ichimura, and Todd (1997, 1998), Lechner (1999) and the comprehensive survey by Heckman, LaLonde, and Smith (1999).

This literature focuses on models with only two potential states, treatment and non-treatment. However, when evaluating European labour market programmes for example a more complex framework appears to be necessary, since the actual choice set of individuals contains more than just two options. Potential participants may or may not participate in one of several different training or employment programmes. This paper extends the conventional two state framework to allow for multiple, mutually exclusive treatments. It shows that all major properties obtained by Rubin (1977) and Rosenbaum and Rubin (1983) also hold in that framework, if suitably refined.[2] The paper also shows that for specific parameters, like the treatment effect on the treated that compares two different programmes for the participants in one of those programmes, the multi-programme nature of the policy can be ignored, because individuals who are not in programmes of interest, are not needed for identification. The paper also sketches a matching estimator that takes account of this multiple treatment structure.

---

1 See for example Heckman (2000), Holland (1986), and Sobel (1994) for an extensive discussion of concepts of causality in statistics, econometrics, and other fields.

2 Parallel to this work similar ideas appeared in Imbens (1999).

2 Notation and definition of the causal effects

2.1 Two treatments

Let Y^1 and Y^0 denote the potential outcomes (*1* denotes treatment, *0* non-treatment). For participants in the treatment the actual (observable) outcome is Y^1, and Y^0 for non-participants. As a notational convention, capital letters indicate quantities of the population, whereas small letters represent their respective quantities in the sample of size *N* (*i=1,...,N*). Additionally, denote variables that are unaffected by treatments[3] - called *attributes* by Holland (1986) - by *X*. Define a binary *assignment* indicator *S*, that determines whether the unit receive the treatment (*S = 1*) or not (*S = 0*). The causal effect, usually defined as the difference of the two potential outcomes, can never be estimated, because the respective *counterfactual* (Y^1 or Y^0) to the observable outcome (Y) is unobservable. However, under certain assumptions average causal effects are identified. For simplicity, this section concentrates entirely on the average treatment effect on the treated:

$$\theta_0 := E(Y^1 - Y^0 \mid S=1) = E(Y^1 \mid S=1) - E(Y^0 \mid S=1). \tag{1}$$

The short hand notation *E(·|S=1)* denotes the mean in the population of all units who participate in the programme (*S=1*). Finally, to make the model's representation of outcomes adequate for causal analysis, the *stable-unit-treatment-value assumption* (SUTVA) has to be satisfied for all members of the population (e.g. Rubin, 1991). SUTVA excludes cross-effects, or general equilibrium effects, among potential programme participants that could occur because of their actual participation decision.[4]

The difficulty with the identification of θ_0 from a large random sample is the term $E(Y^0 \mid S=1)$, because the pair $(y_i^0, s_i = 1)$ is not observable. Much of the literature on causal models in statistics and selectivity models in econometrics is devoted to finding identifying assumptions to estimate $E(Y^0 \mid S=1)$ by using the observable pairs $(y_i^0, s_i = 0)$ in different ways. One frequently used condition states that the assignment is random conditional on a set of covariates (Rubin, 1977). Hence, the assignment is independent (denoted by $\coprod$ in the following) of the potential non-treatment outcome conditional on the value of suitably chosen covariates (conditional independence assumption, CIA):[5]

$$Y^0 \coprod S \mid X = x, \quad \forall x \in \chi. \tag{2}$$

---

3 I will use the terms treatment and programme as synonyms in the remainder of the paper.

4 Assume for the rest of the paper that the typical assumptions of the Rubin model are fulfilled (see Holland, 1986, or Rubin, 1974, for example).

5 See Dawid (1979) for notations, definitions, and implications related to the concept of conditional independence.

χ denotes the part of the attribute space for which the treatment effect is defined. If CIA holds, then $E(Y^0 \mid S=1, X=x) = E(Y^0 \mid S=0, X=x)$.[6] $P^1(x)$ denotes the propensity score that is defined as the participation probability conditional on *x* [*P(S=1|X=x)*]. If $0 < P^1(x) < 1$ holds in χ, then $E(Y^0 \mid S=1) = E[E(Y^0 \mid S=0, X=x) \mid S=1]$ can be estimated in large samples using respective sample analogues.[7]

The condition $0 < P^1(x) < 1$, frequently called *common support condition* (CSC), deserves some further remarks because of its importance for applied work. If there are regions in the attribute space χ where either treated or control observation with these specific attributes have zero probability to occur (for example when all individuals with a specific attribute include in *X* are obliged to participate), the equality $E(Y^0 \mid S=1, X=x) = E(Y^0 \mid S=0, X=x)$ is not helpful for nonparametric identification, because no informative observations $(y_i, s_i = 0, x_i = x)$ exist for these particular values of x. The only way to identify and consistently estimate any effect for such regions, would be to extrapolate from regions of χ that have positive probabilities for both treatment states to occur. Obviously the credibility of these estimates will critically rely on the credibility of the model used to extrapolate. In order to avoid any issue of extrapolation when the true $P^1(x)$ is unknown, the sample used should be restricted to a subspace of χ that has positive empirical probability of both treatment states occurring. Note however, that this procedure implicitly changes the definition of the effects estimated.

