PROGRAMME EVALUATION WITH MULTIPLE TREATMENTS

Markus Frölich

Universität St.Gallen

Abstract. This paper reviews the main identification and estimation strategies for microeconometric policy evaluation. Particular emphasis is laid on evaluating policies consisting of multiple programmes, which is of high relevance in practice. For example, active labour market policies may consist of different training programmes, employment programmes and wage subsidies. Similarly, sickness rehabilitation policies often offer different vocational as well as non-vocational rehabilitation measures. First, the main identification strategies (control-for-confounding-variables, difference-in-difference, instrumental-variable, and regression-discontinuity identification) are discussed in the multiple-programme setting. Thereafter, the different nonparametric matching and weighting estimators of the average treatment effects and their properties are examined.

Keywords. Evaluation; Matching; Treatment effect; Unobservables; Covariate-adjustment; Difference-in-difference

1. Introduction

Programme evaluation is an important tool for informed decision-making with respect to the efficient allocation of resources and for the improvement of existing policies.[1] Evaluation attempts to assess how far a programme has achieved its intended aims. Consider an active labour market programme, such as a vocational training programme for unemployed persons. A principal aim of the programme is to improve the employability of the participants so that they quickly become re-employed and earn higher wages. To assess how far this aim has been fulfilled, it would be necessary to compare the employment situation of the participants after the programme with the employment status they would have had, if they had not participated in the programme. A naive comparison of their employment status before and after the programme is not informative since some of the participants would have found employment even without the programme. A comparison of the re-employment rates of the participants and of the non-participants is futile, too, unless the participants were chosen randomly. Because programme participation is often the result of deliberate decisions, the individuals who decided or were assigned to participate are a selected group, such that a direct comparison of their outcomes with those of the non-participants would lead to selection bias. Instead, a variety of statistical corrections are necessary to

0950-0804/04/02 0181–44 JOURNAL OF ECONOMIC SURVEYS Vol. 18, No. 2

compare only like persons and to identify the average treatment effect, which is the difference between the expected outcome in the case of participation and the expected outcome in the case of non-participation. Whereas earlier evaluation studies often employed parametric selection models, these are increasingly replaced by nonparametric methods that avoid strong functional form assumptions and are, thus, more robust to misspecification. The applicability of these methods depends on the available data and on the specific details of the programme, particularly on the way how programme participation decisions were made.

Whereas most of the literature on programme evaluation has focused on the evaluation of a single programme (see, for example, the surveys by Angrist and Krueger (1999), Heckman *et al*. (1999)), many social policies consist of a variety of different programmes. Active labour market policies, for example, usually comprehend job-search assistance, vocational training programmes, public employment programmes, wage subsidies etc. Evaluating such a diverse policy requires the identification and estimation of many different treatment effects, which makes the analysis more complex. Proper evaluations of policies with multiple programmes have only been carried out recently. This survey gives an overview of possible identification and estimation strategies for the evaluation of policies with multiple programmes.

In Section 2, the concepts of potential and counterfactual outcomes and average treatment effects are introduced. Selection bias is discussed and the various nonparametric strategies to identify treatment effects in the context of multiple treatments are presented, together with a discussion on practical issues regarding their implementation and data requirements. These include controlling-for-confounding-variables, difference-in-difference, instrumental-variable and regression-discontinuity identification. In Section 3, the nonparametric estimation of average treatment effects and of mean counterfactual outcomes is discussed. Different generalized matching and weighting estimators are presented, including an examination of propensity score matching and asymptotic efficiency bounds. Section 4 concludes.

2. Identification of Average Treatment Effects

Policy and programme evaluation is concerned with measuring how far a policy or a programme has achieved its intended aims. A policy is hereafter defined as a bundle of R different programmes. This includes the case of evaluating a single programme ($R = 2$, participation versus non-participation) and evaluating multiple programmes ($R > 2$). One example of policies consisting of multiple programmes are active labour market policies, which often comprise various public employment programmes, on-the-job training, retraining, classroom training, job search assistance, wage subsidies etc. Another example are rehabilitation policies for the re-integration of people with long-term illnesses, which may consist of different forms of vocational workplace training, vocational schooling, medical rehabilitation and social and psychological programmes. In the following, often

the neutral term *treatment* will be used synonymously for *programme*, since the methods presented here are not restricted to the evaluation of social policies but apply similarly to, for example, the evaluation of the effectiveness of medical drugs or of different schooling choices, or of the effects of participation in the military. Since participation in a policy is often voluntary, or since full compliance in a 'mandatory' policy might not always be enforceable, the set of different treatments usually includes a 'no-programme' or 'non-participation' option. As it is assumed that all individuals are untreated before participation in the policy, i.e. that they had not participated previously in the programmes,[2] this 'non-participation' treatment is often special in the sense that it is the treatment most similar to the situation before participation in the policy. To illustrate this asymmetry, the treatment set will be indexed by $r \in \{0, \ldots, R-1\}$, i.e. consisting of a 'non-participation' treatment ($r=0$) and $R-1$ active treatments. In the case of the evaluation of a single programme the treatment set consists of $r=0$ (non-participation) and $r=1$ (participation).

The basic ideas and concepts of the current approaches to causal inference in programme evaluation stem from the statistical analysis of randomized experiments and potential outcomes.[3] The notion of *potential outcomes* was formalized in Neyman (1923), who considered potential yields of crop varieties on different plots of land with the plots *randomly* allocated to the crop varieties. Rubin (1974, 1977) provided a more thorough statistical framework for the concept of potential outcomes and extended it to the analysis of *observational* studies, where the units are not randomly assigned to the treatments.

Let i denote a unit (an individual, a household) which is assigned to one of R mutually exhaustive and exclusive treatments, i.e. each individual participates in exactly one of these treatments. Let

$$Y_i^0, Y_i^1, \ldots, Y_i^{R-1}$$

denote the potential outcomes for this individual. Y_i^0 is the outcome that would be realized (after treatment) if the individual i were assigned to treatment 0. Likewise, Y_i^1 is the outcome that would be realized if the individual i were assigned to treatment 1 and so forth.[4] *Ex ante*, i.e. before participation in the policy, each of these potential outcomes is latent and could be observed if the individual participated in the respective programme. *Ex post*, only the outcome corresponding to the programme in which the individual eventually participated is observed. The other potential outcomes are counterfactual and unobservable by definition.

2.1 *Stable-unit-treatment-value Assumption*

The definition of potential outcomes already made implicit use of the assumption of 'no interference between different units' (Cox, 1958, p. 19) or stable-unit-treatment-value assumption (SUTVA), (Rubin, 1980). It is assumed that the potential outcomes $Y_i^0, Y_i^1, \ldots, Y_i^{R-1}$ of individual i are not affected by the allocation of other individuals to the treatments. Formally, let $\mathbf{D}$ denote a

treatment-allocation *vector*, which indicates for all individuals the programme in which they participate. Let $\mathbf{Y}$ denote the vector of the observed outcomes of all individuals. Define $\mathbf{Y}(\mathbf{D})$ as the potential outcome vector that would be observed if all individuals were allocated to the policy according to the allocation $\mathbf{D}$. Further let $Y_i(\mathbf{D})$ denote the ith element of this potential outcome vector.

The stable-unit-treatment-value assumption states that for any two allocations $\mathbf{D}$ and $\mathbf{D}'$

$$Y_i(\mathbf{D}) = Y_i(\mathbf{D}') \quad \text{if} \quad \mathbf{D}_i = \mathbf{D}'_i,$$

where $\mathbf{D}_i$ and $\mathbf{D}'_i$ denote the ith element of the allocations $\mathbf{D}$ and $\mathbf{D}'$, respectively. In other words, it is assumed that the observed outcome Y_i depends only on the treatment to which individual i is assigned and not on the allocation of other individuals.

This assumption might be invalidated if individuals interact, either directly or through markets. For example, if active labour market programmes change the relative supply of skilled and unskilled labour, all individuals may be affected by the resulting changes in the wage structure. In addition, programmes which affect the labour cost structure, e.g. through wage subsidies, may lead to displacement effects, where unsubsidized workers are laid off and are replaced by subsidized programme participants. Individuals might further be affected by the taxes raised for financing the policy. The magnitude of such market and general equilibrium effects often depends on the scale of the policy, i.e. on the number of participants in the programmes. Departures from SUTVA are likely to be small if only a few individuals participate in the policy, and usually they become larger with increasing numbers of participants. This is the motivation of studies attempting to estimate the general equilibrium effects in the evaluation of active labour market policies by an augmented matching function approach, such as Blanchard and Diamond (1989, 1990) or Puhani (1999). In these studies, observed variations in the scale of the policy over time or geographic location are exploited to estimate the influence of the scale of the policy on the number of unemployed persons who become re-employed. Although these studies provide important insights, their interpretation is often difficult. Apart from using arbitrary parametric specifications, they often do not rest on an explicit causal framework. In many cases, the variations in the policy scale over time are not exogenous, but influenced by the outcomes of the policy in previous periods, which makes it difficult to define a causal effect. Furthermore, disentangling the general equilibrium effects of policies with multiple programmes could be a demanding task.

A different form of interference between individuals can arise due to supply constraints. For example, if the number of programme slots of a certain programme is limited, the availability of the programme for a particular individual depends on how many participants have already been allocated to this programme. Such interaction does not directly affect the potential outcomes and, thus, does not invalidate the microeconometric evaluation approaches discussed subsequently. However, it restricts the set of feasible allocations $\mathbf{D}$ and could

become relevant when trying to change the allocation of participants in order to improve the overall effectiveness of the policy. Supply constraints are often (at least partly) under the control of the programme administration and could be moderated if necessary.

Henceforth, the validity of SUTVA is assumed. Consequently, it is no longer necessary to take account of the full treatment-allocation vector **D**, since the outcome of individual i depends only on the treatment received by himself, which is denoted by a *scalar* variable D_i in the following. Such an approach is warranted if the policy under consideration is rather small in size, if market effects are unlikely, or if the counterfactual world against which the policy is evaluated is such that similar distortions through market and general equilibrium effects would persist, e.g. if the only feasible policy options are to marginally increase or decrease the scale of the policy.

2.2 *Average Treatment Effects and Selection Bias*

The difference between the potential outcome Y_i^r and the potential outcome Y_i^s can be interpreted as the gain or loss that individual i would realize if he participated in programme r relative to what he would realize if he participated in programme s. Thus the difference $Y_i^r - Y_i^s$ is the causal effect of participating in programme r *and not* participating in programme s. In the binary treatment case (i.e. the evaluation of a single programme), the difference $Y_i^1 - Y_i^0$ represents the difference between participating and not participating. Such individual treatment effects (Rubin, 1974) can never be ascertained since only one of the potential outcomes $Y_i^0, Y_i^1, \ldots, Y_i^{R-1}$ can be observed *ex post*: $Y_i^{D_i}$ where $D_i \in \{0, \ldots, R-1\}$ indicates the programme in which individual i actually participated. Therefore programme evaluation seeks to learn about the properties of the potential outcomes in the population. Since only one of the potential outcomes can be observed for each individual, the joint distribution of the potential outcomes $Y^0, \ldots, Y^{R-1}$ is not identified and, consequently, at most the properties of their marginal distributions can be uncovered. A parameter of interest is the *average treatment effect* (ATE)

$$E[Y^r - Y^s], \tag{1}$$

which is the difference between the outcome expected after participation in programme r and the outcome expected after participation in programme s for a person randomly drawn from the population. Analogously, the *average treatment effect on the treated* (ATET)

$$E[Y^r - Y^s | D = r]. \tag{2}$$

is the expected outcome difference for a person randomly drawn from the subpopulation of participants in programme r.

