IFAU – INSTITUTE FOR
LABOUR MARKET POLICY
EVALUATION

Covariate selection for non-parametric estimation of treatment effects

Xavier de Luna
Ingeborg Waernbaum

WORKING PAPER 2005:4

The Institute for Labour Market Policy Evaluation (IFAU) is a research institute under the Swedish Ministry of Industry, Employment and Communications, situated in Uppsala. IFAU's objective is to promote, support and carry out: evaluations of the effects of labour market policies, studies of the functioning of the labour market and evaluations of the labour market effects of measures within the educational system. Besides research, IFAU also works on: spreading knowledge about the activities of the institute through publications, seminars, courses, workshops and conferences; creating a library of Swedish evaluational studies; influencing the collection of data and making data easily available to researchers all over the country.

IFAU also provides funding for research projects within its areas of interest. There are two fixed dates for applications every year: April 1 and November 1. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The authority has a traditional board, consisting of a chairman, the Director-General and eight other members. The tasks of the board are, among other things, to make decisions about external grants and give its views on the activities at IFAU. A reference group including representatives for employers and employees as well as the ministries and authorities concerned is also connected to the institute.

Postal address: P.O. Box 513, 751 20 Uppsala
Visiting address: Kyrkogårdsgatan 6, Uppsala
Phone: +46 18 471 70 70
Fax: +46 18 471 70 71
ifau@ifau.uu.se
www.ifau.se

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

ISSN 1651-1166

Covariate selection for non-parametric estimation of treatment effects[*]

Xavier de Luna[†]and Ingeborg Waernbaum[‡]

January 25, 2005

Abstract

In observational studies, the non-parametric estimation of a binary treatment effect is often performed by matching each treated individual with a control unit which is similar in observed characteristics (covariates). In practical applications, the reservoir of covariates available may be extensive and the question arises which covariates should be matched for. The current practice consists in matching for covariates which are not balanced for the treated and the control groups, i.e. covariates affecting the treatment assignment. This paper develops a theory based on graphical models, whose results emphasize the need for methods looking both at how the covariates affect the treatment assignment and the outcome. Furthermore, we propose identification algorithms to select a minimal set of covariates to match for. An application to the estimation of the effect of a social program is used to illustrate the implementation of such algorithms.

Keywords: Graphical models, Matching estimators, Observational studies, Potential outcomes, Social programs.
JEL: C14

[*] We are grateful to G. Broström, P. Johansson and M. J. Lee for comments that lead to an improved presentation of the paper. We also thank seminar participants at the Institute for Labour Market Policy Evaluation, Uppsala, at the Department of Statistics and Mathematical Statistics, Umeå, at the 20th Nordic Conference on Mathematical Statistics (Jyväskylä, 2004) and at the 6th World Congress of the Bernoulli Society (Barcelona, 2004) for helpful comments. The work reported here was financially supported by the Institute for Labour Market Policy Evaluation and by the Swedish Council for Working Life and Social Research.

[†] Department of Statistics, Umeå University, Umeå S-90187, Sweden. Email: Xavier.deLuna@stat.umu.se

[‡] Department of Statistics, Umeå University. Email: Ingeborg.Waernbaum@stat.umu.se

1 Introduction

The potential outcome framework (also called the Rubin model) was introduced by Rubin (1974) to estimate the effect of a binary treatment on an outcome of interest based on observational data, i.e., where treatment assignment to individuals has not been randomized. In this context, the identification of the average treatment effect can be achieved under specific conditions, see, e.g., Holland (1986). The main assumption for identifiability is that the treatment assignment can be considered as having been randomized when conditioning on a set of pre-treatment variables, also called covariates. In such cases, the average treatment effect can be estimated non-parametricaly with matching estimators, where, for instance, each treated individual is compared to an untreated (control) individual having identical or similar characteristics (values for the covariates); see, e.g. Cochran & Rubin (1973), Rosenbaum (2002) and Imbens (2004). The concept of matching corresponds to the idea of "controlling" or "adjusting" for covariates in parametric regression models.

In typical observational studies, a large amount of covariates is available, describing the individuals to enter the study before being treated. In order to fulfill the condition of randomized-like treatment assignment mentioned above, it is tempting to match for as many covariates as possible. However, adjusting for covariates that are not necessary (overmatching) lowers the quality of the estimation as was noted, e.g., by Rosenbaum (2002, pp.76). This happens because matched individuals are not identical (at least as soon as continuous covariates are involved) but only similar with respect to the covariates matched for. Abadie & Imbens (2002) showed that the bias of matching estimators increases with the number of continuous covariates matched for. In Heckman & Navarro-Lozano (2004) it was noted that the literature on matching provides no guidance on the choice of a minimal set of covariates in order to avoid overmatching.

In this paper we introduce a two-step procedure for the identification of a minimal set of covariates that guarantees the unbiasedness of the estimate of the treatment effect (given that there are no unobserved covariates that need to be matched for) while making sure that no unnecessary covariates are controlled for. In a first step, the variables predicting the treatment are identified. The second step amounts to a search among the variables identified in the first step for variables predicting the outcome for a given treatment assignment. Alternatively, these two steps may be taken in the reverse order with some slight modifications. Note that it is fairly common in practice to only apply the first step described above. However, already Cochran (1965) was aware that one should also look at how the covariates affect the outcome (step 2 above) in order to identify an appropriate set of covariates to match for. In this respect, our results confirm Cochran's early insight.

The paper is organized as follow. Section 2 introduces the models and estimators of interest. Section 3 reviews graphical models and some of their theory in order to apply them to the context of this paper. In particular, a graphical representation of the Rubin model is proposed. In Section 4 a theory is developed allowing us to introduce algorithms for the identification of a minimal set of covariates. Section 5 discusses the practical implementation of the algorithms. The implementation is then illustrated in Section 6 by means of an application to the estimation of a social program effect. Section 7 concludes the paper.

2 Framework: Model and estimators

2.1 The Rubin model

The potential outcome framework of Rubin (1974) is frequently used in observational studies to assess the effect of a binary treatment T ($T = 0$ when not treated, and $T = 1$ when treated) on an outcome of interest. For a given unit/individual, two random variables are defined: the outcome when not treated, Y_0, and the outcome when treated, Y_1. These two variables cannot both be observed because an individual is either non-treated or treated. The estimand of interest is typically an expected value of the difference between the potential outcomes: the average treatment effect, $E(Y_1 - Y_0)$ and/or the average treatment effect on the treated $E(Y_1 - Y_0 \mid T = 1)$.

An assumption (the stable unit treatment value assumption, SUTVA) made in this framework is that the values of Y_1 and Y_0 for the units are the same regardless of the values we observe on T. This means that the treatment assignment to one unit does not affect the value of the potential outcomes on that unit or any other unit. For further discussion on SUTVA see Rubin (1991).

In an experiment, where the treatment assignment is randomized, we have that the treatment assignment and the potential outcomes are independent, denoted

$$Y_1, Y_0 \perp\!\!\!\perp T. \tag{1}$$

Thus, because (1) implies

$$E(Y_1 - Y_0) = E(Y_1 \mid T = 1) - E(Y_0 \mid T = 0), \tag{2}$$

an unbiased estimate of the average treatment effect can be obtained by taking the difference between the sample average of the treated and the sample average of the controls (untreated).

In observational studies (without randomization), (1) does not hold and we need to adjust for differences in the covariates between treated and controls. Adjustment is often carried out by looking at differences in outcomes between treated and controls conditional on some value of the covariates. The following assumptions underlie such procedures, where $\mathbf{X}$ denotes a set of pre-treatment variables observed for each unit:

(A.1) $Y_0 \perp\!\!\!\perp T | \mathbf{X}$,

(A.2) $P(T = 1 \mid \mathbf{X}) < 1$,

(A.3) $Y_1 \perp\!\!\!\perp T|\mathbf{X}$,

(A.4) $P(T = 0 \mid \mathbf{X}) < 1$.