Rosenbaum and Rubin (1983, RR) showed that if CIA is valid, then the estimation problem simplifies. In the case of two treatments, RR found that if the two treatments are independent of the assignment conditional on *X*, then they are also independent conditional on specific functions of *X*, denoted as balancing scores (*b*(*X*)), that fulfil the so-called balancing score property:

$$Y^0 \coprod S \mid X = x, \forall x \in \chi \quad \rightarrow \quad Y^0 \coprod S \mid b(X) = b(x), \ \forall x \in \chi, \qquad \text{(RR)}$$

$$\textit{if} \quad E[P(S=1 \mid X=x) \mid b(X) = b(x)] = P[S=1 \mid X=x] = P^1(x), \ 0 < P^1(x) < 1, \ \forall x \in \chi.$$

Note that the random variable *S* can only be zero or one. In the set-up of RR one particularly important balancing score is the propensity score, because it reduces the dimension of the conditioning vector to one. If the potential non-treatment outcome is independent of the assignment mechanism conditional on $X = x$, then it is also independent of the assignment mechanism conditional on $P^1(X) = P^1(x)$, thus:

---

6 Note that CIA can be seen as overly restrictive, because all what is needed to identify mean effects is conditional mean independence. However, the former has the virtue of making the latter valid for all transformations of the outcome variables. Furthermore, in an application it is usually difficult to argue why conditional mean independence should hold and CIA might nevertheless be violated.

7 It is assumed that the researcher has access to an infinitely large random sample.

$$E[Y^0 \mid S=1, P^1(X)=P^1(x)] = E[Y^0 \mid S=0, P^1(X)=P^1(x)]. \quad (3)$$

Hence, $E(Y^0 \mid S=1) = E\{E[Y^0 \mid S=0, P^1(X)=P^1(x)] \mid S=1\}$ can be used for estimation. When the propensity score is known or can be consistently estimated, the major advantage of this property is the reduction of the dimension of the estimation problem, especially important for nonparametric estimation techniques.[8]

2.2 Many treatments

Consider the outcomes of (*M+1*) different mutually exclusive treatments, denoted by $\{Y^0, Y^1, ..., Y^M\}$. It is assumed that each participant receives exactly one of the treatments (typically the '0' category denotes the case of the treatment type *no treatment*). Therefore, for any participant, only one component of $\{Y^0, Y^1, ..., Y^M\}$ can be observed in the data. The remaining *M* outcomes are counterfactuals in the language of the Rubin model. Participation in a particular treatment *m* is indicated by the variable $S \in \{0,1,...M\}$.

The definitions of average treatment effects used for the case of just two treatments need to be extended. In the following equations, the focus is on a pair-wise comparison of the effects of the treatments *m* and *l*:

$$\gamma_0^{ml} := E(Y^m - Y^l) = EY^m - EY^l; \quad (4)$$

$$\alpha_0^{ml} := E(Y^m - Y^l \mid S=m,l) = E(Y^m \mid S=m,l) - E(Y^l \mid S=m,l); \quad (5)$$

$$\theta_0^{ml} := E(Y^m - Y^l \mid S=m) = E(Y^m \mid S=m) - E(Y^l \mid S=m). \quad (6)$$

γ_0^{ml} denotes the expected (average) effect of treatment *m* relative to treatment *l* for a participant drawn randomly from the population (*N*).[9] Similarly, α_0^{ml} denotes the corresponding effect for a participant randomly selected from the group of participants participating in either *m* or *l*. Note that both average treatment effects are symmetric in the sense that $\gamma_0^{ml} = -\gamma_0^{lm}$ and $\alpha_0^{ml} = -\alpha_0^{lm}$. θ_0^{ml} is the expected effect for an individual randomly drawn from the population of participants in treatment *m* only. If the participants in treatments *m* and *l* differ in a non-random fashion which is related to the outcomes, then $\theta_0^{ml} \neq -\theta_0^{lm}$, i.e. the treatment effects on the treated are not symmetric.[10]

---

8 Efficiency issues involved by conditioning on the propensity score instead of *X* are discussed in detail by Hahn (1998) and Hirano, Imbens and Ridder (2000).