Most of the literature focused on the evaluation of a single programme, where there are only two groups to examine: those who participated in the programme and those who did not. In this situation, three treatment effects are of interest: the average treatment effect on the treated, the average treatment effect on the

non-treated, and the average treatment effect in the population, which is a weighted average of the first two effects. When the policy to be evaluated consists of a variety of different programmes, however, the analysis becomes more complex since a multitude of average treatment effects can be defined. While the comparison between programme r and the 'non-participation' programme: $E[Y^r - Y^0]$ gives the effect between a single programme relative to no programme, it is also of interest to estimate the effects *between* different treatments: $E[Y^r - Y^s]$, for $s \neq 0$. In addition, a comparison between the effect $E[Y^r - Y^s|D = r]$, i.e. for the participants in programme r, and the effect $E[Y^r - Y^s|D = s]$, i.e. for the participants in programme s, indicates whether those who participated in programme s would have gained more from being in programme r relative to those who actually did participate in programme r. In addition, it might be of interest to learn the effect between a programme r and a programme s for those who participated in a third programme: $E[Y^r - Y^s|D = t]$, or for those who participated in either of the two programmes: $E[Y^r - Y^s|D \in \{r, s\}]$, as defined in Lechner (2001a).

The most interesting effect depends on the specific policy context. For example, in the binary treatment case with voluntary participation it may be more informative to know how the programme affected those who participated in it, than how it might have affected those who could have participated but decided not to. In this case, the average treatment effect on the treated $E[Y^1 - Y^0|D = 1]$ would be more interesting than the average effect on the non-participants $E[Y^1 - Y^0|D = 0]$.[5] In the multiple treatment case, usually many effects are relevant for an assessment of the policy.

A further difference between the evaluation of a single programme and of multiple programmes is that many identification strategies that are useful for the evaluation of a single programme are less instructive for the evaluation of multiple treatments. For instance, the difference-in-difference approach (discussed in Section 2.6) is only informative for estimating the effect of a programme r versus the non-participation programme, but is not of help for comparisons between the programmes. Neither do instrumental variables (discussed in Section 2.7) identify the effects between different programmes.

To identify average treatment effects from a sample of past programme participants and non-participants, additional assumptions are required. Let $\{(X_i, D_i, Y_i)\}_{i=1}^n$ be a sample of previous participants, where $Y_i = Y_i^{D_i}$ is the observed outcome and X_i are other individual characteristics. Since Y^r is only observed for the participants in programme r, the data identifies $E[Y^r|D = r]$ and $E[Y^r|X,\ D = r]$ for all r but not $E[Y^r]$ or $E[Y^r|D = s]$. Generally the potential outcomes are different in the various subpopulations

$$E[Y^r|D = r] \neq E[Y^r|D = s] \neq E[Y^r].$$

Consequently, estimating the average treatment effect on the treated (2) by the difference in the subpopulation means $E[Y^r|D = r]$ and $E[Y^s|D = s]$ would give a biased estimate since

$$E[Y^r|D = r] - E[Y^s|D = s] = E[Y^r - Y^s|D = r] + \{E[Y^s|D = r] - E[Y^s|D = s]\}. \quad (3)$$

The second last term in (3) is the proper average treatment effect on the treated (2), whereas the last term in (3) is the *selection bias* (Heckman and Robb, 1985; Manski, 1993). Selection bias arises because the participants in programme r and the participants in programme s are deliberately selected groups that would have different outcomes, even if they were placed into the same programme. In making their programme participation decisions, individuals conjecture about their potential outcomes and base their choice on these guesses. In addition, unobserved character traits, such as health, motivation, ability or work commitment, lead to selection bias if they are correlated with the programme participation decision *and* the potential outcomes (e.g. earnings, employment status). Often programme participation is not completely voluntary but a joint decision of different parties, e.g. an unemployed person and a case worker. Again selection bias arises if the programme participation decision depends either consciously or unconsciously on factors related to the potential outcomes, for example, if case workers assign unemployed persons to particular programmes on the basis of their labour market history.

2.3 *Nonparametric Identification Strategies*

Hence data alone are not sufficient to identify average treatment effects. Conceptual causal models are required, which entail identifying assumptions about the process through which the individuals were assigned to the treatments, or about stability of the outcomes over time, see Pearl (2000). The corresponding minimal identifying assumptions cannot be tested with observational data and their plausibility must be assessed through prior knowledge of institutional details, the allocation process and behavioural theory. Below possible evaluation strategies to identify average treatment effects are presented. Which of these identification strategies, if any, is appropriate depends on the outcome variable of interest. As most policies pursue multiple and often conflicting goals, usually many different outcome variables are of interest, including economic, social, health and psychological indicators as well as programme cost variables. Since selection bias is a phenomenon caused by factors that affect *jointly* the participation decision *and* the potential outcome, selection bias might occur for some outcome variables but not for others. If the effects of a policy should be ascertained with respect to multiple outcome variables (Y^r being a vector), the appropriate identification strategy has to be chosen for each outcome variable on a case by case basis. It may be that a simple evaluation strategy can be used for some outcome variables, for which selection bias seems unlikely (e.g. monetary programme costs), whereas sophisticated evaluation strategies are required for other outcome variables, and finally, it may happen that for some outcome variables, no effect can be identified with the available data.

The *nonparametric identification strategies* discussed below all rely in one way or another on comparing the observed outcomes of one group of individuals with the observed outcomes of another group of individuals to identify average treatment effects. An exception to these comparison group approaches is the

before-after estimator which estimates $E[Y^r - Y^0|D=r]$ by comparing the observed outcomes of the participants before and after the treatment. It relies on the assumption of temporal stability (Holland, 1986), i.e. that the outcome observed before participation is the same as the outcome that would be observed in the 'non-participation' treatment at a later point in time. This assumption is usually not valid if the individual's environment changes over time. For example, Ashenfelter (1978) observed that the earnings of participants in active labour market programmes often had deteriorated recently before participation in the programme. If this decline in earnings represents a transitory labour market shock, it is likely that earnings would have recovered (at least partly) even without participation. The before-after estimator, however, would ascribe all increases in earnings to the programme participation. In this case, the effect of the treatment would be overstated. Particularly, if medium and long term effects of a programme shall be estimated, temporal stability is often not valid. Besides this, the before-after comparison strategy is not very suited for the evaluation of a policy consisting of multiple programmes, since it could only be used to estimate the treatment effect on the treated relative to non-participation, but not for any comparison between the active treatments.

2.4 *Randomized Experiment*

The ideal solution to avoid selection bias due to systematic selection of the participants is to assign individuals randomly to the programmes, as advocated in Fisher (1935). Randomization ensures that the probability to be assigned to a certain treatment is not influenced by the potential outcomes

$$P\left(D = d|Y^0, \ldots, Y^{R-1}\right) = P(D = d)$$

or in the notation of Dawid (1979) that the potential outcomes $Y^0, \ldots, Y^{R-1}$ are statistically independent ($\perp\!\!\!\perp$) of the treatment indicator D

$$Y^0, \ldots, Y^{R-1} \perp\!\!\!\perp D.$$

Random programme assignment ensures that any differences between the treatment groups are by pure chance and not systematic. Consequently, the observed outcomes Y^r among the participants in programme r have the same expected value as the potential outcomes Y^r among the participants in programme s

$$E[Y^r|D = r] = E[Y^r|D = s] = E[Y^r],$$

and selection bias is thus avoided.

Yet implementing a randomized experiment for evaluating social programmes is often not trivial. Participation in a particular policy is often voluntary such that randomization can only be implemented with respect to the individuals who applied for the programme.[6] However, these might be different from the population of interest. Particularly, if randomization covers only parts of the population, the experimental results may not be generalizable to the broader population. Even

if a policy is mandatory and all individuals can be randomly assigned to the treatments, full compliance is often difficult to achieve if participants must exercise some effort during the participation and may refuse their cooperation. Heckman and Smith (1995) discuss different sources that may invalidate the experimental evaluation results. *Randomization bias* occurs if the prospect of randomized allocation alters the pool of potential participants because individuals may be reluctant to apply at all or reduce any preparatory activities such as complementary training due to the fear of being randomized-out (threat of service denial). *Substitution bias* occurs if members of the control group (the randomized-out non-participants) obtain some treatment or participate in similar programmes, e.g. training obtained from private providers. In this case, the experimental evaluation measures only the incremental value of the policy relative to the programmes available otherwise. *Drop-out bias* occurs if individuals assigned to a particular programme do not (or only partly) participate in it. Heckman and Smith (1995) also mention that randomized experiments are expensive, often face political obstacles and may distort the operation of an on-going policy.[7]

2.5 *Control for Confounding Variables*

Even if a randomized experiment would have been feasible, it often simply has not been implemented at the onset of the policy. In this case, only observational data are available and the selection problem must be solved by other means. One approach is to mimic the idea of a randomized experiment and to form comparison groups which are as similar as possible. Accordingly this identification strategy is also called quasi-experimental. The underlying motivation can be illustrated as in Rubin (1974): If two individuals i and j are found that are identical (or very similar) in all their characteristics, then also Y_i^r and Y_j^r should be similar. If one of these individuals takes part in programme r and the other in programme s, and if many such pairs are found, then the difference in observed outcomes could be used as an estimate of the average treatment effect between programme r and programme s.

In fact, for this estimate to be consistent, it is not necessary that individuals i and j are identical in all their characteristics. It suffices if they are identical with respect to all *confounding* variables. The confounding variables are all factors that influence treatment selection *and* the potential outcomes. Hence variables that affect only the treatment choice or only the potential outcomes need *not* to be controlled for. This is analogous to the assumption in the standard econometric simultaneous equations framework that the error term in the outcome equation and the error term in the selection equation are independent, after controlling for the confounding variables X. (The simultaneous equations framework, however, introduces additional functional form assumptions, which are not invoked here.)

This identification strategy rules out, for example, that individuals know their potential outcomes and choose the treatment with the highest outcome. In other words, the probability of choosing a particular programme must not be affected by the potential outcomes. However, treatment selection is allowed to depend on anticipated potential outcomes as long as these potential outcomes are

anticipated on the basis of the exogenous characteristics X, but not on the basis of unobserved characteristics.