For instance, $Y_j \perp\!\!\!\perp T|\mathbf{X}$ denotes that the treatment variable T and the potential outcome Y_j are independent given $\mathbf{X}$; see Dawid (1979) for a general reference on conditional independence.

The estimand of interest may be estimated by noting that, if (A.3-A.4) hold then

$$E(Y_1 \mid \mathbf{X}) = E(Y_1 \mid \mathbf{X}, T = 1),$$

and if (A.1-A.2) hold then

$$E(Y_0 \mid \mathbf{X}) = E(Y_0 \mid \mathbf{X}, T = 0).$$

Hence, assuming (A.1-A.4) we have

$$E(Y_1 - Y_0) = E(E(Y_1 \mid \mathbf{X}, T = 1) - E(Y_0 \mid \mathbf{X}, T = 0)), \tag{3}$$

showing that unbiased estimation of the average treatment effect is possible with the data at hand.

2.2 Estimation by matching

The application of the potential outcome framework to the estimation of average treatment effects may be illustrated with the following simple example.

Assume that we have a random sample of n individuals, some of which have been treated. Assume that the older you are the more likely you are to take the treatment, and that the (expected) response to treatment differs with age. In this setting, we have that (2) (and, hence, (1)) does not hold. On the other hand, we assume that (A.1-A.4) hold for $\mathbf{X} = X_1$, where X_1 denotes age.

Here, because (2) does not hold, we cannot estimate the average treatment effect, $E(Y_1 - Y_0)$, with

$$\frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0i},$$

where Y_{0i} and Y_{1i} are the observed outcome for the n_0 and n_1 individuals in the control and treated group respectively, with $n_0 + n_1 = n$. This estimator is biased because of the confounding variable age.

An unbiased estimator of the average treatment effect must adjust for age. Assuming that (A.1-A.4) hold for $\mathbf{X} = X_1$, an example of a non-parametric estimator based on (3) is then

$$\frac{n}{n_1}\sum_{i=1}^{n_1}\left(Y_{1i} - \hat{Y}_{0i}\right) + \frac{n}{n_0}\sum_{i=1}^{n_1}\left(\hat{Y}_{1i} - Y_{0i}\right), \tag{4}$$

where $\hat{Y}_{0i} = Y_{0j}$, denotes the response of an individual j from the control group who has same (or similar) age as individual i, and $\hat{Y}_{1i} = Y_{1j}$, denotes the response of a treated individual j who has same (or similar) age as i. This estimator is called a matching estimator because each treated unit is matched (with respect to X_1) to a control, and vice versa.

2.3 Bias due to overmatching

Assume that in the above presented example, we have measured another variable X_2 that is associated with the treatment but not with the response. Since the variable is not affecting the response, we still have that assumptions (A.1-A.4) hold for $\mathbf{X} = X_1$, and there is no need to adjust for this extra variable. However, if the analyst does not have prior knowledge on the nature of the association between the variables, she/he may be tempted to control for X_2 as well, when exploring the data and discovering that X_2 is not balanced for (has the same distribution for) the treated and controls.

An estimator such as (4), is biased if matching is not exact, i.e. when matched individuals do not exactly have the same value for the covariates that are matched for. This happens with continuous covariates. Results concerning the large sample properties of matching estimators show that the bias increases with the number of continuous covariates (Abadie & Imbens 2002), stressing the importance of avoiding superfluous conditioning.

In the example discussed, we need to adjust for X_1, but should avoid adjusting for X_2 since it brings in an unecessary bias. If exact matching on X_2 is possible then its use will not introduce a bias, although a loss of efficiency will still be implied (the larger the dimension of $\mathbf{X}$ the more observations are needed to find a match).

In many applications it is not known which covariates one should be adjusting for, and we, therefore, study their identification in the remaining of the paper. For this purpose, we use a graphical representation of the relation between variables.

3 Graphical modeling

3.1 Graphs, causality and conditional independence

In order to develop, in Section 4, a theory for the identification of covariates we will use a graphical representation of certain properties of the Rubin model.

A graph (see, e.g., Lauritzen, 1996) is a pair $\mathcal{G} = (\mathbf{V}, E)$, where $\mathbf{V}$ is a set of nodes representing the variables in the model and E is a set of edges representing the relations between these variables. When two variables $X, Y \in \mathbf{V}$ are such that both the edges XY and YX belong to E, then the edge between X and Y is said undirected (symbolically: $X - Y$). The edge is directed if, e.g., only XY belongs to E (symbolically: $X \rightarrow Y$; we say that X is a parent of Y).

In particular, graphs with only directed edges and which do not contain cycles − called directed acyclic graphs (DAG)− are increasingly used for modeling causal relationships, see, e.g., Pearl (2000), Lauritzen (2001) and Dawid (2002). Basically a DAG is built by drawing directed arrows that stand for causal relations. For instance, the graph $X \rightarrow Y$ is interpreted as X has an effect on Y, or, if X is an attribute (which cannot be intervened upon), Y has no effect on X though Y and X are dependent.

Once a graph $\mathcal{G} = (\mathbf{V}, E)$ is given, a joint probability distribution, $P(\mathbf{v})$, for the variables in $\mathbf{V}$ which is compatible with $\mathcal{G}$ is specified to carry out inference.

Definition 1 *(Markov compatibility, Pearl, 2000, Sec. 1.2) A joint probability distribution $P(\mathbf{v})$ for the variables in $\mathbf{V}$ is compatible with a graph $\mathcal{G} = (\mathbf{V}, E)$ if it admits the factorization*

$$P(x_1, \ldots, x_p) = \prod_i P(x_i|\{parents\ of\ x_i\}),$$

where the set $\{parents\ of\ x_i\} \subseteq \mathbf{V}$ contains the variables having an arrow pointing towards X_i in $\mathcal{G}$.

The set of distributions which are compatible with a DAG $\mathcal{G}$ is characterized by a list of conditional independencies between the variables in $\mathbf{V}$. These independencies can be retrieved directly from the graph by using the d-separation criterion (Verma & Pearl 1990), or the moralization criterion (Lauritzen, Dawid, Larsen & Leimer 1990). In order to describe d-separation we need the notion of a path between two nodes in a graph. A path from X to Y is a sequence of distinct nodes which are connected by an arrow in any direction.

Definition 2 *(d-separation, Pearl, 2000, Sec. 1.2) A path from A to B in a DAG is said to be blocked by $\mathbf{C} \subseteq \mathbf{V}$ if it contains a node N such that either i) $N \in \mathbf{C}$ and arrows in the path are not such that $\rightarrow N \leftarrow$,(then N is called a collider) or ii) $N \notin \mathbf{C}$, there is no arrow such that $N \rightarrow B$ for $C \in \mathbf{C}$, and arrows of the path are such that $\rightarrow N \leftarrow$.*

Then, a set $\mathbf{C}$ is said to d-separate A and B if it blocks all paths between these two nodes.

D-separation is equivalent to conditional independence as follows (see also Figure 1).

Theorem 3 *(Pearl, 2000, Sec. 1.2) For any two variables A and B nodes in a DAG $\mathcal{G}$, for any subset $\mathbf{C}$ of nodes in the same graph, and for all joint probability distribution $P(\mathbf{v})$ for the variables in the graph, we have: $\mathbf{C}$ d-separates A and B if and only if $A \perp\!\!\!\perp B|\mathbf{C}$ for every $P(\mathbf{v})$ compatible with $\mathcal{G}$.*

3.2 Graphical specification of the Rubin model

In the Rubin model of Section 2.1, the notion of causality is implicit. The purpose with the model is to estimate the effect of a treatment (cause) on an outcome of interest; see Holland (1986). Furthermore, the variables which need to be controlled for are pre-treatment and pre-outcome variables. This ensures that the treatment and the outcome do not affect the control variables. The latter causality statements as well as the conditional independence assumptions (A.1) and (A.3) can be translated into two DAGs, see Figure 1, one for each potential outcome.

