DISCUSSION PAPER SERIES

IZA DP No. 91

Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption

Michael Lechner

December 1999

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption

Michael Lechner
University of St. Gallen

Discussion Paper No. 91
December 1999

IZA

P.O. Box 7240
D-53072 Bonn
Germany

Tel.: +49-228-3894-0
Fax: +49-228-3894-210
Email: iza@iza.org

This Discussion Paper is issued within the framework of IZA's research area *Project Evaluation.* Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

IZA Discussion Paper No. 91
December 1999

ABSTRACT

Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption[*]

The assumption that the assignment to treatments is ignorable conditional on attributes plays an important role in the applied statistic and econometric evaluation literature. Another term for it is conditional independence assumption. This paper discusses identification 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 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. The paper also outlines a matching estimator potentially suitable in that framework.

JEL Classification: C30, C40

Keywords: Treatment effects, balancing score, propensity score, causal model, program evaluation, matching

Michael Lechner
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

[*] Financial support from the Deutsche Forschungsgemeinschaft (DFG, project 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, and two 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.

1 Introduction

The prototypical model of the econometric evaluation literature is the following: An individual can choose between two states, like participation in a training program or non-participation in such a program. The potential participant in such a program will get an hypothetical outcome in 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 never identified. Therefore, the lack of identification has to be overcome by plausible, generally untestable assumptions that usually depend heavily on the problem analyzed and the data available. One such assumption is that treatment participation and treatment outcome is 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 Angrist (1998), Dahejia and Wahba (1998, 1999), Heckman, Ichimura, and Todd (1997, 1998), Lechner (1999a) and the very 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 programs 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 perhaps several different training programs or an employment program, or something else. This paper extends the conventional two state framework to allow for multiple mutually exclusive treatments. It shows that all major properties shown by Rubin (1977) and Rosenbaum and Rubin (1983) also hold in that framework, if suitably refined.[2] The paper also sketches a matching estimator that takes account of this multiple treatment structure.

[1] See for example 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 outcomes (*1* denotes treatment, *0* non-treatment). As a notational convention, capital letters indicate quantities of the population or of members of the population, whereas small letters represent their respective quantities in the sample of size *N* (*i=1,...,N*). The units of the sample are supposed to stem from *N* independent draws in this population. Additionally, denote variables that are unaffected by treatments - 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*). For participants in the treatment the actual (observable) outcome is Y^1, and Y^0 for non-participants. The causal effect, for example 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 never observed. However, under certain assumptions the average causal effect, denoted by θ_0 and defined in equation (1), is identified. For simplicity, within this section we concentrate 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 training (*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). Among other things, SUTVA excludes cross-effects, or general equilibrium effects, among potential treatment participants that could occur because of their actual participation decision.

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 somehow using the observable pairs $(y_i^0, s_i = 0)$. One such condition states that the assignment is random conditional on a set of covariates (Rubin, 1977). Hence, the assignment is independent (denoted by $\coprod$) of the potential non-treatment outcome conditional on the value of a covariate set or attribute set (conditional independence assumption, CIA):[3]

$$Y^0 \coprod S \mid X = x,\ \forall x \in \chi. \qquad (2)$$

χ denotes all of the attribute space for which the treatment effect is defined. If CIA holds, then $E(Y^0 \mid S = 1, X = x) =$ $E(Y^0 \mid S = 0, X = x)$. $P^1(x)$ 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.

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 score (*b*(*X*)), that fulfil the so-called balancing score property:

$$Y^0 \coprod S \mid X = x, \forall x \in \chi \quad \Rightarrow \quad Y^0 \coprod S \mid b(X) = b(x),\ \forall x \in \chi,$$

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

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 set 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:

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

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 $\sqrt{N}$ -consistently estimated with a parametric model,

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

then the major advantage of this property is the reduction of the dimension of the estimation problem, especially important for nonparametric estimation techniques.[4]

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 number of participants observed in a random sample to participate in treatment *m* is denoted by N^m $(N = \sum_{m=0}^{M} N^m)$.