9 If a variable *Z* cannot be changed by the effect of the treatment (like time constant personal characteristics of participants), then all what follows is also valid in strata of the data defined by different values of *Z*.

10 This list of treatment effects is not exhaustive, neither with respect to comparisons of types of treatments, nor with respect to populations under consideration.

It is worth noting that $\alpha_0^{ml} = E(Y^m - Y^l \mid S = m,l)$ is a weighted combination of θ_0^{ml} and θ_0^{lm}. The weights are given by the participation probabilities in the respective states *m* and *l*:

$$\begin{aligned}\alpha_0^{ml} &= E(Y^m - Y^l \mid S = m,l) \\ &= E(Y^m - Y^l \mid S = m)P(S = m \mid S = m,l) \; + \; E(Y^m - Y^l \mid S = l)[1 - P(S = m \mid S = m,l)] \\ &= \theta_0^{ml} P(S = m \mid S = m,l) \; - \; \theta_0^{lm}[1 - P(S = m \mid S = m,l)] \; ;\end{aligned}$$

$$P(S = m \mid S = m,l) = \frac{P(S = m)}{P(S = l) + P(S = m)} \; .$$

Note that γ_0^{ml} can be decomposed as follows:

$$\gamma_0^{ml} = EY^m - EY^l = \sum_{j=0}^{M} [E(Y^m \mid S = j) - E(Y^l \mid S = j)]P(S = j) \; . \tag{7}$$

Therefore, the various effects can be ordered in terms of what information is required to identify them. Identification of θ_0^{ml} requires identification of $E(Y^l \mid S = m)$, whereas identification of α_0^{ml} requires identification of $E(Y^l \mid S = m)$ and $E(Y^m \mid S = l)$. Identification of γ_0^{ml} requires either all counterfactuals of Y^m and Y^l, or at least of $E(Y^l \mid S \neq l)$ and $E(Y^m \mid S \neq m)$ (see below). Of course, if $E(Y^l \mid S = j)$ and $E(Y^m \mid S = j)$ is identified for all *j*, then $E(Y^l \mid S \neq l)$ and $E(Y^m \mid S \neq m)$ are identified as well. These considerations imply also that if all θ_0^{ml} $(m,l = 0,...,M)$ are identified, then the other effects are identified as well.

3 Identification and the balancing score

The next step is to define CIA for the case of multiple treatments. This assumption is formalized in expression (8):

$$Y^0, Y^1, ..., Y^M \perp\!\!\!\perp S \mid X = x, \forall x \in \chi \; , \qquad 0 < P^m(x) < 1 \; , \forall m = 0,...,M \; . \tag{8}$$

In this case a generalisation of the balancing score property suggested by Rosenbaum and Rubin (1983) holds as well.

Proposition 1 (*generalized balancing score property*):

$$Y^0, Y^1, ..., Y^M \perp\!\!\!\perp S \mid X = x \qquad \Rightarrow \qquad Y^0, Y^1, ..., Y^M \perp\!\!\!\perp S \mid b(X) = b(x) \; , \; \forall x \in \chi \; ,$$

$$\textit{if } E[P(S = m \mid X = x) \mid b(X) = b(x)] = P[S = m \mid X = x] = P^m(x) \; , \; 0 < P^m(x) < 1 \; ,$$

$$\forall m = 0,...,M \; . \tag{9}$$

The proof is given in Appendix A.

Functions that can be used as balancing scores are for example the vector of attributes X, or the M-dimensional vector of propensity scores $P(x) = [P^1(x),...,P^m(x),...,P^M(x)]$.[11] Note that the dimension is reduced only to the order of M. This means that from the point of view of dimension reduction, using the propensity scores directly, instead of X, as conditioning variables, is only useful when the dimension of X is larger than M.

It is obvious that the form of the CIA stated in expressions (8) and (9) identifies all treatment effects defined in the previous section, because it identifies all counterfactuals $E(Y^l \mid S = m)$ for all combinations of l and m.

The remainder of this section discusses identification for a given pair of treatments m and l. It is shown that the problem can simplify considerably. The following versions of CIA are implied by the general form given in the previous section.

Proposition 2 *(CIA identifies the treatment effects)*

a) If $Y^m, Y^l \coprod S \mid X = x$ and $0 < P^j(x) < 1$ hold for $\forall x \in \chi$ and $\forall j = m,l$, then γ_0^{ml}, γ_0^{lm}, α_0^{ml}, α_0^{lm}, θ_0^{ml}, and θ_0^{lm} are identified.

b) If $Y^m, Y^l \coprod S \mid X = x, S \in \{m,l\}$ and $0 < P^j(x) < 1$ hold for $\forall x \in \chi$ and $\forall j = m,l$, then α_0^{ml}, α_0^{lm}, θ_0^{ml}, and θ_0^{lm} are identified.

c) If $Y^l \coprod S \mid X = x, S \in \{m,l\}$ and $0 < P^l(x) < 1$ hold for $\forall x \in \{\chi \mid P^m(x) > 0\}$ then θ_0^{ml} is identified.