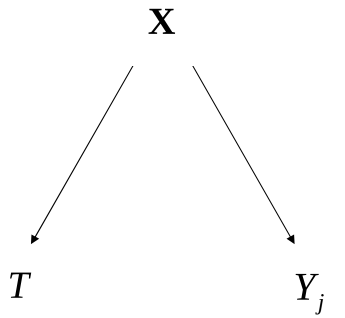

Figure 1: Rubin model displayed as a directed acyclic graph (DAG), $\mathcal{G}_j^R$, where $\mathbf{X}$ is a set of pre-treatment variables, T is the treatment assignment and Y_j, $j = 0, 1$ are the potential outcomes. Here, $\mathbf{X}$ d-separates Y_j from T and hence $Y_j \perp\!\!\!\perp T|\mathbf{X}$. Note that only Y_j is represented in the graph and not both Y_0 and Y_1 because the latter are "complementary" random variables: Either you are in the world where Y_0 is realized or in the world where Y_1 is realized; see Dawid (2002) on the dangers to contemplate the metaphysical joint distribution of Y_0 and Y_1.

We emphazise that we do not consider the DAGs of Figure 1 as equivalent to the Rubin model, but merely as a representation of some of its properties.

In the sequel, we work in the context where a set of pre-treatment variables, $\mathbf{X}$, is available, for which (A.1-A.4) hold. Some of these variables affect the outcomes and/or the treatment. It is, however, unknown which of the pre-treatment variables (if any) affect the outcomes and the treatment. Further, the pre-treatment variables may have an effect on each other in a non-specified fashion, though their causal structure is assumed to be compatible with a DAG.

Formally, let us denote by $\mathcal{G}_j^R$, $j = 0, 1$, the graphs described in Figure 1. We make the following assumption in the sequel:

(A.5) The unspecified arrows within $\mathbf{X}$ are assumed to be such that $\mathcal{G}_j^R$ is a DAG.

Because we do not have knowledge of the full causal structure between the variables involved, $\mathcal{G}_j^R$ is only a partially specified DAG. This is an essential characteristic of our framework since when all the causal relations are known, i.e. the DAG is fully specified, then the covariate selection problem (which as noted earlier is equivalent to finding a d-separator between Y_j and T) is solved (Back-door criterion in Pearl, 2000; see, also, Tian, Paz & Pearl 1998).

4 Theory for the identification of covariates

In Section 2 we saw that in order to estimate an average treatment effect we should match for a set of covariates for which assumptions (A.1-A.4) are fulfilled. We noted further that we should also avoid overmatching. Thus, we want to look for a set of covariates for which (A.1) and (A.3) hold, but such that these assumptions ceases to hold when discarding any of the covariates included in the set. This concept of minimality is formalized in Definition 4 below, followed with the theoretical framework which will allow us later to propose a procedure to identify a minimal set of covariates.

4.1 Minimal d-separators

Definition 4 *Given two nodes A and B in a DAG, $\mathcal{G}$, a set $\mathbf{C}$ that d-separates A from B in $\mathcal{G}$ is said to be minimal if no proper subset of $\mathbf{C}$ d-separates A from B.*

In Tian et al. (1998), this definition is shown to hold if and only if reducing $\mathbf{C}$ by a node cancels its d-separating property.

We now define some subsets of the covariates $\mathbf{X}$ in $\mathcal{G}_j^R$, which have the property of d-separating the treatment T from the potential outcome Y_j. We assume in the sequel that there is at least one path between T and Y_j.

Definition 5 *Let $\mathbf{X}_T$ be the set of nodes in $\mathbf{X}$ that are parents to the treatment variable T in $\mathcal{G}_j^R$, $j = 0, 1$. Let $\mathbf{X}_j$ be the set of nodes in $\mathbf{X}$ that are parents to the potential outcome variable Y_j in $\mathcal{G}_j^R$, $j = 0, 1$.*

The sets $\mathbf{X}_T$ and $\mathbf{X}_j$ in $\mathcal{G}_j^R$ both d-separate Y_j from T since all paths from T must go through $\mathbf{X}_T$ and all paths from Y_j must go through $\mathbf{X}_j$ and no nodes in $\mathbf{X}_T$ or $\mathbf{X}_j$ can unblock a path from T to Y_j. They are, however, not necessarily a minimal d-separator for T and Y_j.

Remark 6 *The set $\mathbf{X}_T$ has the property of d-separating T and $\mathbf{X} \setminus \mathbf{X}_T$ in $\mathcal{G}_j^R$. The corresponding property also holds for $\mathbf{X}_j$. Both of these properties are used later on for identification purposes.*

Definition 7 *Let $\mathbf{Q}_j \subseteq \mathbf{X}_T$ be a minimal d-separator for T and Y_j in $\mathcal{G}_j^R$, $j = 0, 1$.*

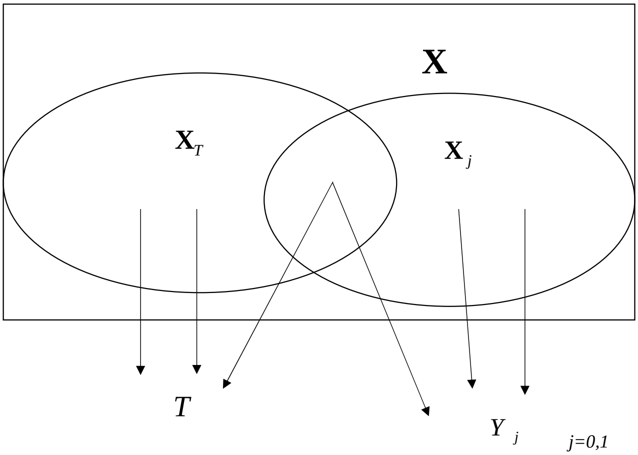

Figure 2: Illustration of the subsets introduced in Definitions 5, 7 and 8.

Definition 8 *Let $\mathbf{Z}_j \subseteq \mathbf{X}_j$ be a minimal d-separator for Y_j and T in $\mathcal{G}_j^R$, $j = 0, 1$.*

We have the following useful results.

Proposition 9 *For $j = 0, 1$, the sets $\mathbf{Q}_j$ and $\mathbf{Z}_j$ are unique as defined in Definition 7 and 8 respectively.*

Proof. We give the proof for $\mathbf{Q}_j$. Assume that there are two distinct subsets $\mathbf{A}$ and $\mathbf{B}$ of $\mathbf{X}_T$, $\mathbf{A} \neq \mathbf{B}$, that are minimal d-separators for T and Y_j. By the definition of minimality we know that $\mathbf{A} \not\subset \mathbf{B}$ and $\mathbf{B} \not\subset \mathbf{A}$. Hence, there must be $a \in \mathbf{A}$ such that $a \notin \mathbf{B}$. Since $a \in \mathbf{A}$ there must be a path from T to Y_j that is blocked by a (reducing $\mathbf{A}$ with a destroys d-separability). Since $a \notin \mathbf{B}$ the path from T to Y_j that is blocked by a must be blocked by some other node $b \in \mathbf{B}$ that is not in $\mathbf{A}$.

This is a contradiction because $\mathbf{X}_T$ contains only parents of T and, therefore, a path from Y_j to T that is blocked by a node in $\mathbf{X}_T$ cannot be blocked by another node in $\mathbf{X}_T$.

■

Corollary 10 *In $\mathcal{G}_j^R$, $\mathbf{Q}_j$ is a d-separator between Y_j and $\mathbf{X}_T \setminus \mathbf{Q}_j$, $j = 0, 1$. Moreover, if $\boldsymbol{\xi}_j \subseteq \mathbf{X}_T$ is a d-separator between Y_j and $\mathbf{X}_T \setminus \boldsymbol{\xi}_j$ of minimum cardinality then $\boldsymbol{\xi}_j = \mathbf{Q}_j$.*

Proof. Assume that there is a path from $\mathbf{X}_T \setminus \mathbf{Q}_j$ to Y_j that is not blocked by $\mathbf{Q}_j$. Since all nodes in $\mathbf{X}_T \setminus \mathbf{Q}_j$ are parents to T there is a path from T to Y_j not blocked by $\mathbf{Q}_j$, which is a contradiction.