An alternative interpretation of this identification strategy is that, given the characteristics X, the programme chosen by a particular individual reveals no information about his potential outcomes:

$$Y^r \perp\!\!\!\perp D|X \quad \forall r. \tag{4}$$

For instance, if more motivated individuals were more inclined to participate in a programme, the mere observation that a particular individual participated would suggest that his motivation is above average. If higher motivation has an effect on the potential outcomes, observing D would also reveal information about Y^r – unless motivation is included in X. If, on the other hand, motivation has no effect on the potential outcomes, there is no need to control for motivation in (4). The assumption (4) is known as *selection on observables* (Barnow *et al*., 1981), *ignorable treatment assignment* (Rosenbaum and Rubin, 1983) or as *conditional independence assumption* (Lechner, 1999).

The confounding variables often include time-varying variables as well. For example, Ashenfelter (1978) noted that the decision to participate in active labour market programmes is highly dependent on the individual's previous earnings and employment histories. Recent negative employment shocks often induce individuals to participate in training programmes. Hence the employment situation in the months before the programme starts is an important determinant of the programme participation decision and is also likely to be correlated with the potential employment outcomes. However, since usually no explicit start date can be observed for the participants in the 'non-participation' treatment, the employment situation in the months before the programme started is undefined for them. To solve this problem, Lechner (1999) suggested drawing hypothetical start dates for the 'non-participants' from the distribution of start dates among the participants.[8] Lechner (2002b) analyzed the assignment of hypothetical start dates further. Instead of drawing dates from the unconditional distribution of start dates, he also considered drawing from the distribution conditional on the confounding variables. This conditional distribution can be simulated by regressing the (logarithm of the) start dates on the covariates and fitting the mean of the conditional distribution at the covariate values of the respective non-participant. In his application both methods led to similar results.

On the other hand, X must *not* include any variables that are itself affected by the policy. These are (endogenous) variables that are caused by the policy and conditioning on such variables would block the part of the causal effect that acts through these variables (Pearl, 2000). For example, if the prospective participants in a particular programme are notified a few weeks in advance about their participation status, the announcement in itself might already have an effect on the individuals' behaviour even before the programme starts. Including any variables in X, that are observed after notification and that are affected by it, would thus omit (part of) the effect of the announcement on the outcome variable, which, arguably, might be considered as part of the total programme effect.

The variables that must and must not be included in X cannot be inferred from the data, nor can their completeness be tested. Knowledge of the institutional details and a conceptual causal model are required to assess which variables are relevant. Hence *a priori*, the selection on observables assumption (4) can neither be regarded as a strong or a weak condition; this depends entirely on the policy specific details, the outcome variable of interest and the available data.[9] However the validity of this assumption should be carefully assessed, since leaving out relevant covariates can change the estimation results considerably and lead to wrong conclusions as, for example, demonstrated in Lechner (2002b).

Generally speaking, identification by the conditional independence assumption (4) is easier to achieve the more bureaucratic, rule-based and deterministic the programme selection process is[10] and the more parties are involved that (truthfully) report their judgements about the individual's characteristics and behaviour (e.g. case worker's and physician's judgements in Frölich *et al.* (2000)). For example, in his analysis of the effects of voluntary participation in the military on civilian earnings, Angrist (1998) takes advantage of the fact that the military is known to screen applicants to the armed forces on the basis of particular characteristics, primarily on the basis of age, schooling and test scores. Hence these characteristics are the principal factors guiding the acceptance decision, and it appears reasonable to assume that among applicants with the same observed characteristics, those who finally enter the military and those who do not are not systematically different. A similar reasoning applies to the effects of schooling, if it is known that applicants to a school or university are screened on the basis of certain characteristics, but that conditional on these characteristics selection is on a first-come/first-serve basis.

On the other hand, if individuals decide largely autonomously, and if no details about their personal traits are available (e.g. in form of truthful self-assessments), validity of the conditional independence assumption is much harder to establish. In this case, longitudinal data containing, for example, past employment and earnings histories can help to proxy typical, though unobserved traits of the individual (e.g. ability, discipline, work commitment, health status). Such a very informative longitudinal dataset is used, for example, in Gerfin and Lechner (2002) for the evaluation of active labour market policies in Switzerland.

If the conditional independence assumption (4) is valid, the potential outcomes conditional on X are identified because

$$E[Y^r|X, D = r] = E[Y^r|X, D = s] = E[Y^r|X].$$

The average treatment effect (1) and the average treatment effect on the treated (2) can be obtained by weighting these outcomes by the distribution of X in the respective population. By the law of iterated expectations, the average treatment effect is identified as

$$\begin{aligned} E[Y^r - Y^s] &= E[Y^r] - E[Y^s] \\ &= E[E[Y^r|X]] - E[E[Y^s|X]] \\ &= \int (E[Y^r|X = x, D = r] - E[Y^s|X = x, D = s]) \cdot f_X(x)dx, \quad (5) \end{aligned}$$

where $f_X(x)$ is the density of X in the population. Since $E[Y^r|X, D=r]$ and $E[Y^s|X, D=s]$ can be estimated from observed data, the average treatment effect can be obtained by estimating the expected outcome conditional on X in both treatment groups and weighting them accordingly by the distribution of X in the full population.

Analogously, the average treatment effect on the treated is identified as

$$\begin{aligned} E[Y^r - Y^s|D=r] &= E[Y^r|D=r] - E[E[Y^s|X, D=r]|D=r] \\ &= E[Y^r|D=r] - \int E[Y^s|X=x, D=s] \cdot f_{X|D=r}(x)dx, \end{aligned} \tag{6}$$

where $f_{X|D=r}(x)$ denotes the density of X among the participants in programme r. The former term is identified by the sample mean outcome of the participants in programme r, and the latter term can be estimated by adjusting the average outcomes in treatment group s for the distribution of X among the participants in r.

The requirement of a *common support* has been neglected in the discussion so far. Although the conditional independence assumption (4) identifies the conditional potential outcomes $E[Y^r|X=x]$ through observations on participants in programme r, this identification holds only for all x for which there is a positive probability that participants in programme r are observed with characteristics x. Let

$$S^r = \{x : f_{X|D=r}(x) > 0\}, \tag{7}$$

denote the support of X among the participants in programme r, which can also be expressed as

$$S^r = \{x : p^r(x) > 0\},$$

where $p^r(x) = P(D=r|X=x)$ is the probability that an individual with characteristics x participates in programme r.[11] For any $x \notin S^r$ the expected outcome $E[Y^r|X=x, D=r]$ is not identified, since it is impossible to observe any participant in programme r with characteristics x. Let S denote the support of X in the population, i.e. $S=\{x : f_X(x)>0\}$, which is the union of all treatment group supports: $S=\cup S^r$. The average treatment effect on the treated $E[Y^r - Y^s|D=r]$ is only identified, if $S^r \subseteq S^s$, i.e. if any x with positive mass among the participants in treatment r belongs also to the support of the treatment subpopulation s. Identification of the average potential outcome $E[Y^r]$ requires even that $S^r=S$, i.e. that each individual has also a positive probability of being observed in programme r. Analogously, the identification of the average treatment effect $E[Y^r - Y^s]$ requires $S^r=S^s=S$. In the case of randomized experiments these conditions are automatically satisfied (for the population on which randomization took place) since each individual has a positive probability of being randomized into any of the available programmes. With observational studies, however, this is often not the case. For example, in active labour market programmes being unemployed is usually a central condition for eligibility. Thus employed persons

cannot be participants as they are not eligible and, hence, no counterfactual outcome is defined for them. In these cases, it might be adequate to concentrate on the part of the population for which the effect can be identified and to redefine the average treatment effect (1) as

$$\underset{S^r \cap S^s}{E}[Y^r - Y^s] = \underset{X|X \in (S^r \cap S^s)}{E} E[Y^r - Y^s|X]$$

and the average treatment effect on the treated (2) as

$$\underset{S^r \cap S^s}{E}[Y^r - Y^s|D = r],$$

where $E_{S^r \cap S^s}$ refers to the expected outcome with respect to the common support, i.e. with respect to the part of the population which has characteristics X belonging to the supports S^r and S^s. If most of the population mass belongs to the common support, these re-defined treatment effects are likely to be close to the treatment effects for the full population. Furthermore, if the potential outcomes are bounded random variables, the intervals in which the average outcomes $E[Y^r]$ and $E[Y^s]$ lie can be bounded, which directly implies bounds on the average treatment effect $E[Y^r - Y^s]$. For a further discussion about bounding the effects when $S^r \neq S^s$ see Lechner (2001b). For the remainder of this paper, conditioning on the common support is usually kept implicit to ease notation.[12]

A way to 'verify' the validity of the conditional independence assumption (4), known as the 'pre-programme test', is based on observed outcomes before treatment participation (Heckman and Robb, 1985). To distinguish pre- and post-treatment outcomes, the notation has to be extended to take explicit account of time. Redefine $Y_t^0, \ldots, Y_t^{R-1}$ as the potential outcomes at a time t after the assignment to treatment. Let $Y_\tau^0, \ldots, Y_\tau^{R-1}$ be the potential outcomes at a time τ *before* the assignment to treatment and $Y_\tau = Y_\tau^D$ the observed outcome. Suppose that the outcomes before the assignment to the treatment are not affected by the subsequent treatment, i.e. that the outcome before treatment is the same regardless of the programme in which the individual eventually participates:

$$Y_\tau^0 = Y_\tau^1 = \ldots = Y_\tau^{R-1}. \tag{8}$$

Hence the observed outcome before treatment $Y_\tau = Y_\tau^D$ is no longer contingent on the treatment.

Validity of the conditional independence assumption (4) in period t means that all confounding variables are included in X such that, conditional on X, any differences between the treatment groups are unsystematic or at least not related to the potential outcomes $Y_t^0, \ldots, Y_t^{R-1}$. If the 'non-participation' outcomes Y_t^0 at time t and at time τ are related, it is also reasonable that the set of confounding factors contains the same variables at time t and τ. This would imply that, conditional on X, also the outcomes observed *before* treatment are not systematically different between the treatment groups

$$E[Y_\tau|X, D = r] = E[Y_\tau|X, D = s] = E[Y_\tau|X]. \tag{9}$$

On the other hand, if the conditional independence assumption is invalid, there are confounding factors, which are not included in X, that influence programme selection as well as the potential outcomes and cause selection bias: $E[Y_t^r|X, D = r] \neq E[Y_t^r|X, D = s]$. Again, if the 'non-participation' outcome Y_t^0 is related over time, it is likely that these factors would also generate systematic differences between the treatment groups in earlier time periods:

$$E[Y_\tau|X, D = r] \neq E[Y_\tau|X, D = s] \neq E[Y_\tau|X].$$

Large differences in the pre-programme outcomes between the treatment groups would thus cast doubts on the validity of the conditional independence assumption (4) and a formal test statistic can be derived as in Heckman and Robb (1985). Yet, it is not a proper test of the conditional independence assumption since its justification requires additional untestable assumptions about the relationship between pre- and post-treatment outcomes.