Part a) is the strongest form of CIA and the implied common support requirement (CSC). This form of the assumption identifies all counterfactuals $E(Y^m \mid S = j)$ and $E(Y^l \mid S = j)$, because it implies $E(Y^m \mid X = x, S = j) = E(Y^m \mid X = x, S = m)$ and $E(Y^l \mid X = x, S = j) = E(Y^l \mid X = x, S = l)$ for $\forall j = 0,...,M$.

In part b) CIA is relaxed to hold only for the subpopulations participating in either treatment m or l. Hence, the assumptions of part b) identify in general only the counterfactuals $E(Y^m \mid S = l)$ and $E(Y^l \mid S = m)$ and thus α_0^{ml}, α_0^{lm} as well as θ_0^{ml} and θ_0^{lm} are identified. Note that this assumption is implied by from the previous version of CIA, because independence of potential outcomes and assignments in the population implies independence in any subpopulation defined by assignment categories.

---

11 There are only M linearly independent probabilities, because of adding-up.

Finally in part c) of Proposition 2, CIA is further relaxed to hold only for the potential outcome Y^l. In that case it identifies only $E(Y^l \mid S = m)$ and thus only θ_0^{ml}.

The fact that for the identification of α_0^{ml}, θ_0^{ml}, and θ_0^{lm} CIA is only necessary in the subsample of participants in treatments *m* and *l* also implies that only this subsample is necessary for the empirical analysis.

Corollary *(sample reduction properties)*

a) If the conditions of Proposition 1b) hold, then only the subsample of participants in treatments *m* or *l* is needed to identify α_0^{ml}, α_0^{lm}, θ_0^{ml}, and θ_0^{lm}.

b) If the conditions of Proposition 1c) hold, then only the subsample of participants in treatments *m* or *l* is needed to identify θ_0^{ml}.

The crucial point here is to note again that identification is achieved by the equalities $E(Y^m \mid X = x, S = l) = E(Y^m \mid X = x, S = m)$ and $E(Y^l \mid X = x, S = m) = E(Y^l \mid X = x, S = l)$. $E(Y^m \mid X = x, S = m)$ and $E(Y^l \mid X = x, S = l)$ are identified from the subsamples of participants in *m* and *l*, respectively.

The *Corollary* means in practise that if the interest is only in a specific pairwise-effect for corresponding subpopulations, CIA allows one to delete the other treatments and their participants from consideration. In other words, for the estimation of α_0^{ml}, θ_0^{ml}, and θ_0^{lm} we can ignore the existence of multiple treatments.

The following proposition provides balancing score properties corresponding to the version of CIA given in Proposition 2 and proposes concrete balancing scores. To ease notation, the focus is on balancing scores with minimal dimension, e.g. the propensity score. Of course the proposition holds also for all balancing scores that are as least as fine as the propensity scores given.

Proposition 3 *(balancing score property)*

a) If $Y^m, Y^l \coprod S \mid X = x$ and $0 < P^j(x) < 1$ hold for $\forall x \in \chi$ and $\forall j = 0,...,M$, it follows that $Y^m, Y^l \coprod S \mid [P^1(X) = P^1(x),...,P^M(X) = P^M(x)]$. γ_0^{ml}, γ_0^{lm}, α_0^{ml}, α_0^{lm}, θ_0^{ml}, and θ_0^{lm} are identified.

b) If $Y^m, Y^l \coprod S \mid X = x, S \in \{m,l\}$ and $0 < P^j(x) < 1$ hold for $\forall x \in \chi$ and $\forall j = m,l$, it follows that $Y^m, Y^l \coprod S \mid P^{l|ml}(X) = P^{l|ml}(x), S \in \{m,l\}$. $P^{l|ml}(x) :=$ $P(S = l \mid S \in \{m,l\}, X = x) = \dfrac{P^l(x)}{P^l(x) + P^m(x)}$. α_0^{ml}, α_0^{ml}, θ_0^{ml}, and θ_0^{lm} are identified.

c) If $Y^l \coprod S \mid X = x, S \in \{m,l\}$ and $0 < P^j(x) < 1$ hold for $\forall x \in \chi$ and $\forall j = m,l$, it follows that $Y^l \coprod S \mid P^{l|ml}(X) = P^{l|ml}(x), S \in \{m,l\}$. $P^{l|ml}(x) := P(S = l \mid S \in \{m,l\}, X = x)$. θ_0^{ml} is identified.

Part a) is a version of the general balancing score property. It has however not of minimal dimension, as can be seen from the considerations following below (Imbens, 1999).