To prove the second part of the corollary consider a set $\boldsymbol{\xi}_j$ of minimum cardinality such that it d-separates Y_j and $\mathbf{X}_T \setminus \boldsymbol{\xi}_j$. A path from Y_j to T has to go through either $\boldsymbol{\xi}_j$ or $\mathbf{X}_T \setminus \boldsymbol{\xi}_j$. Paths from Y_j to T that go through $\mathbf{X}_T \setminus \boldsymbol{\xi}_j$ are blocked by $\boldsymbol{\xi}_j$ since $\boldsymbol{\xi}_j$ d-separates Y_j and $\mathbf{X}_T \setminus \boldsymbol{\xi}_j$, and paths from Y_j to T that go through $\boldsymbol{\xi}_j$ are blocked by $\boldsymbol{\xi}_j$ (since $\boldsymbol{\xi}_j$ does not contain a collider in a path from T to Y_j). Hence, $\boldsymbol{\xi}_j$ d-separates Y_j and T.

Since $\boldsymbol{\xi}_j$ is a set of minimum cardinality in $\mathbf{X}_T$ such that it d-separates Y_j and $\mathbf{X}_T \setminus \boldsymbol{\xi}_j$ we know that $\boldsymbol{\xi}_j$ cannot be reduced by a node, thereby implying (by the first part of the corollary) that it is a minimal d-separator for T and Y_j. By Proposition 9, $\mathbf{Q}_j \subseteq \mathbf{X}_T$ is unique as a minimal d-separator for T and Y_j and, therefore, $\boldsymbol{\xi}_j = \mathbf{Q}_j$. ■

Corollary 11 *In $\mathcal{G}_j^R$, $\mathbf{Z}_j$ is a d-separator between T and $\mathbf{X}_j \backslash \mathbf{Z}_j$, $j = 0, 1$. Moreover, if $\boldsymbol{\xi}_j \subseteq \mathbf{X}_j$ is a d-separator for T and $\mathbf{X}_j \setminus \boldsymbol{\xi}_j$ of minimum cardinality then $\boldsymbol{\xi}_j = \mathbf{Z}_j$.*

Proof. Similar to Corollary 10 ■

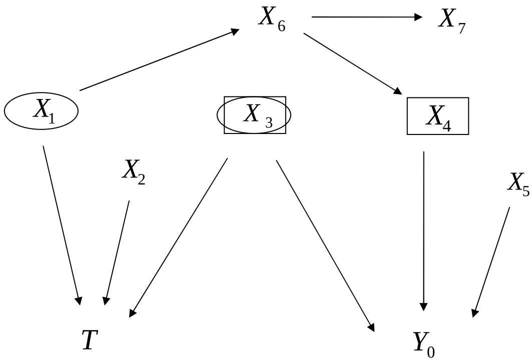

Figure 3: Graph with set $\mathbf{Q}_0$ marked with ellipses and $\mathbf{Z}_0$ marked with rectangles.

Example 12 *Figure 3 presents an illustrative example. Here $\mathbf{X}_T = \{X_1, X_2, X_3\}$ and $\mathbf{X}_0 = \{X_3, X_4, X_5\}$. The set $\mathbf{Q}_0 = \{X_1, X_3\}$ is a d-separating set between Y_0 and $\mathbf{X}_T \setminus \mathbf{Q}_0 = \{X_2\}$. Similarly $\mathbf{Z}_0 = \{X_3, X_4\}$ d-separates T from $\mathbf{X}_0 \setminus \mathbf{Z}_0 = \{X_5\}$. Both $\mathbf{Q}_0$ and $\mathbf{Z}_0$ are minimal d-separators for Y_0 and T.*

Note that in Example 12 the set $\{X_3, X_6\}$ is also a minimal d-separator for Y_0 and T. This illustrates the fact that the sets defined in Definition 7 and 8 are not the only existing d-separators. They are, however, unique as subsets of $\mathbf{X}_T$ and $\mathbf{X}_j$, and, most importantly, they are identifiable without making assumptions on the causal structure within $\mathbf{X}$ as we shall see in the following section.

4.2 Identification algorithms

In the previous section we have defined subsets of the covariates d-separating the treatment from the potential outcomes. By Theorem 3, all the definitions and results of the previous section specifying d-separators can be translated into conditional independence statements valid for all the distributions compatible with the graphs $\mathcal{G}_j^R$, $j = 0, 1$. In particular, we have from Definitions 7 and 8 that

$$Y_0 \perp\!\!\!\perp T|\mathbf{Q}_0 \text{ and } Y_1 \perp\!\!\!\perp T|\mathbf{Q}_1$$

and

$$Y_0 \perp\!\!\!\perp T|\mathbf{Z}_0 \text{ and } Y_1 \perp\!\!\!\perp T|\mathbf{Z}_1.$$

In other words, assumption (A.1) holds for the sets $\mathbf{Q}_0$ and $\mathbf{Z}_0$ and assumption (A.3) holds for the sets $\mathbf{Q}_1$ and $\mathbf{Z}_1$. These sets are minimal in the sense that the conditional independence would not hold anymore if one variable were taken away from them.

Our purpose is to identify these minimal sets. However, because the above conditional independence statements involve partially unobserved variables (the potential outcomes), they cannot be utilized. On the other hand, the identification can be achieved by the two algorithms proposed in Table 1.

Table 1: Identification of covariate sets $\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{Z}_0$ and $\mathbf{Z}_1$.

Algorithm A: Identification of $\mathbf{Q}_0$ and $\mathbf{Q}_1$	
Step 1.	Identify $\mathbf{X}_T$ such that $(T \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_T|\mathbf{X}_T)$ holds.
Step 2.	For $j = 0, 1$:
	Identify $\mathbf{Q}_j \subseteq \mathbf{X}_T$ such that $(Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j|\mathbf{Q}_j, T = j)$ holds.
Algorithm B: Identification of $\mathbf{Z}_0$ and $\mathbf{Z}_1$	
For $j = 0, 1$:	
Step 1.	Identify $\mathbf{X}_j$ such that $(Y_j \perp\!\!\!\perp \mathbf{X} \setminus \mathbf{X}_j|\mathbf{X}_j, T = j)$ holds.
Step 2.	Identify $\mathbf{Z}_j \subseteq \mathbf{X}_j$ such that $(T \perp\!\!\!\perp \mathbf{X}_j \setminus \mathbf{Z}_j|\mathbf{Z}_j)$ holds.

The identifiability of $\mathbf{X}_T$ in Step 1 of Algorithm A is a consequence of the properties of the set $\mathbf{X}_T$ described in Remark 6, while the identifiability of $\mathbf{Q}_0$ and $\mathbf{Q}_1$ in Step 2 is due to Corollary 10 and to the fact that, for $j = 0, 1$,

$$Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j|\mathbf{Q}_j \Leftrightarrow Y_j \perp\!\!\!\perp \mathbf{X}_T \setminus \mathbf{Q}_j|\mathbf{Q}_j, T = j. \tag{5}$$

This result allows us to specify an algorithm based solely on observed variables, since given $T = j$, Y_j is observed.

Similarly, for Algorithm B, the identifiability of $\mathbf{X}_j$, $j = 0, 1$, in Step 1 is due to the properties of the set $\mathbf{X}_j$ described in Remark 6 and to (5) where $\mathbf{Q}_j$ is replaced by $\mathbf{X}_j$ and $\mathbf{X}_T$ is replaced by $\mathbf{X}$, while $\mathbf{Z}_0$ and $\mathbf{Z}_1$ are identified in Step 2 by Corollary 11.

In Section 5 we discuss how both algorithms can be implemented with data.