Furthermore, the application of the pre-programme test requires to find a time period τ where Y_τ is neither a confounding variable itself nor a variable already causally influenced by the programme. The former condition is routinely violated if previous Y_τ influence the participation decision D. For example, the past employment situation is often a strong determinant of participation in active labour market programmes. Hence Y_τ itself is a confounding variable and as such must be included in X as a conditioning variable.

The latter condition is violated if anticipation of programme participation changes the individual's behaviour even before the programme starts. For example, if an unemployed person gets informed that he is assigned to a particular labour market programme, he might immediately adjust his job-search intensity or any complementary training activities. This implies that the potential outcomes differ even before the beginning of the programme:

$$Y_\tau^0 \neq Y_\tau^1 \neq \ldots \neq Y_\tau^{R-1}. \tag{10}$$

Accordingly the observed outcome $Y_\tau = Y_\tau^D$ depends on the treatment eventually received. In this case, the equality (9) no longer holds.

In these cases the pre-programme test cannot be applied and this discussion highlights the importance of taking account of the time structure of the outcome variable. In different periods of time τ *before* the start of the programme, the variable Y_τ can be a confounder (in which case it must be included in X), an outcome variable causally affected through programme anticipation (in which case it must *not* be included in X), or neither confounding nor causally affected (in which case the pre-programme test can be applied).

2.6 *Difference in Difference – Predictable Bias Assumption*

In many evaluation settings it may not be feasible to observe all confounding variables. In these cases the evaluation strategy has to cope with selection on unobserved variables. Nevertheless, average treatment effects may still be identified either through an instrumental variable (see below) or if the average selection

bias can be estimated from pre-treatment outcomes. This latter approach is based on a similar motivation as the pre-programme test: If systematic differences in the pre-programme outcomes between different treatment groups occur, these differences may not only indicate that not all confounding variables have been included, but may further be useful to predict the magnitude of selection bias in the post-programme outcomes.

If X does not contain all confounding variables, adjusting for the differences in the X distributions, analogously to (6), will not yield a consistent estimate of the average treatment effect on the treated because

$$E\left[Y_t^r|D=r\right] - \int E\left[Y_t^s|X=x, D=s\right] \cdot f_{X|D=r}(x)dx$$
$$\neq E\left[Y_t^r|D=r\right] - \int E\left[Y_t^s|X=x, D=r\right] \cdot f_{X|D=r}(x)dx = E\left[Y_t^r - Y_t^s|D=r\right]$$

since $E[Y_t^s|X, D=r] \neq E[Y_t^s|X, D=s]$. The difference

$$\int (E[Y_t^s|X=x, D=r] - E[Y_t^s|X=x, D=s]) \cdot f_{X|D=r}(x)dx$$

is the systematic bias in the potential outcome Y_t^s in period t that still remains even after adjusting for the different distributions of X.

Pre-programme outcomes might help to estimate this systematic bias with respect to the 'non-participation' outcome Y_t^0. Therefore the following discussion centers on the identification of average treatment effects on the treated relative to non-participation: $E[Y_t^r - Y_t^0|D=r]$.[13] Define the *average selection bias*

$$B_t = \int (E[Y_t^0|X=x, D=r] - E[Y_t^0|X=x, D=0]) \cdot f_{X|D=r}(x)dx$$

as the systematic outcome difference between the group of non-participants ($D=0$) and the group of participants ($D=r$) if both groups would participate in treatment 0. If, for example in the evaluation of active labour market programmes, the individuals who decided to participate were on average more able, it is likely that their labour market outcomes would also have been better even without participation in the programme. In this case, the average selection bias B_t would be positive. If the potential outcome in the case of non-participation Y_t^0 is related over time, it is likely that these differences between the treatment groups would also persist in other time periods including periods before the start of the programme. In other words, the more able persons would also had enjoyed better labour market outcomes in periods before treatment.

If the pre-programme outcome in period τ is not causally affected by the programme, so that (8) holds, the 'non-participation' outcomes $Y_\tau^0 = Y_\tau$ are observed for the different treatment groups and the corresponding average selection bias in period τ

$$B_\tau = \int (E[Y_\tau|X=x, D=r] - E[Y_\tau|X=x, D=0]) \cdot f_{X|D=r}(x)dx$$

is identified from the observed pre-programme data.

Assuming that the average selection bias is stable over time

$$B_t = B_\tau \tag{11}$$

the average treatment effect on the treated is identified as

$$\begin{aligned} E[Y_t^r - Y_t^0|D=r] &= E[Y_t^r|D=r] - \left(\int E[Y_t^0|X=x, D=0] \cdot f_{X|D=r}(x)dx + B_t \right) \\ &= (E[Y_t^r|D=r] - E[Y_\tau|D=r]) \\ &\quad - \int (E[Y_t^0|X=x, D=0] - E[Y_\tau|X=x, D=0]) f_{X|D=r}(x)dx. \end{aligned} \tag{12}$$

This resembles a difference-in-difference type estimator adjusted for the distribution of the X covariates, which is further discussed in Section 3.[14]

The bias-stability assumption (11) is not strictly necessary. Instead, it suffices if B_t can be consistently estimated from the average selection biases in pre-programme periods (*predictable-bias assumption*). If (causally unaffected) pre-programme outcomes are observed for many periods, the average selection bias can be estimated in each period and any regular trends observed in $\hat{B}_\tau, \hat{B}_{\tau-1}, \hat{B}_{\tau-2}, \ldots$ may lead to better predictions of the bias B_t than simply estimating B_t by the selection bias in period τ, as the bias-stability assumption (11) would suggest.

Loosely speaking, the predictable-bias-assumption (with the bias-stability-assumption as a special case) is weaker than the conditional independence assumption (4) since it allows that $B_t \neq 0$, whereas the conditional independence assumption requires $B_t = 0$. However, both assumptions are not nested because B_t may be zero while $\hat{B}_\tau, \hat{B}_{\tau-1}, \hat{B}_{\tau-2}, \ldots$ may be unable to predict $B_t = 0$.[15] A further difference occurs if the pre-programme outcomes Y_τ are themselves confounders, i.e. influencing the treatment selection decision and the post-programme outcomes. If, in addition, all other confounding variables are observed, the independence assumption (4) would be valid conditional on the pre-programme outcomes *and* the other confounders. This would imply zero selection bias ($B_t = 0$) and the applicability of the control-for-confounding-variables approach. The difference-in-difference approach, on the other hand, would introduce selection bias ($B_t \neq 0$) by not conditioning on the pre-programme outcome Y_τ (i.e. not including Y_τ in X).

A weakness of the difference-in-difference approach is that it does not entail any theoretical guidelines for deciding which variables (if any at all) should be included in the conditioning set X.[16] Heckman *et al.* (1997, 1998) and Smith and Todd (2000) consider a stronger version of the bias-stability-assumption (11), which requires that the bias is stable not only on average but for any possible value of X

$$E[Y_t^0 - Y_\tau|X, D=r] = E[Y_t^0 - Y_\tau|X, D=0]. \tag{13}$$

This stronger assumption demands that all variables that affect the increase (growth) in the non-participation outcome over time ($Y_t^0 - Y_\tau$) and the selection

to treatment 0 or r are included in X. Although this stronger assumption does not help to identify the average treatment effect on the treated, it may be useful in the search to identify the relevant conditioning variables X: because if (13) is true then also (11) holds.

2.7 *Instrumental Variables Identification*

An alternative strategy to handle selection on the basis of unobserved characteristics exploits the identifying power of an instrumental variable, which is a variable that influences the probability to participate in a particular treatment but has no effect on the potential outcomes. It affects the observed outcome only indirectly through the participation decision. Causal effects can be identified through a variation in this instrumental variable since the effect of this variation is entirely channeled via the programme selection. Here only the basic ideas are outlined and illustrated for the binary treatment case ($R=2$) with a binary instrumental variable $Z \in \{0, 1\}$. Causal inference through instrumental variables has been analyzed in greater detail by Imbens and Angrist (1994), Angrist *et al*. (1996), Heckman and Vytlacil (1999) and Imbens (2001), among others.

A fundamental result of instrumental variables identification is that average treatment effects can only be identified with respect to the subpopulation that could be induced to change programme status by a variation in the instrumental variable (Imbens and Angrist, 1994). For subpopulations that would participate in the same treatment regardless of a hypothetical exogenous change in the value of Z, their counterfactual outcomes are not identified. This is similar to the common support restriction discussed above. Hence an average treatment effect for the full population could only be identified if all individuals change programme status with a variation in Z. Otherwise, only a *local average treatment effect* (LATE) for the subpopulation responsive to Z is identified (Imbens and Angrist, 1994).

Define $D_{i,Z_i=0}$ as the programme participation status $D \in \{0, 1\}$ that would be observed for individual i if Z_i were set exogenously to the value 0. Define $D_{i,Z_i=1}$ analogously as the participation status that would be observed if Z_i were set to 1. Hence $D_{i,Z_i=0}$ and $D_{i,Z_i=1}$ are *potential* participation indicators, and let D_{i,Z_i} denote the *observed* value of D_i for individual i, i.e. the participation decision corresponding to the realized value Z_i. According to the potential participation behaviour, the population of all individuals can be partitioned into 4 subpopulations: Individuals for whom $D_{i,Z_i=0}=D_{i,Z_i=1}=1$ always participate in the programme regardless of the value of the instrumental variable. On the other hand, individuals with $D_{i,Z_i=0}=D_{i,Z_i=1}=0$ never participate. Individuals for whom $D_{i,Z_i=0}=0$ and $D_{i,Z_i=1}=1$ participate in the programme only if the instrument takes the value 1 and do not participate otherwise. These individuals 'comply' with their instrument assignment and are denoted compliers. Finally, individuals with $D_{i,Z_i=0}=1$ and $D_{i,Z_i=1}=0$ participate only if the instrument takes the value 0 and are called defiers. Thus each individual can be classified

either as an always-participant, a never-participant, a complier or a defier. Let τ_i denote the participation-type of individual i:

Definition of types		
$\tau_i = a$	if $D_{i,Z_i=0} = 1$ and $D_{i,Z_i=1} = 1$	Always − participant
$\tau_i = n$	if $D_{i,Z_i=0} = 0$ and $D_{i,Z_i=1} = 0$	Never − participant
$\tau_i = c$	if $D_{i,Z_i=0} = 0$ and $D_{i,Z_i=1} = 1$	Complier
$\tau_i = d$	if $D_{i,Z_i=0} = 1$ and $D_{i,Z_i=1} = 0$	Defier.