Part b) and c) need no explicit proof because the proofs for the binary cases already given in Rosenbaum and Rubin (1983) apply. The major implication is that even for the case of multiple treatments a reduction of the dimension to **one** is possible.

Indeed, a similar reduction is also possible for the identification of γ_0^{ml}. For details, the reader is referred to Imbens (1999), but the following paragraphs gives the reasoning why this is possible.

To discuss identification it is useful to rewrite equation (4) in the following way:

$$\begin{aligned}
\gamma_0^{ml} &= EY^m - EY^l \\
&= E(Y^m \mid S = m)P(S = m) + E(Y^m \mid S \neq m)P(S \neq m) \\
&- E(Y^l \mid S = l)P(S = l) + E(Y^l \mid S \neq l)P(S \neq l) \\
&= E(Y^m \mid S = m)P(S = m) + E[E(Y^m \mid X, S = m) \mid S \neq m]P(S \neq m) \\
&- E(Y^l \mid S = l)P(S = l) + E[E(Y^l \mid X, S = l) \mid S \neq l]P(S \neq l).
\end{aligned}$$

Hence (8) identifies $\gamma_0^{m,l}$ as long as $P^m(x)\,P^l(x) > 0$, since it implies $E(Y^j \mid X = x, S = j) = E(Y^j \mid X = x, S \neq j)$, $j = m,l$.

Defining a new random variable $\tilde{S}^j = \underline{1}(S = j)$, the following two conditions that follow from (8) are sufficient to identify $\gamma_0^{m,l}$:

$$Y^j \coprod \tilde{S}^j \mid X = x, \ \tilde{S}^j = \underline{1}(S = j), \quad \forall x \in \chi, \forall j = m,l. \tag{10}$$

Based on these conditions a balancing score property can be deduced:

$$\begin{aligned}
&Y^j \coprod \tilde{S}^j \mid X = x, \forall x \in \chi \quad \rightarrow \quad Y^j \coprod \tilde{S}^j \mid b^j(X) = b^j(x), \forall x \in \chi, \\
&\textit{if } E[P^j(x) \mid b^j(X) = b^j(x)] = P^j(x), \ 0 < P^j(x) < 1, \quad j = m,l.
\end{aligned} \tag{11}$$

Hence, $[P^m(x), \ P^l(x)]$ is a balancing score. Expression (11) corresponds again to the binary case considered by Rosenbaum and Rubin (1983) and given in expression (RR). The fact that it is applied twice - for *m* as well as for *l* - is not essential. Expression (11) leads to $E(Y^j \mid b^j(x), S = j) = E[Y^j \mid b^j(x), S \neq j]$, $j = m,l$. As for the binary case the balancing scores of minimum dimension are the marginal choice probabilities, hence γ_0^{ml} could be rewritten as follows:

$$\gamma_0^{ml} = E(Y^m \mid S=m)P(S=m) + E[E(Y^m \mid P^m(X), S=m) \mid S \neq m]P(S \neq m)$$
$$- E(Y^l \mid S=l)P(S=l) + E[E(Y^l \mid P^l(X), S=l) \mid S \neq l]P(S \neq l)\,.$$

Thus the dimension of the estimation problem is reduced to one.

Therefore, methods to reduce the dimension of the condition set to one are available for all parameters of interest.

4 Estimation

There are many ways to estimate the parameters defined and identified before. One example is kernel regression as performed by Brodaty, Crepon, and Fougere (2000, this volume). The following suggestion is in line with the conventional matching estimators used in the case of two treatments only (see for example Rosenbaum and Rubin, 1985).

a) Estimation of $P(S=j)$

The first set of components are the conditional and unconditional probabilities of the type $P(S=j)$ and $P(S=j \mid S \in \{k,j\}) = \dfrac{P(S=j)}{P(S=j)+P(S=k)}$ ($j \neq k$). Consistent estimates can be obtained by using the respective cell frequencies**.**

b) Estimation of $E(Y^j \mid S=j)$

$E(Y^j \mid S=j)$ can be estimated by the mean of the outcomes of units observed in category *j*.

c) Estimation of $E\{E[(Y^j \mid P^{j|kj}(X), S=j] \mid S=k\}$ **(** $k \neq j$ **)**

In this case the following matching estimator is feasible:

In the **first step** estimate a probability model to obtain consistent estimates of the choice probabilities $P_N^j(x)$ and $P_N^{j|kj}(x)$ (or $P_N^k(x)$ and $P_N^j(x)$) that form the respective balancing scores. For choosing that model a priori knowledge is important. For example, if the choices are ordered, like in a dose-response set-up, an ordered choice model would be appropriate.[12] In other cases a multinomial logit or

12 See Imbens (1999).

a more flexible model like a multinomial probit or a semiparametric model may be the appropriate tool.