4.3 Parameters of interest

As noted in Section 2.1 the typical parameter of interest is the average treatment effect $E(Y_1 - Y_0)$. Using the minimal d-separators identified with Algorithms A and B, this effect can be identified as follows.

$$\begin{aligned} E(Y_1 - Y_0) &= E(E(Y_1 \mid \mathbf{V}_1)) - E(E(Y_0 \mid \mathbf{V}_0)) \\ &= E(E(Y_1 \mid T = 1, \mathbf{V}_1)) - E(E(Y_0 \mid T = 0, \mathbf{V}_0)), \end{aligned}$$

where either $\mathbf{V}_1 = \mathbf{Q}_1$ and $\mathbf{V}_0 = \mathbf{Q}_0$ or $\mathbf{V}_1 = \mathbf{Z}_1$ and $\mathbf{V}_0 = \mathbf{Z}_0$.

In some applications, it is of interest to study the impact of a treatment on the subpopulation consisting of the treated units. The parameter of interest is then the average treatment effect on the treated, $E(Y_1 - Y_0 \mid T = 1)$. Then, under assumptions (A.1-A.2),

$$\begin{aligned} E(Y_1 - Y_0 \mid T = 1) &= E(Y_1 | T = 1) - E(E(Y_0 | T = 1, \mathbf{V}_0) | T = 1) \\ &= E(Y_1 | T = 1) - E(E(Y_0 | T = 0, \mathbf{V}_0) | T = 1), \end{aligned} \tag{6}$$

and we only need to identify $\mathbf{V}_0 = \mathbf{Q}_0$ or $\mathbf{V}_0 = \mathbf{Z}_0$.

5 Practical implementation

Assume that we have a random sample of individuals for which the variables of the graph $\mathcal{G}_j^R$ have been observed. We can think of two main approaches to implement Algorithms A and B based on such data. Note that because the final purpose is to estimate a treatment effect non-parametricaly, the identification of the covariates must be done by avoiding distributional assumptions where possible.

The first and most general approach consists in using a non-parametric test of conditional independence. Recent results have been obtained in this area by Su & White (2003), who propose an empirical likelihood ratio procedure. For instance, Step 1 of Algorithm B could in theory be implemented by examination of all possible subsets $\boldsymbol{\xi}$ of $\mathbf{X}$, and selection of the subset of minimum cardinality for which the null hypothesis $Y_0 \perp\!\!\!\perp \mathbf{X} \setminus \boldsymbol{\xi} | \boldsymbol{\xi}, T = 0$ is not rejected. There are two main problems with this procedure when the cardinality of $\mathbf{X}$ is large (which is the typical case in applications). Because many subsets of covariates must be visited it is difficult to keep control of the overall size of the testing procedure. Moreover, the non-parametric test cited above is in fact applicable only to sets of low dimension due to the curse of dimensionality (Bellman 1961), as noted by Su & White (2003).

The second approach is obtained by making one important simplification. We assume that conditional independence in the mean is sufficient for conditional independence in distribution,

$$E(A|B,C) = E(A|B) \Rightarrow A \perp\!\!\!\perp C|B,$$

for all the conditional independence statements of Table 1.

While this is a restrictive assumption in general, it is not restrictive neither for Step 1, Algorithm A, nor for Step 2, Algorithm B, because the variable T is binary.

Let us consider Step 1 of Algorithm B as an illustration. By the above assumption, this step is simplified to

$$\text{Identify } \mathbf{X}_0 \text{ such that } E(Y_0|\mathbf{X}, T=0) = E(Y_0|\mathbf{X}_0, T=0) \text{ holds.}$$

The identification of $\mathbf{X}_0$ is then a usual covariate selection issue in a regression framework $Y_0 = g(\mathbf{X}_0, T=0) + \varepsilon$. We do not want to make model assumptions, and, hence, the function $E(Y_0|\mathbf{X}_0, T=0) = g(\mathbf{X}_0, T=0)$ should be estimated non-parametricaly. However, because of the large number of covariates typically available and to avoid the curse of dimensionality, we suggest the use of a regression function polynomial in the covariates (linear in the parameters). Other possibilities exist, including generalized additive models, projection pursuit regression, etc.; see, e.g., Hastie & Tibshirani (1990) and the references therein. The distribution of the error term ε can also be left unspecified by using, e.g., least squares. There is a large literature on how to select covariates in such a regression framework, see, e.g., Miller (1990), McQuarrie & Tsai (1998) and Burnham & Anderson (2000). Scoring methods such as Akaike's information criteria, cross-validation, and many others, are most popular because they avoid the multiple testing problem. With such methods, a score is computed for each candidate model, and the model with maximum score is preferred.

The above discussion is valid for all the steps in Algorithms A and B, thereby yielding the translation of Algorithms A and B given in Table 2.

6 Application: estimation of a social program effect

6.1 The Lalonde data

We use a data set first analyzed by Lalonde (1986). Lalonde studied data from the National Supported Work Demonstration (NSW) where 297 individuals were assigned to participate in a training program and 425 were assigned to be controls. The NSW program was operated in ten locations across the United States. The intention of the training that included counselling and work experience was to provide means for workers without job skills to get access to the labour market.

Table 2: Identification of covariates sets $\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{Z}_0$ and $\mathbf{Z}_1$.

Algorithm A': Identification of $\mathbf{Q}_0$ and $\mathbf{Q}_1$	
Step 1.	For all $\boldsymbol{\xi} \subseteq \mathbf{X}$, fit a model for $E\left[T\|\boldsymbol{\xi}\right]$, and select $\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}$ which yields maximum score.
Step 2.	For $j = 0, 1$: For all $\boldsymbol{\xi} \subseteq \hat{\boldsymbol{\xi}}$, fit a model for $E\left[Y_j\|\boldsymbol{\xi}, T = j\right]$, and select $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}$ which yields maximum score.
Algorithm B': Identification of $\mathbf{Z}_0$ and $\mathbf{Z}_1$	
	For $j = 0, 1$:
Step 1.	For all $\boldsymbol{\xi} \subseteq \mathbf{X}$, fit a model for $E\left[Y_j\|\boldsymbol{\xi}, T = j\right]$, and select $\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}_j$ which yields maximum score.
Step 2.	For all $\boldsymbol{\xi} \subseteq \hat{\boldsymbol{\xi}}_j$, fit a model for $E\left[T\|\boldsymbol{\xi}\right]$, and select $\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}_j$ which yields maximum score.

The participation in the program was randomized and, therefore, the effect of the program on income could be estimated by the difference in mean income between participants and non-participants. Two additional data sets consisting of non-experimental groups not participating in a training program were used for comparisons to the original experiment: the Panel Study of Income Dynamics (PSID) and Westat's Matched Current Population Survey-Social Security Administration File (CPS). Lalonde (1986) extracted subsets from the PSID and CPS data that were similar to the experimental group in some of the characteristics. The subsets are referred to as CPS 2, CPS 3, PSID 2 and PSID 3 in the following data analysis. The data analysed in this section is a subset of the NSW experimental group of the Lalonde data set that was selected by Dehejia & Wahba (1999, 2002) and also studied in Smith & Todd (2005), and Abadie & Imbens (2002). Dehejia & Wahba (1999) motivates the use of this subset instead of the full experimental treatment group by the fact that an additional variable (earnings 1974) can be taken into account that was not available for the full sample. In total there are ten covariates: age (years), education (years in school), black (0 or 1), hispanic (0 or 1), married (0 or 1), no high-school degree (0 or 1), earnings 1974 ($), earnings 1975 ($), unemployed 1974 (0 or 1), unemployed 1975 (0 or 1). The effect of the program is measured on the response variable earnings 1978 ($). For a more detailed description of the variables see Lalonde (1986) or Dehejia & Wahba (1999). The data is available at `http://www.columbia.edu/ rd247/nswdata.html.`

6.2 Description of the study

The main purpose of this case study is to illustrate the use of our identification algorithms for the estimation of treatment effects with observational data. We, therefore, ignore the randomized control group. We have, on the other hand, six different non-randomized control groups which leads us to conduct six different analyses. In line with the previous studies on the Lalonde data we aim at estimating the average treatment effect on the treated, $E(Y_1 - Y_0|T = 1)$. As a consequence, merely $\mathbf{Z}_0$ or $\mathbf{Q}_0$ need to be identified, see Section 4.3.