The definitions of average treatment effects used for the case of just two treatments need to be extended.[5] In the following equations, the focus is on a pair-wise comparison of the effects of the treatments *m* and *l*:

$$\gamma_0^{m,l} = E(Y^m - Y^l) = EY^m - EY^l; \tag{4}$$

$$\alpha_0^{m,l} = E(Y^m - Y^l | S = m, l) = E(Y^m | S = m, l) - E(Y^l | S = m, l); \tag{5}$$

$$\theta_0^{m,l} = E(Y^m - Y^l | S = m) = E(Y^m | S = m) - E(Y^l | S = m). \tag{6}$$

$\gamma_0^{m,l}$ denotes the expected (average) effect of treatment *m* relative to treatment *l* for a participant drawn randomly from the population (*N*).[6] Similarly, $\alpha_0^{m,l}$ denotes the same 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 $\gamma_0^{m,l} = -\gamma_0^{l,m}$ and $\alpha_0^{m,l} = -\alpha_0^{l,m}$.[7] $\theta_0^{m,l}$ is the same expected effect for an individual randomly drawn from the population of

[4] The trade-offs involved by conditioning on $P^1(X)$ instead of *X* are discussed in detail by Hahn (1998).

[5] 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).

[6] 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*.

participants in treatment *m* only. Note that if the participants in treatments *m* and *l* differ in a non-random fashion, then $\theta_0^{m,l} \neq -\theta_0^{l,m}$, i.e. these treatment effects on the treated are not symmetric.[8]

It is worth noting that $\alpha_0^{m,l} = E(Y^m - Y^l | S = m,l)$ is a weighted combination of $\theta_0^{m,l}$ and $\theta_0^{l,m}$. The weights are given by the participation probabilities in the respective states *m* and *l*:

$$\alpha_0^{m,l} = E(Y^m - Y^l | S = m,l)$$

$$= E(Y^m - Y^l | S = m) P(S = m | S = m,l) + E(Y^m - Y^l \mid S = l)[1 - P(S = m \mid S = m,l)]$$

$$= \theta_0^{m,l} P(S = m | S = m,l) \text{ - } \theta_0^{l,m}[1 - P(S = m \mid S = m,l)];$$

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

3 Identification and the balancing score

3.1 Conditional independence assumption

In this paper identification is considered for a particular assumption that plays a prominent role in evaluation studies, namely the conditional independence assumption (Rubin, 1977). The conditional independence assumption (CIA) states that the potential treatment outcomes are independent of the assignment mechanism for any given value of a vector of attributes (X) in a particular attribute space χ.[9] This assumption is formalized in expression (7):

$$Y^0, Y^1, \ldots, Y^M \amalg S | X = x, \forall x \in \chi. \qquad (7)$$

[7] For $m = l$, all effects are of course zero.

[8] Note that this list of treatment effects is by no means exhaustive, neither with respect to comparisons of types of treatments, nor with respect to populations under consideration. These issues will be further explored in future work.

[9] Note that CIA can be seen as overly restrictive, since 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.

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

$$Y^0, Y^1, \ldots, Y^M \coprod S | X = x \qquad \rightarrow \qquad Y^0, Y^1, \ldots, Y^M \coprod S | b(X) = b(x),\ \forall x \in \chi,$$

$$if\ E[P(S = m \mid X = x) \mid b(X) = b(x)] = P[S = m \mid X = x] = P^m(x),\ 0 < P^m(x) < 1, \forall m = 0, \ldots, M. \quad (8)$$

The proof of this property 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 $P(x) = [P^1(x), \ldots, P^m(x), \ldots, P^M(x)]$.[10] 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 shown in the next section that versions of (7) exist that are technically less restrictive but nevertheless sufficient to identify the various treatment effects. Their main advantage will be to reduce the dimension of the estimation problem still further.

3.2 Identification and balancing scores

This section discusses the identification of $\theta_0^{m,l}$ and $\gamma_0^{m,l}$ from an infinitely large random sample. In such a sample all participation probabilities are identified. Therefore, and since it is shown in section 2 that $\alpha_0^{m,l} = \theta_0^{m,l} P(S = m | S = m, l) - \theta_0^{l,m} P(S = l | S = m, l)$, there is no need to address the identification of $\alpha_0^{m,l}$ explicitly. $\alpha_0^{m,l}$ is identified whenever $\theta_0^{m,l}$ and $\theta_0^{l,m}$ are identified.

3.2.1 The effect for the population ($\gamma_0^{m,l}$)

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

[10] Note that there are only *M* linearly independent probabilities, because of adding-up.

$$\gamma_0^{m,l} = EY^m - EY^l$$

$$= E(Y^m \mid S = m)P(S = m) + E(Y^m \mid S \neq m)P(S \neq m)$$

$$- E(Y^l \mid S = l)P(S = l) + E(Y^l \mid S \neq l)P(S \neq l)$$

$$= E(Y^m \mid S = m)P(S = m) + \underset{X}{E}[E(Y^m \mid X, S = m) \mid S \neq m]P(S \neq m)$$