In the **second step** $E\{E[(Y^j \mid \hat{P}_N^j(X), S = j] \mid S \neq j\}$ or $E\{E[(Y^j \mid \hat{P}_N^{j|kj}(X), S = j] \mid S = k\}$ needs to be estimated when using the probabilities as balancing scores. There are several ways to proceed. First, one could obtain a parametric, semi-parametric or a non-parametric regression estimate of the expectation conditional on the respective one or two dimensional balancing scores. The outer expectation could then be estimated by averaging that function with respect to the empirical distribution function of X in the respective subpopulation.

An alternative is to estimate both expectations in one step by using a matching estimator. The idea of the simplest version of such an estimator is to find for every participant in *k* or *(not j)* one participant in *j* that has (almost) the same balancing score. Taking the mean of the outcome variable for these matched comparison observations gives the desired estimate. Note that standard matching procedures typically use each control observation (here $S = j$) only once, because the number of comparison observations is typically much larger than the treated observations (necessary to get 'good' matches). However, for the case of many treatments each group will act as a treated group as well as a comparison group. Therefore, requiring the number of comparison observations to be larger than the number of treated observations does not make sense. Thus, one needs to rely on matching algorithms that use single observations more than once. Appendix B gives an estimator and its approximate variance using such an approach. This estimator is also used in an empirical study by Frölich, Heshmati and Lechner (2000), Gerfin and Lechner (2000) and Lechner (2000a, b). The latter two papers address several practical concerns that could arise with this kind of matching estimator.

5 Conclusion

The Rubin causal model has been the working horse in the evaluation literature. However, a model that allows for more than two treatments is necessary to evaluate the different types of active labour market policies in European countries, for example. This paper extends the classical Rubin model to the case of many treatments and discusses various definitions of the causal effects. It also discusses the identification of these effects under the conditional independence assumption. It is shown that the so-called balancing score properties of the model with two treatments can be extended. Furthermore, a sample reduction property is derived a by-product. Finally, the paper shows that feasible non-parametric estimators such as matching can be devised by exploiting the dimension reducing effect of using this balancing score property. First experiences with this approach, as contained in Brodaty, Crepon, and Fougere (2000, this volume), Frölich, Heshmati and Lechner (2000), Gerfin and Lechner (2000), Larson (2000), and Lechner (2000a, b) underline its usefulness in applied microeconometric analyses of active labour market policies.

Appendix A: Technical appendix

In the following it will be shown that Proposition 1 of the main part of the paper is correct:

$$Y^0, Y^1, ..., Y^M \perp\!\!\!\perp S \mid X = x \text{ (CIA)} \quad \rightarrow \quad Y^0, Y^1, ..., Y^M \perp\!\!\!\perp S \mid b(X) = b(x), \ \forall x \in \chi,$$

$$\text{if } E[P(S = m \mid X = x) \mid b(X) = b(x)] = P(S = m \mid X = x) = P^m(x), \ 0 < P^m(x) < 1,$$

$$\forall m = 0, \ldots, M. \qquad (9)$$

Proof:

Let $F(\cdot)$ denote the joint distribution function of S and the potential outcomes, then the following equation holds generally:

$$F(Y^0, Y^1, ..., Y^M, S \mid X) = F(S \mid Y^0, Y^1, ..., Y^M, X) F(Y^0, Y^1, ..., Y^M \mid X).$$

CIA can be expressed in terms of the distribution of S conditional on the potential outcomes:

$$F(S \mid Y^0, Y^1, ..., Y^M, X) \overset{CIA}{=} F(S \mid X). \qquad (A.1)$$

If the balancing score property given in (9) holds, then it is also true that:

$$F(S \mid Y^0, Y^1, ..., Y^M, b(X)) \overset{!}{=} F(S \mid b(X)) = F(S \mid X). \qquad (A.2)$$

Since S is discrete random variable with $M+1$ possible values, $F(S|X)$ is a discrete function with M+1 values for every given value of X. Hence, (A.2) can be reformulated in terms of probabilities:

$$P[S = m \mid Y^0, Y^1, ..., Y^M, b(X)] \overset{!}{=} P[S = m \mid b(X)] = P(S = m \mid X), \ \forall m = 0, ..., M. \qquad (A.3)$$