6.2.1 Matching

We match each treated unit to a control unit in a given control group. For each treated unit a subset of controls with the same values on the indicator covariates is selected. The unit is then matched with the control that is closest in terms of the remaining continuous covariates. To measure the closeness between a treated and a control unit we use the Mahalanobis distance see, e.g., Gu & Rosenbaum (1993).

$$d(\mathbf{X}_t, \mathbf{X}_c) = (\mathbf{X}_t - \mathbf{X}_c)^T \mathbf{S}^{-1} (\mathbf{X}_t - \mathbf{X}_c),$$

where $\mathbf{X}_t$ is the covariate vector for the treated unit, $\mathbf{X}_c$ is the covariate vector for the control unit and $\mathbf{S}$ is the pooled sample covariance matrix. Matching is done with replacement so each control unit can be used as a match more than once which increases the set of potential matches and thereby the matching quality. The treatment effect of the treated is estimated as the mean of the differences in the response variable between the treated and the matched units.

6.2.2 Covariate selection

We want to control on a set of covariates which is as small as possible. We, therefore, estimate the parameter of interest by first identifying the sets $\mathbf{Q}_0$ and $\mathbf{Z}_0$ by means of the algorithms in Table 2. Note that Dehejia & Wahba (1999) used a set of covariates which were balanced for treated and controls. In $\mathcal{G}_0^R$, this corresponds to adjusting for the set $\mathbf{X}_T$.

In order to implement the algorithms we need to regress T and $Y_j|T = j$ on the covariates. When T is the response, we use logistic regression. We allow for non-linearity by using a polynomial logistic regression, that is

$$\log\left(\frac{P(T = 1|\mathbf{X})}{1 - P(T = 1|\mathbf{X})}\right) = f(\mathbf{x}),$$

where $f(\mathbf{x})$ is a second order polynomial function. The response $Y_j|T = j$ (earnings 1978) is split into two variables; U_j (one if zero earnings and zero otherwise) and $W_j \equiv (Y_j|T = j, U_j = 1)$. A two step regression is then performed to identify the covariates affecting $Y_j|T = j$. First a second order polynomial logistic regression is used to explain U_j. Then, a second order polynomial regression is used to fit lnW_j (modelling income on the log scale is customary in the related literature). The covariates found to explain U_j and W_j are pooled together as the variates explaining $Y_j|T = j$.

A covariate selection procedure is needed in all the regressions described above in order to implement the algorithms of Table 2. We use a forward stepwise procedure where the AIC criterion is used to enter covariates. More precisely, we use the function `stepAIC` available in the software R (R Development Core Team 2004). All our computations are performed with R.

The reservoir of variables available for selection are the ten covariates from the Dehejia & Wahba subset of the Lalonde data described in Section 6.1.

6.3 Results

We focus our presentation and discussion on the results of the study performed with the CPS 3 control group. We report in the Appendix (Tables 5-8) the results for the other groups. The units in CPS 3 are a subset of CPS 1 consisting of all unemployed in 1976 whose income in 1975 was below the poverty level. We have 185 treated units and 429 controls.

Table 3 shows the covariates selected by the different steps of the algorithms A' and B', and the resulting estimated treatment effects on the treated. Moreover, mean and median Mahalanobis distances for both steps of algorithm A' and B' are reported. For these four covariate sets the treatment effect of the treated varies from 45 to 1646. More interesting is that the mean Mahalanobis distance decreases from 0.125 to 0.107 and from 0.563 to 0.133 when matching for the smaller sets $\mathbf{Q}_0$ and $\mathbf{Z}_0$ instead of $\mathbf{X}_T$ and $\mathbf{X}_0$ respectively. Thus, as expected, by matching for less variables better matches are obtained. Moreover, one may here prefer $\mathbf{Q}_0$ to $\mathbf{Z}_0$ since it provides better matches (compare 0.107 to 0.133).

In Table 4 we display the sample means and the standardized bias for the different covariates calculated as $\bar{x}_d/\sqrt{\left(s_t^2 + s_c^2\right)/2}$ where $\bar{x}_d$ is the mean difference in the covariate between the treated and control group and s_t^2 and s_c^2 are the variances within the treated and control groups before matching. When comparing the standardized bias we see that the quality of the matches improves for all the variables selected by the algorithms.

The sets of covariates selected by the two algorithms are different. This is to be expected in general. Drawing a graph representing the selection procedures as in Figure 4 provides a useful diagnostic tool to evaluate the sets $\mathbf{Q}_0$ and $\mathbf{Z}_0$ obtained. For instance, we see that algorithm A' does not select the variable RE75 in $\mathbf{Z}_0$ although according to the graph it should be in a d-separator for T and Y_0. Again $\mathbf{Q}_0$ should be preferred to $\mathbf{Z}_0$, since Figure 4 indicates that $\mathbf{Z}_0$ may lead to undermatching (two few covariates are matched for). This discussion highlights the usefulness of running both algorithms and of looking at all information obtained.

Table 3: Covariates $\mathbf{X}_T$, $\mathbf{Q}_0$, $\mathbf{X}_0$ and $\mathbf{Z}_0$ selected by algorithms A' (step 1 and 2) and B' (step 1 and 2) respectively, with resulting mean and median Mahalanobis distances, treatment effects on the treated (TT). Standard deviations divided by $\sqrt{185}$ given in parentheses.

	CPS 3			
Covariate	$\mathbf{X}_T$	$\mathbf{Q}_0$	$\mathbf{X}_0$	$\mathbf{Z}_0$
age			√	√
education			√	√
married	√	√		
black	√	√	√	√
hispanic	√			
nodegree				
RE74			√	
RE75	√	√	√	
U74	√	√	√	√
U75				
Mean distance	0.125	0.107	0.563	0.133
Median distance	0	0	0.170	0.042
TT	257	563	45	1646
	(728)	(712)	(740)	(622)

Table 4: Sample means and standardized bias before and after matching for control group CPS 3 (matched variables highlighted in bold).

	Sample means			Standardized bias			
Covariate	Treated	Control	Pre-matching	$\mathbf{X}_T$	$\mathbf{Q}_0$	$\mathbf{X}_0$	$\mathbf{Z}_0$
age	25.82	28.03	-0.242	-0.020	0	**0.040**	**-0.011**
education	10.35	10.24	0.045	0.022	0.018	**-0.059**	**-0.015**
married	0.84	0.20	1.668	**0**	**0**	-0.204	-0.324
black	0.06	0.14	-0.277	**0**	**0**	**0**	**0**
hispanic	0.19	0.51	-0.720	**0**	0.127	0.054	0.018
nodegree	0.71	0.60	0.235	0.160	0.103	0.148	0.137
RE74	2096	5619	-0.596	0.015	0.041	**0.042**	0.004
RE75	1532	2466	-0.287	**0.070**	**0.061**	**0.113**	0.018
U74	0.71	0.26	-0.998	**0**	**0**	**0**	**0**
U75	0.60	0.31	-0.602	0	0	0.011	-0.158

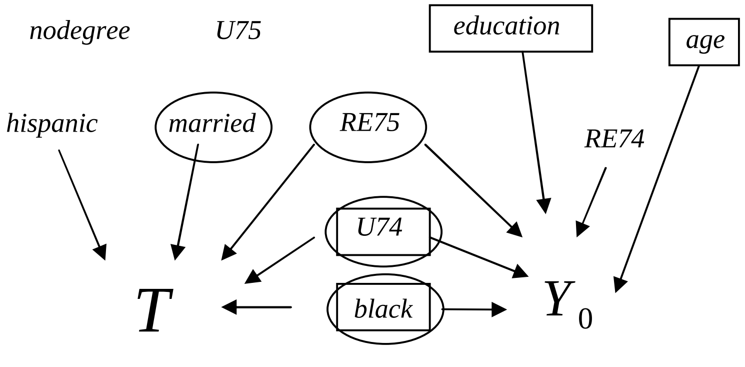

Figure 4: Graph illustrating the results from algorithm A' and B' with control group CPS 3. Covariates with arrows pointing towards T represent $\mathbf{X}_T$, and covariates with arrows pointing towards Y_0 represent $\mathbf{X}_0$. Selected set $\mathbf{Q}_0$={married, black, RE75, U74} marked with ellipses and $\mathbf{Z}_0$={age, education, black, U74} marked with rectangles.