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

Hence (7) 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 (7) are sufficient to identify $\gamma_0^{m,l}$:

$$Y^j \coprod \tilde{S}^j \mid X = x,\ \tilde{S}^j = \underline{1}(S = j), \quad \forall x \in \chi, \forall j = m,l\,. \qquad (9)$$

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

$$Y^j \coprod \tilde{S}^j \mid X = x, \forall x \in \chi \qquad ➔ \qquad Y^j \coprod \tilde{S}^j \mid b^j(X) = b^j(x), \forall x \in \chi\,,$$

$$\textit{if}\ E[P^j(x) \mid b^j(X) = b^j(x)] = P^j(x)\,,\ 0 < P^j(x) < 1\,, \quad j = m,l\,. \qquad (10)$$

Expression (10) corresponds 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. Hence no further proof is necessary.

Expression (10) 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 $\gamma_0^{m,l}$ could be rewritten as follows:

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

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

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

3.2.2 The effect for participants in *m* ($\theta_0^{m,l}$)

As before it is useful to rewrite equation (6) to discuss identification:

$$\theta_0^{m,l} = E(Y^m \mid S=m) - E(Y^l \mid S=m)$$

$$= E(Y^m \mid S=m) - \underset{X}{E}[E(Y^l \mid X, S=l) \mid S=m]$$

Hence, (7) identifies $\theta_0^{m,l}$ as long as $P^m(x)\, P^l(x) > 0$, since it implies $E(Y^l \mid X=x, S=l)$ = $E(Y^l \mid X=x, S=m)$.[11]

To derive a balancing score of dimension one again note that (7) implies the independence of Y^l and *S* within any restricted choice set defined by values of *S*. Therefore, the following condition holds:

$$Y^l \amalg S \mid [X=x, S=l,m], \quad \forall x \in \chi. \tag{11}$$

'S=l,m' is a short hand notation for the event *'S = l or S = m'*. (11) is sufficient to identify $\theta_0^{m,l}$, because it implies $E(Y^l \mid X=x, S=l)$ = $E(Y^l \mid X=x, S=m)$. (11) also implies the following balancing score property:

$$Y^l \amalg S \mid [X=x, S=l,m],\ \forall x \in \chi \quad \Rightarrow \quad Y^l \amalg S \mid [b^{l|ml}(X=x), S=l,m],\ \forall x \in \chi,$$

[11] $P^m(x)\, P^l(x) > 0$ implies $1 > P(S=m \mid x, S=m,l) = P^m(x)/[P^m(x)+P^l(x)] > 0$.

$$if\ E[P(S=l \mid X=x, S=l,m) \mid b^{l|ml}(X) = b^{l|ml}(x)] = P(S=l \mid X=x, S=l,m) = P^{l|ml}(x),$$

$$0 < P^{l|ml}(x) < 1. \qquad (12).$$

Again, since the only population of interest is the one with *S* = *m* or *S* = *l*, the proof of this balancing score property is the same as in the binary case considered by Rosenbaum and Rubin (1983). Hence no proof is given here.

Expression (12) leads to $E(Y^l \mid b^{l|ml}(x), S=l)$ = $E[Y^l \mid b^{l|ml}(x), S=m]$. Contrary to the case considered in the previous section, the balancing score of minimum dimension is the conditional choice probability $P^{l|ml}(x)$, so that $\theta_0^{m,l}$ could be expressed as follows:

$$\theta_0^{m,l} = E(Y^m \mid S=m) + \underset{P^{l|ml}(X)}{E}[E(Y^l \mid P^{l|ml}(X), S=l) \mid S=m].$$

Again, the dimension of the estimation problem is reduced to one. In the cases when the conditional choice probabilities are more difficult to obtain than the marginal ones, it may be attractive to condition on $P^l(X)$ and $P^m(X)$ instead of $P^{l|ml}(X)$. This also identifies $\theta_0^{m,l}$ because $P^l(X)$ together with $P^m(X)$ is finer than $P^{l|ml}(X)$.[12]

4 Potential estimators

To obtain consistent estimates of the treatment effects discussed above consistent estimates of their components are needed. 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).

[12] $E[P^{l|ml}(X) \mid P^l(X), P^m(X)]$ = $E[\frac{P^l(X)}{P^l(X)+P^m(X)} \mid P^l(X), P^m(X)]$ = $P^{l|ml}(X)$.