(14)

Since the individuals of type always-participant and of type never-participant cannot be induced do change treatment state through a variation in the instrumental variable, the impact of D on Y can at most be ascertained for the subpopulation of compliers and defiers. To analyze identification of local average treatment effects by instrumental variables, the potential outcomes framework needs to be extended: Define $Y_{i,Z_i}^{D_i=0}$ and $Y_{i,Z_i}^{D_i=1}$ as the *potential* outcomes for individual i, where $Y_{i,Z_i}^{D_i=0}$ is the outcome that would be observed for individual i if D_i were set to 0, and $Y_{i,Z_i}^{D_i=1}$ is the outcome that would be observed if individual i were assigned to the programme. Define $Y_{i,Z_i=0}^{D_i}$ and $Y_{i,Z_i=1}^{D_i}$ as the outcomes that would be observed if the instrument Z_i were set to 0 or 1, respectively. Similarly, $Y_{i,Z_i=0}^{D_i=0}$, $Y_{i,Z_i=0}^{D_i=1}$, $Y_{i,Z_i=1}^{D_i=0}$, $Y_{i,Z_i=1}^{D_i=1}$ are the outcomes that could be observed if the instrument Z_i *and* the participation indicator D_i were set exogenously. The conceptual difference between the potential outcomes $Y_{i,Z_i=0}^{D_i=0}$, $Y_{i,Z_i=0}^{D_i=1}$, $Y_{i,Z_i=1}^{D_i=0}$, $Y_{i,Z_i=1}^{D_i=1}$ and the potential outcomes $Y_{i,Z_i=0}^{D_i}$, $Y_{i,Z_i=1}^{D_i}$ is that in the former case Z and D are fixed by external intervention, whereas in the latter case only Z is set exogenously and D_i is determined by the participation behaviour of individual i. In other words, the former outcomes isolate the direct effect of the instrument Z on Y, while the latter embed the direct effect and the indirect effect of Z on Y via the treatment participation D. Finally, $Y_i = Y_{i,Z_i}^{D_i}$ is the observed outcome where Z_i and D_i are the realized values for individual i.

From Table (14) already it can be seen, why instrumental variables are not particularly suited for the evaluation of multiple programmes. For instance, if the policy consisted of three programmes (i.e. $D \in \{0,1,2\}$), 9 different types of individuals could be defined: individuals who always participate in programme 0, 1 or 2, respectively; individuals who switch from 0 to 1, from 0 to 2, from 1 to 0, etc. A single instrument would not suffice to estimate the various conceivable effects between the programmes. One might think that simply eliminating all individuals with $D_i = 2$ would permit identifying the local average treatment effect between programme 0 and 1. However, deleting all individuals for whom $D_i = 2$ is *observed* would not achieve this, because the remaining sample would still contain all individuals that would participate in programme 2 if their instrument were switched. For identifying the local treatment effect between programme 0 and 1, it would be necessary to eliminate also those individuals for whom $D_i = 2$ *would*

be observed if the instrument were changed. This, however, is not feasible. The following discussion, therefore, continues with the evaluation of a single programme ($R = 2$).

With these definitions the expected value of the outcome variable Y can be written as the expected value of Y in the four subpopulations defined by (14) weighted by the relative size of these subpopulations:

$$\begin{aligned} E[Y] = {} & E[Y_i|\tau_i = a] \cdot P(\tau_i = a) \\ & + E[Y_i|\tau_i = n] \cdot P(\tau_i = n) \\ & + E[Y_i|\tau_i = c] \cdot P(\tau_i = c) \\ & + E[Y_i|\tau_i = d] \cdot P(\tau_i = d). \end{aligned} \tag{15}$$

Analogously, the conditional expectation of Y given the instrument can be written as:

$$\begin{aligned} E[Y|Z = 0] = {} & E\left[Y_{i,Z_i}^{D_i}|Z_i = 0, \tau_i = a\right] \cdot P(\tau_i = a|Z_i = 0) \\ & + E\left[Y_{i,Z_i}^{D_i}|Z_i = 0, \tau_i = n\right] \cdot P(\tau_i = n|Z_i = 0) \\ & + E\left[Y_{i,Z_i}^{D_i}|Z_i = 0, \tau_i = c\right] \cdot P(\tau_i = c|Z_i = 0) \\ & + E\left[Y_{i,Z_i}^{D_i}|Z_i = 0, \tau_i = d\right] \cdot P(\tau_i = d|Z_i = 0) \\ = {} & E\left[Y_{i,Z_i=0}^{D_i=1}|Z_i = 0, \tau_i = a\right] \cdot P(\tau_i = a|Z_i = 0) \\ & + E\left[Y_{i,Z_i=0}^{D_i=0}|Z_i = 0, \tau_i = n\right] \cdot P(\tau_i = n|Z_i = 0) \\ & + E\left[Y_{i,Z_i=0}^{D_i=0}|Z_i = 0, \tau_i = c\right] \cdot P(\tau_i = c|Z_i = 0) \\ & + E\left[Y_{i,Z_i=0}^{D_i=1}|Z_i = 0, \tau_i = d\right] \cdot P(\tau_i = d|Z_i = 0) \end{aligned}$$

and

$$\begin{aligned} E[Y|Z = 1] = {} & E\left[Y_{i,Z_i=1}^{D_i=1}|Z_i = 1, \tau_i = a\right] \cdot P(\tau_i = a|Z_i = 1) \\ & + E\left[Y_{i,Z_i=1}^{D_i=0}|Z_i = 1, \tau_i = n\right] \cdot P(\tau_i = n|Z_i = 1) \\ & + E\left[Y_{i,Z_i=1}^{D_i=1}|Z_i = 1, \tau_i = c\right] \cdot P(\tau_i = c|Z_i = 1) \\ & + E\left[Y_{i,Z_i=1}^{D_i=0}|Z_i = 1, \tau_i = d\right] \cdot P(\tau_i = d|Z_i = 1). \end{aligned}$$

Suppose

[*Assumption 1: Unconfounded participation type*] that the relative size of the subpopulations always-participants, never-participants, compliers and defiers is independent of the instrument:

$$P(\tau_i = t|Z_i = 0) = P(\tau_i = t|Z_i = 1) \qquad \text{for } t \in \{a, n, c, d\}. \tag{16}$$

Suppose further
[*Assumption 2: Mean exclusion restriction*] that the potential outcomes are mean independent of the instrumental variable Z in each subpopulation

$$E\left[Y_{i,Z_i}^{D_i=0}|Z_i=0,\tau_i=t\right]=E\left[Y_{i,Z_i}^{D_i=0}|Z_i=1,\tau_i=t\right] \qquad \text{for } t\in\{n,c,d\}$$
$$E\left[Y_{i,Z_i}^{D_i=1}|Z_i=0,\tau_i=t\right]=E\left[Y_{i,Z_i}^{D_i=1}|Z_i=1,\tau_i=t\right] \qquad \text{for } t\in\{a,c,d\}. \tag{17}$$

With this exclusion restriction the expected value of Y given Z is independent of Z in the always- and in the never-participant subpopulation. Hence when taking the difference $E[Y|Z=1]-E[Y|Z=0]$ the respective terms for the always- and for the never-participants cancel:

$$\begin{aligned}E[Y|Z=1]-E[Y|Z=0]&\\ =\left(E\left[Y_{i,Z_i=1}^{D_i=1}|Z_i=1,\tau_i=c\right]-E\left[Y_{i,Z_i=0}^{D_i=0}|Z_i=0,\tau_i=c\right]\right)\cdot P(\tau_i=c)&\\ +\left(E\left[Y_{i,Z_i=1}^{D_i=0}|Z_i=1,\tau_i=d\right]-E\left[Y_{i,Z_i=0}^{D_i=1}|Z_i=0,\tau_i=d\right]\right)\cdot P(\tau_i=d)&\end{aligned}$$

and with the mean exclusion restriction for the compliers and the defiers:

$$=\left(E\left[Y_{i,Z_i}^{D_i=1}-Y_{i,Z_i}^{D_i=0}|\tau_i=c\right]\right)\cdot P(\tau_i=c)-\left(E\left[Y_{i,Z_i}^{D_i=1}-Y_{i,Z_i}^{D_i=0}|\tau_i=d\right]\right)\cdot P(\tau_i=d). \tag{18}$$

The difference $E[Y|Z=1]-E[Y|Z=0]$ thus represents the difference between the average treatment effect on the compliers (who switch into treatment as a reaction on a change in the instrument from 0 to 1) and the average treatment effect on the defiers (who switch out of treatment). An estimate of (18) is not very informative since, for example, an estimate of zero could be the result of a treatment without effect as well as the result of a treatment with a large impact but offsetting flows of compliers and defiers. Hence the exclusion restriction (17) is not sufficient to isolate a meaningful treatment effect. However, as (18) indicates, a treatment effect could be identified if either no compliers $P(\tau_i=c)=0$ or no defiers $P(\tau_i=d)=0$ existed. If an instrumental variable is found that affects *all* individuals in the 'same direction', e.g. that either induces individuals to switch into participation or leaves their participation status unchanged, but does not induce any individual to switch out of treatment, the average treatment effect on the responsive subpopulation is identified.
[*Assumption 3: Monotonicity*] Suppose that the subpopulation of defiers has relative size zero

$$P(\tau=d)=0, \tag{19}$$

or in other words, that $D_{i,Z_i=1}\geq D_{i,Z_i=0}$ for all i. Suppose further
[*Assumption 4: Existence of compliers*] that the instrumental variable does have an impact on treatment choice, i.e. that the subpopulation of compliers exists

$$P(\tau=c)>0. \tag{20}$$

With these additional assumptions, the expression (18) can be written as

$$E\left[Y_{i,Z_i}^{D_i=1} - Y_{i,Z_i}^{D_i=0} | \tau_i = c\right] = \frac{E[Y|Z=1] - E[Y|Z=0]}{P(\tau = c)},$$

where $E\left[Y_{i,Z_i}^{D_i=1} - Y_{i,Z_i}^{D_i=0} | \tau_i = c\right]$ is the average treatment effect in the subpopulation of compliers.

Noting further that $P(D=1|Z=0) = P(\tau = a) + P(\tau = d)$ and $P(D=1|Z=1) = P(\tau = a) + P(\tau = c)$, it follows with (19) that the relative size of the subpopulation of compliers is identified as

$$P(\tau = c) = P(D = 1|Z = 1) - P(D = 1|Z = 0).$$

Hence the average treatment effect on the compliers is identified as

$$E\left[Y_{i,Z_i}^{D_i=1} - Y_{i,Z_i}^{D_i=0} | \tau_i = c\right] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}. \quad (21)$$

This is the average treatment effect on those individuals who are induced to switch into the programme due to the instrumental variable. Since the subpopulation of compliers is not identified, it often may be difficult to interpret this treatment effect. As the complier subpopulation is defined through the instrumental variable, any local average treatment effect is directly tied to its instrumental variable and cannot be interpreted on its own. For example, if the instrumental variable Z represents the size of a programme (e.g. the number of available slots), the local average treatment effect would represent the impact of the programme if it were extended from size z_0 to size z_1 on the subpopulation which would participate only in the enlarged programme.