(A.3) will be proofed as follows:

$$P[S = m \mid Y^0, Y^1, ..., Y^M, b(X)] = E\{P[S = m \mid Y^0, Y^1, ..., Y^M, X] \mid Y^0, Y^1, ..., Y^M, b(X)\}$$

$$= E\{P[S = m \mid X] \mid Y^0, Y^1, ..., Y^M, b(X)\}$$

If the balancing score $b(X)$ is at least as fine as the propensity score $P[S = m \mid X]$, i.e. $E[P(S = m \mid X = x) \mid b(X)] = P[S = m \mid X = x]$, then $E[P(S = m \mid X = x) \mid b(X)]$ does not depend on the potential outcomes, hence:

$$E\{P[S = m \mid X] \mid Y^0, Y^1, ..., Y^M, b(X)\} = E\{P[S = m \mid X] \mid b(X)\}$$

$$= P[S = m \mid b(X)] = P(S = m \mid X), \quad \forall m = 0, ..., M.$$

Therefore, $b(X) = [P^1(X), ..., P^M(X)]$ is a valid balancing score. q.e.d.

Appendix B: An example of a matching estimator and an approximation of its variance

Table B.1 gives a condensed description of a matching protocol that could be used in practise.

Table B.1: A matching protocol for the estimation of θ_0^{ml}

Step 1	Specify and estimate a multinomial choice model to obtain $[\hat{P}_N^0(X), \hat{P}_N^1(X), ..., \hat{P}_N^M(X)]$.
Step 2	Estimate the expectations of the outcome variables conditional on the respective balancing scores. For a given value of *m* and *l* the following steps are performed: a) Compute $\hat{P}_N^{l\|ml}(X) = \frac{\hat{P}_N^l(X)}{\hat{P}_N^l(X) + \hat{P}_N^m(X)}$ or use $[\hat{P}_N^m, \hat{P}_N^l(X)]$ directly. Alternatively step 1 may be omitted and the conditional probabilities may be directly modeled (as in the binary case). b) Choose one observation in the subsample defined by participation in *m* and delete it from that pool. c) Find an observation in the subsample of participants in *l* that is as close as possible to the one chosen in step a) in terms of $\hat{P}_N^{l\|ml}(X)$ or $[\hat{P}_N^m, \hat{P}_N^l(X)]$. In the case of using $[\hat{P}_N^m, \hat{P}_N^l(X)]$ 'closeness' can be based on the Mahalanobis distance. Do not remove that observation, so that it can be used again. d) REPEAT b) and c) until no participant of *m* is left. e) Using the matched comparison group formed in c), compute the respective conditional expectation by the sample mean. Note that the same observations may appear more than once in that group.
Step 3	Repeat step 2 for all combinations of *m* and *l*.
Step 4	Compute the estimate of the treatment effects using the results of step 3 and compute their covariance matrix (see below).

Note: If the aim is to estimate only γ_0^{ml} then the algorithm changes in an obvious way.

Suppose that the matching protocol used gives an estimator for $E(Y^l \mid S = m)$ of the following type:

$$\hat{E}_N(Y^l \mid S = m) = \sum_{i \in l} w_i^m y_i^l \ .$$

The weight functions fulfil $\sum_{i\in l} w_i^m = N^m$, $\forall l = 1,...,M$. w_i^m denotes the number of observations in *m*, to which observation *i* is matched. N^m denotes the number of observations in treatment *m*.

Using this notation we get the following estimators for the various treatment effects:

$$\hat{\theta}_N^{ml} = \frac{1}{N^m}\sum_{i\in m} y_i^m - \frac{1}{N^m}\sum_{i\in l} w_i^m y_i^l \;;$$

$$\hat{\gamma}_N^{ml} = \sum_{j=0}^{M}\left[\left(\frac{1}{N^j}\sum_{i\in m} w_i^j y_i^m - \frac{1}{N^j}\sum_{i\in l} w_i^j y_i^l\right)P(S=j)\right]; \qquad \hat{\gamma}_N^{ml} = -\hat{\gamma}_N^{lm};$$

$$\hat{\alpha}_N^{ml} = \hat{\theta}_N^{ml} P(S=m \mid S=m \text{ or } S=l) - \hat{\theta}_N^{lm} P(S=l \mid S=m \text{ or } S=l); \quad \hat{\alpha}_N^{ml} = -\hat{\alpha}_N^{lm}.$$

To derive the variances of these estimators the weights and the probabilities are assumed to be fixed and the observations are assumed to be independent. The first assumption is obviously an approximation since the weights are estimated in the algorithm given in Table B.1. We also assume that the variances of the observable outcome variables are the same within a particular treatment, as well as that they do not depend on the values of the balancing scores.

$$Var(\hat{\theta}_N^{ml}) = \frac{1}{N^m} Var(Y^m \mid S=m) + \frac{\sum_{i\in l}(w_i^m)^2}{(N^m)^2} Var(Y^l \mid S=l).$$