7 Discussion

Balancing all observed covariates, e.g., by matching, between the treated and the control groups in observational studies is frequently encouraged in the literature. However, all covariates do not need to be balanced for, and balancing for unnecessary covariates (overmatching) has a cost in terms of increased bias. An ad-hoc covariate selection procedure was proposed in Cochran (1965), see also Rosenbaum (2002, pp.77), where the need for looking also at how covariates affect the outcome was recognized.

The theoretical results of this paper emphasize the need for methods looking both at how the covariates affect the treatment assignment (balancing property) and how they affect the outcome. Furthermore, we are able to propose identification algorithms which are formally justified.

Of the two algorithms proposed, only one may be used in practice. However, running both of the two algorithms is advantageous since it allows us to do a cross-checking of the results. The d-separating sets obtained by the two algorithms can be evaluated by drawing graphs corresponding to the one in Figure 4. This evaluation can help to decrease the risk of undermatching (using too few covariates). Of course, subject matter information, when available, must be used when determining the covariates to be matched for.

The theoretical results of this paper have been deduced under assumption (A.5). We believe this assumption to be over-restrictive and conjecture that it is sufficient to assume that the unspecified arrows within $\mathbf{X}$ be such that $\mathcal{G}_j^R$ is a chain graph; see, e.g., Lauritzen (1996).

References

Abadie, A. & Imbens, G. (2002), 'Simple and bias-corrected matching estimators for average treatment effects', *NBER Technical Working paper* (No. 283). National Bureau of Economic Research, Cambridge Massachusetts.

Bellman, R. E. (1961), *Adaptive Control Processes*, Princeton University Press, Princeton.

Burnham, K. & Anderson, D. (2000), *Model Selection and Inference: A practical Information-Theoretic Approach*, Springer-Verlag, New York.

Cochran, W. (1965), 'The planning of observational studies of human populations (with discussion)', *Journal of the Royal Statistical Society Series A* **128**, 134–155.

Cochran, W. & Rubin, D. B. (1973), 'Controlling bias in observational studies: A review', *Sankhya Series A* **35**, 417–446.

Dawid, A. P. (1979), 'Conditional independence in statistical theory', *Journal of the Royal Statistical Society Series B* **41**, 1–31.

Dawid, A. P. (2002), 'Influence diagrams for causal modelling and inference', *International Statistical Review* **70**, 161–189.

Dehejia, R. & Wahba, S. (1999), 'Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs', *Journal of the American Statistical Association* **94**, 1053–1062.

Dehejia, R. & Wahba, S. (2002), 'Propensity score matching methods for nonexperimental causal studies', *The Review of Economics and Statistics* **84**, 151–161.

Gu, X. S. & Rosenbaum, P. R. (1993), 'Comparison of multivariate matching methods: Structures distances and algortihms', *Journal of Computational and Graphical Statistics* **2**, 405–20.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Heckman, J. & Navarro-Lozano, S. (2004), 'Using matching, instrumental variables and control functions to estimate economic choice models', *The Review of Economics and Statistics* **86**, 30–57.

Holland, P. W. (1986), 'Statistics and causal inference', *Journal of the American Statistical Association* **81**, 945–960.

Imbens, G. W. (2004), 'Nonparametric estimation of average treatment effects under exogeneity: A review', *The Review of Economics and Statistics* **86**, 4–29.

Lalonde, R. J. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *American Economic Review* **76**, 604–620.

Lauritzen, S. (1996), *Graphical Models*, Oxford University Press, Oxford.

Lauritzen, S. (2001), 'Causal inference from graphical models', In Barndorff-Nielsen, O.E., Cox, D. R. and Klüppelberg, C. eds. ,*Complex Stochastic Systems*, London: Chapman and Hall, pp. 63-107.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. & Leimer, H.-G. (1990), 'Independence properties of directed markov fields', *Networks* **20**, 491–505.

McQuarrie, A. & Tsai, C. (1998), *Regression and Time Series Model Selection*, World Scientific, River Edge, NJ.

Miller, A. (1990), *Subset Selection in Regression (2nd edition)*, Chapman and Hall, New York.

Pearl, J. (2000), *Causality*, Cambridge University Press, Cambridge.

R Development Core Team (2004), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org*

Rosenbaum, P. R. (2002), *Observational Studies (2nd edition)*, Springer, New York.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**, 688–701.

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.

Smith, J. A. & Todd, P. E. (2005), 'Does matching overcome Lalonde's critique of nonexperimental estimators?', *Journal of Econometrics* **125**, 305–353.

Su, L. & White, H. (2003), Testing conditional independence via empirical likelihood. Paper 2003-14. Department of Economics, UCSD.

Tian, J., Paz, A. & Pearl, J. (1998), Finding minimal D-separators, Technical Report 980007. UCLA Computer Science Department. Los Angeles: University of California.

Verma, T. & Pearl, J. (1990), 'Causal networks: semantics and expressiveness', In R.D. Schachter, T.S. Levitt, L. N. Kanal, and J.F. Lemmer eds., *Uncertainty in artificial intelligence IV*, North-Holland: Amsterdam, pp. 69-76.

A More tables

We display here the results obtained with all the non-randomized control groups available as described in Section 6.

Table 5: Covariates $\mathbf{X}_T$ and $\mathbf{Q}_0$ selected by algorithm A' (step 1 and 2) with resulting mean and median Mahalanobis distances, treatment effects on the treated (TT) for CPS subsets. Standard deviations divided by $\sqrt{185}$ given in parentheses.

	CPS 1		CPS 2		CPS 3	
Covariate	$\mathbf{X}_T$	$\mathbf{Q}_0$	$\mathbf{X}_T$	$\mathbf{Q}_0$	$\mathbf{X}_T$	$\mathbf{Q}_0$
age	√	√				
education	√	√				
married	√	√	√	√	√	√
black	√	√	√	√	√	√
hispanic	√		√		√	
nodegree	√	√				
RE74	√	√				
RE75	√	√	√	√	√	√
U74	√	√	√	√	√	√
U75						
Mean distance	0.211	0.182	0.006	0.004	0.125	0.107
Median distance	0.042	0.041	0	0	0	0
TT	2099	2335	661	640	257	563
	(695)	(701)	(753)	(749)	(728)	(712)

Table 6: Covariates $\mathbf{X}_0$ and $\mathbf{Z}_0$ selected by algorithm B' (step 1 and 2) with resulting mean and median Mahalanobis distances, treatment effects on the treated (TT) for CPS subsets. Standard deviations divided by $\sqrt{185}$ given in parentheses.

	CPS 1		CPS 2		CPS 3	
Covariate	$\mathbf{X}_0$	$\mathbf{Z}_0$	$\mathbf{X}_0$	$\mathbf{Z}_0$	$\mathbf{X}_0$	$\mathbf{Z}_0$
age	√	√	√		√	√
education	√	√	√		√	√
married	√	√	√	√		
black	√	√	√	√	√	√
hispanic			√	√		
nodegree	√	√	√			
RE74	√	√	√		√	
RE75	√	√	√	√	√	
U74	√	√	√	√	√	√
U75			√			
Mean distance	0.182	0.182	NA*	0.006	0.563	0.133
Median distance	0.041	0.041	NA*	0	0.170	0.042
TT	2335	2335	NA*	661	45	1646
	(701)	(701)		(753)	(740)	(622)

*Not available since exact matching for indicator variables was not possible.