A central condition for the identification of the local average treatment effect (21) is the mean exclusion restriction (17). This mean exclusion restriction combines two conceptually distinct assumptions, which can be seen by rewriting (17) for the potential outcome $Y_{i,Z}^{D_i=1}$ as

$$E\left[Y_{i,Z_i=0}^{D_i=1} | Z_i = 0, \tau_i = t\right] = E\left[Y_{i,Z_i=1}^{D_i=1} | Z_i = 0, \tau_i = t\right] = E\left[Y_{i,Z_i=1}^{D_i=1} | Z_i = 1, \tau_i = t\right] \quad (22)$$

for $t \in \{a, c, d\}$. The first equality sign corresponds to an exclusion assumption on the *individual level*. It is assumed that the potential outcome $Y_i^{D_i=1}$ for individual i is unaffected by an exogenous change in Z_i. It rules out any direct systematic impact of Z on the potential outcomes on an individual level (and this assumption is satisfied for instance if $Y_{i,Z_i=0}^{D_i=1} = Y_{i,Z_i=1}^{D_i=1}$). The second equality sign in (22) represents an unconfoundedness assumption on the *population level*. It assumes that the potential outcome $Y_{i,Z_i=1}^{D_i=1}$ is identically distributed in the subpopulation of individuals for whom the instrument Z_i takes the value 0 and in the subpopulation of individuals with $Z_i = 1$. This assumption rules out selection effects on the population level. Hence the second part of the mean exclusion restriction (22) refers to the composition of individuals for whom $Z = 1$ or $Z = 0$ is observed, whereas the first part refers to how the instrument affects the outcome Y of a particular individual.

The second part of the mean exclusion assumption (22) is trivially satisfied if the instrument Z is randomly assigned. Nevertheless randomization of Z does not guarantee that the exclusion assumption holds on the individual level. On the other hand, if Z is chosen by the individual itself, unconfoundedness of Z (on the population level) is unlikely to hold. For example, Card (1995) uses college proximity as an instrument for estimating the returns to schooling. Living closer to a college is likely to induce some children to obtain more college education. Although it appears reasonable that distance to the nearest college by itself does not affect the subsequent potential labour market outcomes of the child (first part of (22)), it might be that the families who decide to reside nearer or farther to a college are rather different. In this case the instrumental variable is subject to self-selection and the mean exclusion assumption (17) is unlikely to be satisfied.

In situations where the instrumental variable is not randomly assigned, the instrument Z might also be confounded with the potential participation indicators $D_{i,Z_i=0}$ and $D_{i,Z_i=1}$, thus invalidating the unconfounded participation-type assumption (16). E.g. the composition of always-participants, never-participants and compliers might be different among families who decide to reside close to a college than among those who live distant to a college. To identify an average treatment effect when the instrument Z itself is confounded with the potential participation indicators $D_{i,Z_i=0}$, $D_{i,Z_i=1}$ or with the potential outcomes $Y_{i,Z_i}^{D_i=0}$, $Y_{i,Z_i}^{D_i=1}$, it is necessary to consider extended versions of Assumptions 1 and 2 where the unconfounded-participation type and mean exclusion assumptions are required to hold only conditional on all confounding variables X. These extensions are examined, for example, in Frölich (2002).[17]

Identification of treatment effects on the basis of Assumptions 1 to 4 has been applied, for example, in Hearst *et al.* (1986) and Angrist (1990) to estimate the effects of participating in the Vietnam war on mortality and civilian earnings, respectively. A suited instrumental variable was provided through the US conscription policy during the years of the Vietnam war, which conscripted individuals on the basis of randomly drawn birth dates. Imbens and van der Klaauw (1995) used variations in the compulsory conscription policy in the Netherlands during World War II to estimate the effect of veteran status on earnings. Angrist and Krueger (1991) estimated the returns to schooling using the quarter of birth as an instrumental variable for educational attainment. According to US compulsory school attendance laws, compulsory education ends when the pupil reaches a certain age, and thus, the month in which termination of the compulsory education is reached depends on the birth date. Since the school year starts for all pupils in summer/autumn, the minimum education varies with the birth date, which can be exploited to estimate the impact of an additional year of schooling on earnings.

2.8 *Regression-discontinuity Design*

A particular type of instrumental variable identification is exploited in the regression-discontinuity design. This approach uses discontinuities in the programme selection process to identify a causal effect. Suppose a (continuous)

variable Z influences an outcome variable Y and also another variable D, which itself affects the outcome variable Y. Hence, Z has a direct impact on Y as well as an indirect impact on Y via D. This latter impact represents the causal effect of D on Y, which can be identified if the direct and the indirect impacts of Z on Y can be told apart. In the case that the direct impact of Z on Y is known to be smooth but the relationship between Z and D is discontinuous, any discontinuities (jumps) in the observed relationship between Z and Y at locations where the relation Z to D is discontinuous can be attributed to the indirect impact of Z on Y via D.

This idea has been utilized by Thistlethwaite and Campbell (1960) to estimate the effect of receiving a National Merit Award on subsequent career aspirations. Since the Award is only granted if a test score Z exceeds a certain threshold z_0, the treatment status D (Award granted: $D=1$, not granted: $D=0$) depends in a discontinuous way on the test score Z. Let Y^1 and Y^0 denote the corresponding potential outcomes. Certainly the test score Z influences not only D but also affects the potential outcomes directly. Hence Z is not a proper instrumental variable, since the exclusion restriction is not satisfied. Nevertheless, in a small neighbourhood around the discontinuity at z_0, the direct impact of Z on the potential outcomes is likely to vary only a little with Z. Hence *locally* the instrumental variable assumptions (exclusion restriction, monotonicity) are satisfied, and the difference between the mean outcome for individuals just above the threshold z_0 and the mean outcome for individuals just below the threshold represents the causal effect $E[Y^1 - Y^0|Z=z_0]$. Again, this is a kind of local average treatment effect since it is identified only for the subpopulation of individuals with test score equal to z_0.

In the above example programme participation status $D=1(Z>z_0)$ is a deterministic function of Z, which is also called a *sharp design* (Trochim, 1984) since *all* individuals change programme participation status exactly at z_0. This requires a strictly rule-based programme selection process (such as age limits or other eligibility criteria). For example, Hahn *et al.* (1999) analyze the effect of antidiscrimination laws on the employment of minority workers by exploiting the fact that only firms with more than 15 employees are subject to these antidiscrimination laws.

Often, however, the participation decision is not completely determined by Z, even in a rule-based selection process. Case workers may have some discretion about whom they offer a programme, or they may base their decision also on criteria that are unobserved to the econometrician. Additionally, individuals offered a programme may decline participation. In this *fuzzy design* not all individuals would change programme participation status from $D=0$ to $D=1$ if Z were increased from $z_0-\varepsilon$ to $z_0+\varepsilon$. Rather, the relation between Z and D may be discontinuous at z_0 only on average. In the *fuzzy* design the *expected* value of D given Z (which is the probability of treatment receipt) is supposed to be discontinuous at z_0:

$$\lim_{\varepsilon\to 0} E[D|Z=z_0+\varepsilon] \neq \lim_{\varepsilon\to 0} E[D|Z=z_0-\varepsilon]. \tag{23}$$

For example, van der Klaauw (2002) analyzes the effect of financial aid offers to college applicants on their probability of subsequent enrollment. College applicants are ranked according to their test score achievements into a small number of categories. The amount of financial aid offered depends largely on this classification. Yet, the financial aid officer also takes other characteristics into account, which are not observed by the econometrician. Hence the treatment assignment is not a deterministic function of the test score Z, but the conditional expectation function $E[D|Z]$ displays jumps because of the test-score rule.[18]

Hahn *et al*. (2001) analyze nonparametric identification in the case of a fuzzy regression-discontinuity design, where $D \in \{0, 1\}$ is a random function of Z but $E[D|Z]$ is discontinuous at z_0.[19] Since Z may also influence the potential outcomes directly, the treatment effect is not identified without further assumptions. Supposing that the direct influence of Z on the potential outcomes is continuous, the potential outcomes change little when Z is varied within a small neighbourhood. Under a localized version of the unconfounded-participation-type assumption, the exclusion restriction and the monotonicity assumption (discussed in the previous section on instrumental variables identification), they show that the average treatment effect on the *local compliers* is identified as

$$\lim_{\varepsilon \to 0} E\left[Y^1 - Y^0 | D(Z = z_0 + \varepsilon) = 1, D(Z = z_0 - \varepsilon) = 0\right] = \frac{\lim_{\varepsilon \to 0} E[Y|Z = z_0 + \varepsilon] - \lim_{\varepsilon \to 0} E[Y|Z = z_0 - \varepsilon]}{\lim_{\varepsilon \to 0} E[D|Z = z_0 + \varepsilon] - \lim_{\varepsilon \to 0} E[D|Z = z_0 - \varepsilon]}. \quad (24)$$

The *local compliers* is the group of individuals whose Z value lies in a small neighbourhood of z_0 and whose treatment status D would change from 0 to 1 if Z were changed exogenously from $z_0 - \varepsilon$ to $z_0 + \varepsilon$. As a special case, in the sharp design all individuals are locally compliers and change their treatment status at z_0. Thus the denominator of (24) would be 1.

The regression-discontinuity approach permits identification of a treatment effect under weak conditions. In particular a type of instrumental variable assumption needs to hold only *locally*. On the other hand, the average treatment effect is identified only for the local compliers. And due to its local nature of identification, no $\sqrt{n}$-consistent estimator can exist for estimating it.

2.9 *Bounds*

In evaluation settings where none of the above outlined identification strategies is feasible, it might still be possible to estimate intervals wherein the average treatment effects lie. Such bounds on the treatment effects have been derived by Manski (1989, 1990, 1997) in a series of papers. Consider the simplest case where no further information is available except that the outcome variables Y^r and Y^s have bounded support: $Y^r \in [\underline{Y}, \bar{Y}]$, $Y^s \in [\underline{Y}, \bar{Y}]$. With these bounds on the support of Y^r, the expected outcome of Y^r is bounded to lie in the interval

$$E[Y^r] = E[Y^r|D = r] \cdot P_{D=r} + E[Y^r|D \neq r] \cdot P_{D\neq r}$$
$$\in \left[E[Y^r|D = r]P_{D=r} + P_{D\neq r}\underline{Y}, E[Y^r|D = r]P_{D=r} + P_{D\neq r}\bar{Y}\right],$$

where $P_{D=r} = P(D = r)$ is shorthand notation for the size of the subpopulation participating in programme r. The width of this interval decreases with $P(D = r)$ since the expected outcome is only identified in this subpopulation. Using an analogous argument for $E[Y^s]$, the average treatment effect $E[Y^r - Y^s]$ can be bounded to lie in the interval

$$E[Y^r - Y^s] \in [E[Y^r|D = r]P_{D=r} - E[Y^s|D = s]P_{D=s} + P_{D\neq r}\underline{Y} - P_{D\neq s}\bar{Y},$$
$$E[Y^r|D = r]P_{D=r} - E[Y^s|D = s]P_{D=s}$$
$$+ P_{D\neq r}\bar{Y} - P_{D\neq s}\underline{Y}].$$

The width of this interval is

$$(\bar{Y} - \underline{Y})(P(D \neq r) + P(D \neq s))$$
$$= (\bar{Y} - \underline{Y})(2 - P(D = r) - P(D = s))$$

and narrows for larger probabilities to participate in programme r or s. However, even in the binary treatment case, where $P(D = r) + P(D = s) = 1$, the interval width is $(\bar{Y} - \underline{Y})$ and a zero treatment effect cannot be ruled out. Hence without further assumptions, such as monotonous instrumental variables as in Manski (1997) or Manski and Pepper (2000), these bounds on the treatment effects provide only limited information.