It is useful to reformulate this estimator in the following way to obtain the variance of $\hat{\gamma}_N^{m,l}$:

$$\hat{\gamma}_N^{ml} = \sum_{i\in m} y_i^m \sum_{j=0}^{M}\left[\frac{w_i^j}{N^j}P(S=j)\right] - \sum_{i\in l} y_i^l \sum_{j=0}^{M}\left[\frac{w_i^j}{N^j}P(S=j)\right];$$

$$Var(\hat{\gamma}_N^{ml}) = \sum_{i\in m}\left[\sum_{j=0}^{M}\frac{w_i^j}{N^j}P(S=j)\right]^2 Var(Y^m \mid S=m) + \sum_{i\in l}\left[\sum_{j=0}^{M}\frac{w_i^j}{N^j}P(S=j)\right]^2 Var(Y^l \mid S=l).$$

It is again useful to reformulate the estimator ($P(S=m \mid S\in\{m,l\}) =: P^{m|ml}$) to obtain the variance of $\hat{\alpha}_N^{m,l}$:

$$\hat{\alpha}_N^{ml} = \sum_{i\in m} y_i^m\left[\frac{1}{N^m}P^{m|m,l} + \frac{w_i^l}{N^l}(1-P^{m|ml})\right] - \sum_{i\in l} y_i^l\left[\frac{1}{N^l}P^{l|ml} + \frac{w_i^m}{N^m}(1-P^{l|ml})\right];$$

$$Var(\hat{\alpha}_N^{ml}) = \sum_{i\in m}\left[\frac{1}{N^m}P^{m|m,l} + \frac{w_i^l}{N^l}(1-P^{m|m,l})\right]^2 Var(Y^m \mid S=m) +$$

$$+ \sum_{i\in l}\left[\frac{1}{N^l}P^{l|ml} + \frac{w_i^m}{N^m}(1-P^{l|ml})\right]^2 Var(Y^l \mid S=l).$$

Since this estimator does take account of the fact that the weights are computed based on estimated quantities and of the fact of matching itself, a bootstrap may be used as an alternative. However, results by Lechner (2000b) show little difference between the bootstrap variances and the approximate variances.

References

Angrist, J.D. (1998): "Estimating Labor Market Impact of Voluntary Military Service Using Social Security Data ", *Econometrica*, 66, 249-288.

Brodaty, Th., B. Crepon, and D. Fougere (2000): "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France", 1986-1988, this volume.

Dehejia, R., and S. Wahba (1998): "Propensity Score Matching Methods for Non-experimental Causal Studies", *NBER working paper*, 6829.

Dawid, A. P. (1979): "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society Series B*, 41, 1-31, with discussion.

Dehejia, R. H. and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes", *Journal of the American Statistical Association*, 94, 1053-1062.

Frölich, M., A. Heshmati, and M. Lechner (2000):" A Microeconometric Evaluation of Rehabilitation of Long-term Sickness in Sweden", *discussion paper, 2000-04*, Department of Economics, University of St. Gallen.

Gerfin, M. and M. Lechner (2000): "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland", *discussion paper, 2000-10,* Department of Economics, University of St. Gallen.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66, 315-331.

Heckman, J. J. (2000): „Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective“, *Quarterly Journal of Economics*, 115, 45-97.

Heckman, J.J., H. Ichimura, and P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from a Job Training Programme", *Review of Economic Studies*, 64, 605-654.

Heckman, J.J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.

Heckman, J.J., R.J. LaLonde, and J.A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs", forthcoming in O. Ashenfelter and D. Card (eds.): *Handbook of Labor Economics*, Vol. III A, chapter 31, 1865-2097.

Hirano, K., G. W. Imbens, and G. Ridder (2000): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *NBER technical working papers*, 251.

Holland, P.W. (1986): "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81, 945-970, with discussion.

Imbens, G. (1999): "The Role of the Propensity Score in Estimating Dose-Response Functions", *NBER technical working paper*, 0237, forthcoming in *Biometrica*.

Larsson, L. (2000): "Evaluation of Swedish youth labour market programmes", *IFAU discussion paper*, 2000:1.

Lechner, M. (1999): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification", *Journal of Business & Economic Statistics*, 17, 74-90.

Lechner, M. (2000a): "Programme Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labour Market Policies", *discussion paper, 2000-01*, Department of Economics, University of St. Gallen.

Lechner, M. (2000b): "Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods", *discussion paper 2000-14*, Department of Economics, University of St. Gallen.

Rosenbaum, P.R. and D.B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrica*, 70, 41-50.

Rosenbaum, P.R. and D.B. Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", *The American Statistician*, 39, 33-38.

Roy, A.D. (1951): "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*, 3, 135-146.

Rubin, D.B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66, 688-701.

Rubin, D.B. (1977): "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, 1-26.

Rubin, D.B. (1991): "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism", *Biometrics*, 47, 1213-1234.