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

The first set of components are the conditional and unconditional probabilities of the type $P(S=j)$ and $P(S=j \mid S=k \text{ } or \text{ } S=j)=\frac{P(S=j)}{P(S=j)+P(S=k)}$ $(j \neq k)$. 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 $E\{E[(Y^j \mid b^j(X), S=j] \mid S \neq j\}$ **and** $E\{E[(Y^j \mid b^{j|kj}(X), S=j] \mid S=k\}$ **(**$k \neq j$**)**

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 $\hat{P}_N^j(x)$ and $\hat{P}_N^{j|kj}(x)$ (or $\hat{P}_N^k(x)$ and $\hat{P}_N^j(x)$) that form the respective balancing scores. For the choice of 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.[13] In other cases a multinomial logit or a more flexible model like a multinomial probit or a semiparametric model may be the appropriate choice.

In the **second step** $E\{E[(Y^j \mid \hat{P}_N^j(X), S=j] \mid S \neq j\}$, $E\{E[(Y^j \mid \hat{P}_N^{j|kj}(X), S=j] \mid S=k\}$ or $E\{E[(Y^j \mid \hat{P}_N^j(X), \hat{P}_N^k(X), S=j] \mid S=k\}$ needs to be estimated when using the probabilities as balancing scores. There are several options to proceed. First, one could get 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

[13] See Imbens (1999).

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 much sense. Thus, one needs to rely on matching algorithms that use single observations more than once. Appendix B gives an estimator and its variance using such an approach. This estimator is also used in an empirical study by Lechner (1999b). Two practical concerns could arise with this kind of matching estimator. First, it may be that the respective distributions of the scores do not overlap. This can be checked by comparing the distributions of the respective balancing scores in the respective subsamples. Second, due to the multiple use of single observations, it could be that a few observations are ’over-used’ in the sense of unnecessary inflating the variance of the estimator. This can easily be checked. If this phenomenon appears, a more sophisticated version of matching is called for.

5 Conclusion

The Rubin causal model has been the working horse in the evaluation literature. However, a model that allows for more than two treatment possibilities is necessary to evaluate the different types of active labour market policies in European countries, for example. The paper extends the classical Rubin model to the case of many treatments and discusses various measures 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 to that model as well. Finally, the paper shows that feasible non-parametric estimators such as matching can be devised by exploiting the dimension reducing effect of using that balancing score property.

Appendix A: Proof of the balancing score property

In the following it will be shown that the claim made in (8) of the main part of the paper is correct:

$$Y^0,Y^1,\ldots,Y^M \amalg S|X=x \text{ (CIA)} \qquad ➔ \qquad Y^0,Y^1,\ldots,Y^M \amalg S|b(X)=b(x),\ \forall x\in\chi,$$

$$if\ E[P(S=m \mid X=x) \mid b(X)=b(x)] = P[S=m \mid X=x] = P^m(x),\ 0<P^m(x)<1,\ \forall m=0,\ldots,M\ .(8)$$

Proof:

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

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

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

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

If the balancing score property given in (8) holds, then it is also true that:

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

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:

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

(A.3) will be proofed as follows:

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

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

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

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

$$= P[S=m|b(X)] = P(S=m|X), \qquad \forall m=0,\ldots,M.$$

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

Appendix B: Matching estimators and their variances

Table B.1 gives a condensed description of a matching protocol that could be used in practise.[14]

[14] For an application see Lechner (1999b).

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

Step 1	Specifiy 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 $\hat{P}_N^{l\|ml}(X) = \dfrac{\hat{P}_N^l(X)}{\hat{P}_N^l(X) + \hat{P}_N^m(X)}$ 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 modelled (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 a) and b) until no participant in *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 $\gamma_0^{m,l}$ or γ_0^m, then the algorithm simplifies in an obvious way.

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

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

The weight functions fulfil $\sum\nolimits_{i \in l} w_i^m = N^m, \ \forall\, l = 1, ..., M$. N^m denotes the number of observations in treatment *m*.

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

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

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

$$\hat{\alpha}_N^{m,l} = \hat{\theta}_N^{m,l} P(S=m|S=m \, or \, S=l) - \hat{\theta}_N^{l,m} P(S=l|S=m \, or \, S=l); \quad \hat{\alpha}_N^{m,l} = -\hat{\alpha}_N^{l,m} .$$

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.

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

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

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

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

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

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

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

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

References

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

Dahejia, R., and S. Wahba (1998): "Propensity Score Matching Methods for Nonexperimental Causal Studies", *NBER working paper*, 6829.

Dahejia, R., and S. Wahba (1999): "Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programmes *", Journal of the American Statistical Association*, forthcoming.

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

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., 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.

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.

Lechner, M. (1999a): "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. (1999b): "Propensity Score Matching and the Bias when Treatment Heterogeneity is Ignored", *mimeo*.

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.