3. Estimation of Mean Counterfactual Outcomes

After a proper identification strategy has been established, the treatment effects of interest can be estimated. Although identification is the fundamental and crucial task in programme evaluation, the choice of an appropriate estimator can still make a difference. Keeping in mind that collection of informative and reliable data on participants and non-participants is often costly and time-consuming, an estimator should be chosen which uses the available information in an efficient way. This seems to be particularly important in programme evaluation where the interpretation and the consequences of evaluation studies often hinge on whether a programme effect is estimated as statistically significant or not. Even in the case where two estimators generate the same point estimates, the interpretation of these estimates depends on their variability, such that the estimate according to the more precise estimator is more likely to be statistically significant. Since insignificant programme effects are often interpreted as 'no effect', the choice of the estimator affects ceteris paribus the odds of the evaluation study's conclusions.

An important element in the estimation of programme effects is the conditional expected potential outcome Y^s weighted by the distribution of X among the participants in programme r

$$(EY)_{s|r} = \int E[Y^s|X = x, D = s] \cdot f_{X|D=r}(x)dx. \tag{25}$$

Under the control-for-confounding-variables approach, the term (25) equals the average counterfactual outcome for the participants in programme r:

$$E[Y^s|D = r]$$

since

$$E[Y^s|D = r] = E[E[Y^s|X, D = s]|D = r].$$

The expression (25) represents the adjustment of the expected potential outcome Y^s for the distribution of the confounding variables among the participants in programme r. The estimation of (25) is the decisive part for the estimation of average treatment effects on the treated $E[Y^r - Y^s|D = r]$ since $E[Y^r|D = r]$ can be estimated simply by the sample mean of the participants in programme r.

The estimation of objects like (25) is also central to the estimation of average treatment effects $E[Y^r] - E[Y^s]$ and average treatment outcomes $E[Y^s]$ because $E[Y^s]$ can be written as

$$E[Y^s] = E[Y^s|D = s] \cdot P(D = s) + E[Y^s|D \neq s] \cdot P(D \neq s),$$

where $E[Y^s|D = s]$, $P(D = s)$ and $P(D \neq s)$ can be estimated by sample means. The estimation of the average counterfactual outcome

$$\begin{aligned} E[Y^s|D \neq s] &= E[E[Y^s|X, D = s]|D \neq s] \\ &= \int E[Y^s|X = x, D = s] \cdot f_{X|D\neq s}(x)dx \end{aligned} \tag{26}$$

proceeds by weighting the conditional expectation function $E[Y^s|X]$ by the density $f_{X|D\neq s}$, i.e. by the density of X among the population not participating in programme s. This is analogous to (25) with $f_{X|D=r}$ replaced by the density function $f_{X|D\neq s}$.

In addition, also the difference-in-difference or predictable-bias approach, discussed above, relies on weighting conditional expectation functions by density functions of other subpopulations to estimate the selection bias in different periods. In particular, estimates of $\int E\left[Y_l^0|X = x, D = 0\right] f_{X|D=r}(x)dx$ and $\int E[Y_\tau|X = x, D = 0] f_{X|D=r}(x)dx$ are required in (12). Thus the estimation of objects like (25) forms a crucial building block for many programme evaluation estimators. In the following, various nonparametric estimators of (25) and their properties are discussed.

The general framework for nonparametric *covariate-distribution adjustment* can be characterized as follows: Data on an outcome variable Y^s and covariates X are sampled randomly from a 'source' population (e.g. the participants in programme s). In addition, a second sample, drawn from a 'target' population (e.g. the participants in programme r), is available, which contains only information on the covariates X but not on their potential outcome variable Y^s. Denote the observations sampled from the source population by $\{(Y_{si}, X_{si})\}_{i=1}^{n_s}$, with $f_s(x)$ the

density of X_{si}. Denote the sample drawn from the target population by $\{X_{rj}\}_{j=1}^{n_r}$, with density function f_r. Suppose further that the support of X in the target population is a subset of the support of X in the source population, to ensure that the conditional expectation function in the source population is identified at every X_{rj}.[20] This notation is shorthand for the sample of participants in programme s: $\{(Y_{si}, X_{si})\}_{i=1}^{n_s} \equiv \{(Y_i^s, X_i)|D_i = s\}_{i=1}^{n}$ and the sample of participants in programme r: $\{X_{rj}\}_{j=1}^{n_r} \equiv \{X_i|D_i = r\}_{i=1}^{n}$, where $f_s \equiv f_{X|D=s}$ and $f_r \equiv f_{X|D=r}$ and n_s and n_r are the number of observed participants in programme s and programme r, respectively. (For the estimation of average counterfactual outcomes (26) the target population needs to be redefined correspondingly such that the target sample $\{X_{rj}\}_{j=1}^{n_r}$ represents all observations that participated in any programme except s.) Nonparametric covariate-distribution adjustment proceeds by weighting the conditional expectation function $E[Y^s|X]$ in the source population by the density f_r in the target population.

3.1 *Generalized Matching Estimators*

A class of estimators of (25) is obtained by replacing the target distribution f_r by its empirical distribution function in the target sample and estimating nonparametrically the conditional expectation function $m_s(x) = E[Y^s|X = x,\ D = s]$ from the source sample. This yields the *generalized matching estimator*

$$\frac{1}{n_r}\sum_{j=1}^{n_r}\hat{m}_s(X_{rj}), \tag{27}$$

where $\hat{m}_s(x)$ is an estimate of $m_s(x)$. This estimator adjusts the conditional expected outcome for the distribution of X in the target population by evaluating and averaging $m_s(x)$ only at the values X_{rj} that are observed in the target sample. A variety of estimators to estimate $m_s(X_{rj})$ from the source sample $\{Y_{si}, X_{si}\}_{i=1}^{n_s}$ have been suggested:

A simple and common method to implement the estimator (27) is based on *pair-matching* (Rubin, 1974). Pair-matching proceeds by finding for each observation of the target sample an observation of the source sample with identical (or very similar) covariates X. These 'matched' source sample observations mirror the covariate distribution of the target sample and their average outcome provides an estimate of (25). In other words, pair-matching estimates $\hat{m}_s(X_{rj})$ by the observed outcome Y_{si} of that source sample observation i which is most similar in its covariates X_{si} to X_{rj}. Building on this idea, alternative estimators estimate $\hat{m}_s(X_{rj})$ by a weighted mean of the observed outcomes of those source sample observations that are similar to X_{rj}, and accordingly (27) is called a generalized matching estimator.

Whereas pair-matching assigns zero weights to all observations except the closest observation to X_{rj}, parametric regression-based matching estimators use all observations of the source sample regardless of their similarity to X_{rj}. In particular, least squares regression estimates $\hat{m}_s(x)$ as $x'\hat{\beta}$ with $\hat{\beta} = (\mathbf{X}_s'\mathbf{X}_s)^{-1}(\mathbf{X}_s'\mathbf{Y}_s)$, where $\mathbf{X}_s$ is the matrix of all X_{si}' and $\mathbf{Y}_s$ is the column vector containing all Y_{si}. Hence the imputed value at X_{rj} is $\hat{m}_s(X_{rj}) = X_{rj}'\hat{\beta}$ and the *least squares matching* estimator of (25) is

$$\bar{X}_r'(\mathbf{X}_s'\mathbf{X}_s)^{-1}(\mathbf{X}_s'\mathbf{Y}_s), \tag{28}$$

where $\bar{X}_r = (1/n_r)\Sigma X_{rj}$ is the average of the covariates in the target sample.

In between these two extremes, pair-matching and least-squares matching, exist a variety of nonparametric estimators of $m_s(X_{rj})$, which take account only of the source sample observations that lie in a neighbourhood of X_{rj} and downweight observations according to their dissimilarity to X_{rj}. These include k-nearest neighbours and local polynomial regression, which lead to the *k-NN matching* and the *local polynomial matching* estimator, respectively. Consistency of the matching estimator requires consistent estimation of $m_s(x)$, which in turn requires that the local neighbourhood of X_{rj} shrinks with increasing sample size. Hence the least squares matching estimator is in general inconsistent, whereas pair-matching and *k-NN* matching and local polynomial matching with an appropriately chosen bandwidth value are consistent. Thus in principle, implementation of the generalized matching estimator is straightforward by choosing a consistent nonparametric regression estimator and averaging the imputed values $\hat{m}_s(X_{rj})$ for the target observations.

In practice, however, nonparametric covariate-distribution adjustment often needs to be performed with respect to a high-dimensional X vector. For example, in the control-for-confounding-variables approach, X includes all variables that affect the participation decision and the potential outcome. Nonparametric estimation of the regression function $m_s(x) = E[Y^s|X = x,\ D = s]$ becomes then rather difficult since the convergence rate of nonparametric regression estimators decreases with the number of (continuous) covariates (Stone, 1980). Pair-matching, for example, would require to find for each target sample observation a source sample observation which is identical (or very similar) in all characteristics. To circumvent this dimensionality problem, similarity between observations is often measured through a multivariate distance metric, which maps the mismatch in the characteristics onto the real line. One such metric is the Mahalanobis distance $\langle \cdot,\cdot \rangle_M$, which weights the distance in the covariates by the inverse of their variance matrix

$$\langle X_{si}, X_{rj} \rangle_M = (X_{si} - X_{rj})' \left[\widehat{\mathbf{Var}}(X)\right]^{-1} (X_{si} - X_{rj}),$$

where $\widehat{\mathbf{Var}}(X)$ is an estimate of the variance of X in the source or the target population or a weighted average of these. Pair-matching on the Mahalanobis distance proceeds by finding for each target sample observation X_{rj} the source sample observation with the smallest distance $\langle X_{si}, X_{rj} \rangle_M$.

3.2 *The Propensity Score*

However, the choice of a distance metric (such as the Mahalanobis distance) to reduce the dimensionality is rather ad-hoc. Besides the fact that a different distance metric might have produced rather different results, it is not even guaranteed that the estimate is consistent. Nevertheless, a distance metric based on the balancing score property of the propensity score has favourable theoretical properties. For the case of binary treatment evaluation ($R = 2$), Rosenbaum and

Rubin (1983) showed that conditional independence (4) of programme selection and potential outcomes given X also implies independence conditional on the (one-dimensional) probability to participate in the programme given X, which they called the *propensity score*. Imbens (2000) and Lechner (2001a) generalized this result to the evaluation of multiple treatments ($R > 2$), where the appropriate propensity score is the probability to participate in treatment s for an individual who participates either in treatment r or s and has characteristics X.