Table 7: Covariates $\mathbf{X}_T$ and $\mathbf{Q}_0$ selected by algorithm A' (step 1 and 2) with resulting mean and median Mahalanobis distances, treatment effects of the treated (TT)for PSID subsets. Standard deviations divided by $\sqrt{185}$ given in parentheses.

	PSID 1		PSID 2		PSID 3	
Covariate	$\mathbf{X}_T$	$\mathbf{Q}_0$	$\mathbf{X}_T$	$\mathbf{Q}_0$	$\mathbf{X}_T$	$\mathbf{Q}_0$
age	√	√	√	√	√	√
education	√	√	√	√		
married	√	√	√	√	√	
black	√	√	√		√	√
hispanic	√	√	√		√	
nodegree					√	
RE74						
RE75	√	√	√	√		
U74	√	√	√	√	√	√
U75	√	√	√	√	√	√
Mean distance	0.737	0.737	NA*	0.858	NA*	0.099
Median distance	0.172	0.172	NA*	0.217	NA*	0.010
TT	2373	2373	NA*	2276	NA*	1635
	(741)	(741)		(737)		(757)

*Not available since exact matching for indicator variables was not possible.

Table 8: Covariates $\mathbf{X}_0$ and $\mathbf{Z}_0$ selected by algorithm B' (step 1 and 2) with resulting mean and median Mahalanobis distances, treatment effects of the treated (TT) for PSID subsets. Standard deviations for PSID divided by $\sqrt{185}$ given in parentheses.

	PSID 1		PSID 2		PSID 3	
Covariate	$\mathbf{X}_0$	$\mathbf{Z}_0$	$\mathbf{X}_0$	$\mathbf{Z}_0$	$\mathbf{X}_0$	$\mathbf{Z}_0$
age	√	√	√	√	√	√
education	√	√	√	√		
married	√	√	√	√		
black	√	√			√	√
hispanic	√	√				
nodegree	√					
RE74	√		√	√	√	
RE75	√	√	√	√		
U74	√	√			√	√
U75	√	√	√		√	√
Mean distance	NA*	0.737	0.203	0.141	0.174	0.099
Median distance	NA*	0.172	0.132	0.114	0.041	0.010
TT	NA*	2373	2236	1579	2099	1635
		(741)	(747)	(739)	(700)	(757)

*Not available since exact matching for indicator variables was not possible.

Publication series published by the Institute for Labour Market Policy Evaluation (IFAU) – latest issues

Rapporter/Reports

2004:1 Björklund Anders, Per-Anders Edin, Peter Fredriksson & Alan Krueger “Education, equality, and efficiency – An analysis of Swedish school reforms during the 1990s”

2004:2 Lindell Mats “Erfarenheter av utbildningsreformen Kvalificerad yrkesutbildning: ett arbetsmarknadsperspektiv”

2004:3 Eriksson Stefan & Jonas Lagerström ”Väljer företag bort arbetslösa jobbsökande?

2004:4 Forslund Anders, Daniela Fröberg & Linus Lindqvist ”The Swedish activity guarantee”

2004:5 Franzén Elsie C & Lennart Johansson “Föreställningar om praktik som åtgärd för invandrares integration och socialisation i arbetslivet”

2004:6 Lindqvist Linus ”Deltagare och arbetsgivare i friårsförsöket”

2004:7 Larsson Laura ”Samspel mellan arbetslöshets- och sjukförsäkringen”

2004:8 Ericson Thomas ”Personalutbildning: en teoretisk och empirisk översikt”

2004:9 Calmfors Lars & Katarina Richardson ”Marknadskrafterna och lönebildningen i landsting och regioner”

2004:10 Dahlberg Matz & Eva Mörk ”Kommunanställda byråkraters dubbla roll”

2004:11 Mellander Erik, Gudmundur Gunnarsson & Eleni Savvidou ”Effekter av IT i svensk industri”

2004:12 Runeson Caroline ”Arbetsmarknadspolitisk översikt 2003”

2004:13 Nordström Skans Oskar ”Har ungdomsarbetslöshet långsiktiga effekter?”

2004:14 Rooth Dan-Olof & Olof Åslund ”11 september och etnisk diskriminering på den svenska arbetsmarknaden”

2004:15 Andersson Pernilla & Eskil Wadensjö ”Hur fungerar bemanningsbranschen?”

2004:16 Lundin Daniela ”Vad styr arbetsförmedlarna?”

2004:17 Forslund Anders, Per Johansson & Linus Lindqvist ”Anställningsstöd – en väg från arbetslöshet till arbete?”

2004:18 Jönsson Annelis & Lena Rubinstein Reich ”Invandrade akademiker som lärare i den svenska skolan”

Working Papers

2004:1 Frölich Markus, Michael Lechner & Heidi Steiger “Statistically assisted programme selection – International experiences and potential benefits for Switzerland”

2004:2 Eriksson Stefan & Jonas Lagerström “Competition between employed and unemployed job applicants: Swedish evidence”

2004:3 Forslund Anders & Thomas Lindh “Decentralisation of bargaining and manufacturing employment: Sweden 1970–96”

2004:4 Kolm Ann-Sofie & Birthe Larsen “Does tax evasion affect unemployment and educational choice?

2004:5 Schröder Lena “The role of youth programmes in the transition from school to work”

2004:6 Nilsson Anna “Income inequality and crime: The case of Sweden”

2004:7 Larsson Laura & Oskar Nordström Skans “Early indication of program performance: The case of a Swedish temporary employment program”

2004:8 Larsson Laura “Harmonizing unemployment and sickness insurance: Why (not)?”

2004:9 Cantoni Eva & Xavier de Luna “Non-parametric adjustment for covariates when estimating a treatment effect”

2004:10 Johansson Per & Mårten Palme “Moral hazard and sickness insurance: Empirical evidence from a sickness insurance reform in Sweden”

2004:11 Dahlberg Matz & Eva Mörk “Public employment and the double role of bureaucrats”

2004:12 van den Berg Gerard J, Maarten Lindeboom & Peter J Dolton “Survey non-response and unemployment duration”

2004:13 Gunnarsson Gudmundur, Erik Mellander & Eleni Savvidou “Human capital is the key to the IT productivity paradox”

2004:14 Nordström Skans Oskar “Scarring effects of the first labour market experience: A sibling based analysis”

2004:15 Ericson Thomas ”The effects of wage compression on training: Swedish empirical evidence”

2004:16 Åslund Olof & Dan-Olof Rooth ”Shifting attitudes and the labor market of minorities: Swedish experiences after 9-11”

2004:17 Albrecht James, Gerard J van den Berg & Susan Vroman "The knowledge lift: The Swedish adult education program that aimed to eliminate low worker skill levels"

2004:18 Forslund Anders, Per Johansson & Linus Lindqvist "Employment subsidies – A fast lane from unemployment to work?"

2004:19 Zijl Marloes, Gerard J van den Berg & Arjan Heyma "Stepping-stones for the unemployed: The effect of temporary jobs on the duration until regular work"

2005:1 Ericson Thomas "Personnel training: a theoretical and empirical review"

2005:2 Lundin Martin "Does cooperation improve implementation? Central-local government relations in active labour market policy in Sweden"

2005:3 Carneiro Pedro, James J Heckman & Dimitriy V Masterov "Labor market discrimination and racial differences in premarket factors"

2005:4 de Luna Xavier & Ingeborg Waernbaum "Covariate selection for non-parametric estimation of treatment effects"

Dissertation Series

2002:1 Larsson Laura "Evaluating social programs: active labor market policies and social insurance"

2002:2 Nordström Skans Oskar "Labour market effects of working time reductions and demographic changes"

2002:3 Sianesi Barbara "Essays on the evaluation of social programmes and educational qualifications"

2002:4 Eriksson Stefan "The persistence of unemployment: Does competition between employed and unemployed job applicants matter?"

2003:1 Andersson Fredrik "Causes and labor market consequences of producer heterogeneity"

2003:2 Ekström Erika "Essays on inequality and education"