Suppose that the potential outcome Y^s conditional on X is identically distributed in source and target population

$$Y^s \perp\!\!\!\perp D | X, D \in \{r, s\}, \tag{29}$$

which is a slightly weaker version of the conditional independence assumption (4). Define the (one-dimensional) propensity score $p^{s|rs}(x)$ as the probability of belonging to the target population instead of belonging to the source population

$$p^{s|rs}(x) = P(D = s | X = x, D \in \{r, s\}) = \frac{p^s(x)}{p^r(x) + p^s(x)}, \tag{30}$$

where $p^r(x) = P(D = r | X = x)$ and $p^s(x) = P(D = s | X = x)$. Then, as shown by Lechner (2001a), conditional independence on X implies conditional independence on the propensity score $p^{s|rs}$

$$\Longrightarrow Y^s \perp\!\!\!\perp D | p^{s|rs}(X), D \in \{r, s\}. \tag{31}$$

Proof. The proof is adopted from Lechner (2001a). Since conditional independence in (31) is required only with respect to the subpopulations r and s, the participation indicator $D \in \{r, s\}$ is binary and all that needs to be shown is that

$$P\Big(D = s | Y^s, p^{s|rs}(X), D \in \{r, s\}\Big) = P\Big(D = s | p^{s|rs}(X), D \in \{r, s\}\Big).$$

By the law of total probability and using (29):

$$\begin{aligned}
&P\Big(D = s | Y^s, p^{s|rs}(X), D \in \{r, s\}\Big) \\
&= E\Big[P(D = s | Y^s, X, D \in \{r, s\}) | Y^s, p^{s|rs}(X), D \in \{r, s\}\Big] \\
&= E\Big[p^{s|rs}(X) | Y^s, p^{s|rs}(X), D \in \{r, s\}\Big] \\
&= p^{s|rs}(X) \\
&= E\Big[p^{s|rs}(X) | p^{s|rs}(X), D \in \{r, s\}\Big] \\
&= E\Big[P(D = s | X, D \in \{r, s\}) | p^{s|rs}(X), D \in \{r, s\}\Big] \\
&= P\Big(D = s | p^{s|rs}(X), D \in \{r, s\}\Big).
\end{aligned}$$

The intuition behind (31) is that the propensity score balances the distribution of X in the source and the target population. In other words, conditional on $p^{s|rs}$ the distribution of X is identical in the source and the target population[21]

$$X \perp\!\!\!\perp D | p^{s|rs}(X), D \in \{r, s\}.$$

Although an observation of the source sample and an observation of the target sample with the same propensity score value $p^{s|rs}$ do not necessarily have the same X value (thus preventing the application of (29)), the probability that X equals a particular value is the same for both observations and the conditional independence assumption (29) can be invoked separately at every possible X value.

The validity of (31) implies that the counterfactual outcome $E[Y^s|D=r]$ can be estimated consistently by solely adjusting the distribution of the propensity score $p^{s|rs}$:

$$\begin{aligned} E[Y^s|D=r] &= E\Big[E\Big[Y^s|p^{s|rs}(X), D=s\Big]|D=r\Big] \\ &= \int E\Big[Y^s|p^{s|rs}, D=s\Big] \cdot f_{p^{s|rs}|D=r}(p^{s|rs}) \cdot dp^{s|rs}, \end{aligned} \tag{32}$$

where $f_{p^{s|rs}|D=r}$ is the distribution of the propensity score in the target population. Hence matching on the one-dimensional propensity score $p^{s|rs}$ instead on X gives a consistent estimator of the counterfactual mean (25). For example, if pair-matching is used, it suffices to find pairs of participants and non-participants that have the same propensity score. They no longer need to be identical on all X covariates.

An analogous relationship holds for the estimation of the counterfactual mean $E[Y^s|D \neq s]$ (26) where the target population consists of all subpopulations which do not participate in programme s. The appropriate propensity score is $p^s(x)$ since $P(D=s|X=x) + P(D \neq s|X=x)$ add up to one in (30), and consequently a propensity score matching estimator based on p^s

$$\begin{aligned} E[Y^s|D \neq s] &= E[E[Y^s|p^s(X), D=s]|D \neq s] \\ &= \int E[Y^s|p^s, D=s] \cdot f_{p^s|D \neq s}(p^s) \cdot dp^s \end{aligned}$$

is consistent.

Remarkably, propensity score matching can even be used for estimating (25) in situations where the conditional independence assumption is not valid. The equality of propensity score matching and matching on X

$$\begin{aligned} &\int E[Y^s|X=x, D=s] \cdot f_{X|D=r}(x) dx \\ &= \int E[Y^s|p^{s|rs}, D=s] \cdot f_{p^{s|rs}|D=r}(p^{s|rs}) \cdot dp^{s|rs} \end{aligned} \tag{33}$$

is a mechanical result of the balancing property of the propensity score and independent of any properties of the potential outcomes. As a consequence,

propensity score matching can also be used in the difference-in-difference or predictable-bias evaluation approach, which is often pursued when the conditional independence assumption appears to be controversial. Hence $E[E[Y_t^0 - Y_\tau|X, D=0]|D=r]$ in (12) can be estimated by propensity score matching as

$$\begin{aligned}
&E\left[E\left[Y_t^0 - Y_\tau|X, D=0\right]|D=r\right] \\
&= E\left[E\left[Y_t^0 - Y_\tau|p^{0|r0}(X), D=0\right]|D=r\right] \\
&= \int E\left[Y_t^0|p^{0|r0}(X), D=0\right] \cdot f_{p^{0|r0}|D=r}(p^{0|r0}) \cdot dp^{0|r0} \\
&\quad - \int E\left[Y_\tau|p^{0|r0}(X), D=0\right] \cdot f_{p^{0|r0}|D=r}(p^{0|r0}) \cdot dp^{0|r0},
\end{aligned}$$

where $p^{0|r0}$ is the appropriate propensity score. Notice that panel data is not required since the covariate adjustment can proceed separately for Y_t^0 and Y_τ.

Propensity score matching circumvents the dimensionality problem since the nonparametric regression needs to be performed only with respect to the one-dimensional propensity score and thus avoids the so-called 'curse of dimensionality'. For this reason, propensity score matching has been used in many applied evaluation studies, e.g. Brodaty *et al.* (2001), Dehejia and Wahba (1999), Frölich *et al.* (2000), Gerfin and Lechner (2002), Heckman *et al.* (1997, 1998), Jalan and Ravallion (2002), Larsson (2000), Lechner (1999), Puhani (1999) etc. However, in most cases the propensity scores themselves are unknown and need to be estimated consistently. Parametric estimation of the propensity scores for the evaluation of multiple treatments is discussed in Lechner (2002a), who compares binary probit models, multinomial logit models and simulated multinomial probit models. Semiparametric estimation of the propensity score is analyzed in Todd (1999). Propensity score matching proceeds then with respect to the estimated propensity score.

Although matching on the propensity score balances the distribution of X in source and target sample and thus provides a consistent estimate of the counterfactual mean (25), it may not be the most precise estimator in finite samples, as the components of X might affect the propensity score $p^{s|rs}(x)$ and the conditional expectation function $m_s(x) = E[Y^s|X=x, D=s]$ to different degrees. Some covariates may affect strongly the conditional expectation $m_s(x)$ but have only little weight among the determinants of the participation probability $p^{s|rs}(x)$, whereas other covariates may be important determinants of $p^{s|rs}(x)$ but have little impact on $m_s(x)$. In this case, observations with a similar propensity score value are also likely to be similar with respect to the latter covariates, but may not be so with respect to the former covariates, since their influence on $p^{s|rs}$ is small. Hence observations with identical propensity score values may be very dissimilar with respect to the main determinants of $m_s(x)$. However, as the main purpose of matching is to balance particularly the covariates that are highly influential on the potential outcome, conditioning on the propensity score may not be the most efficient method in finite samples. To achieve a balancing of the relevant variables in finite samples, matching

on the principal covariates determining $m_s(x)$ or on the propensity score and a subset of covariates might be more appropriate. The latter refers to the *augmented propensity score* approach, where matching proceeds on a vector $(p^{s|rs}, \tilde{X})$ consisting of the propensity score $p^{s|rs}$ and a subset of covariates $\tilde{X}$, which are important determinants of $m_s(x)$ but might be 'under-represented' in $p^{s|rs}$. For example, in the evaluation of active labour market programmes, it might occur that programme assignment decisions are largely driven by the employment offices' case workers whereas subsequent labour market programmes depend mainly on individual characteristics. The use of the augmented propensity score has already been suggested by Rosenbaum and Rubin (1983) in their analysis of balancing scores that are 'finer' than the propensity score. All the above discussed balancing properties hold as well with the augmented propensity score, as can easily be seen by repeating the proof. The augmented propensity score has, for example, been used in Lechner (1999), with respect to time-varying and time-invariant covariates, and in Lechner (2002a), where he compares propensity score matching on $p^{s|rs}(x)$ to matching on $(p^s(x), p^r(x))$.[22]

3.3 *The Re-weighting Estimator*

An alternative estimation strategy to adjust for the differences in the covariate composition among source and target population relies on weighting the observed outcomes by the density ratio of X, which is considered in Horvitz and Thompson (1952), Imbens (2000), Hirano *et al.* (2003) and Ichimura and Linton (2001). Since observations (Y^s, X) at X locations where the density $f_s(x)$ in the source population is large are relatively over-represented and observations where $f_s(x)$ is small are relatively under-represented, a weighted average of Y^s should downweight the former observations and upweight the latter observations by the ratio f_r/f_s of the density of X in the target and the source population. Rewriting the object of interest (25)

$$\int E[Y^s|X=x, D=s]\cdot f_r(x)dx = \int E[Y^s|X=x, D=s]\cdot\frac{f_r(x)}{f_s(x)}f_s(x)dx$$
$$= E\left[Y^s\frac{f_r(X)}{f_s(X)}|D=s\right]$$

suggests the *re-weighting estimator*

$$\frac{1}{n_s}\sum_{i=1}^{n_s} Y_{si}\cdot\frac{\hat{f}_r(X_{si})}{\hat{f}_s(X_{si})} \tag{34}$$

as an alternative estimator of (25), where the covariate densities f_s and f_r can be estimated from the source and the target sample, respectively. By multiplying the observations Y^s of the source sample with the density ratio f_r/f_s the estimator rectifies the relative over/under-representation of source sample observations at large/small values of f_s.

The re-weighting estimator can also be written in terms of the propensity score by noting that the propensity score ratio equals the density ratio times the size ratio of the subpopulations:[23]

$$\frac{p^{s|rs}(X)}{1-p^{s|rs}(X)} = \frac{f_s(X)}{f_r(X)}\frac{P(D=s)}{P(D=r)}. \tag{35}$$

The relative size of the source population to the target population, $P(D=s)/P(D=r)$, can be consistently estimated by n_s/n_r if sampling from the source and the target population was done with the same probability. Accordingly (25) can also be expressed as

$$\int E[Y^s|X=x, D=s]\cdot f_r(x)dx = E\left[Y^s\cdot\frac{1-p^{s|rs}(X)}{p^{s|rs}(X)}\frac{P(D=s)}{P(D=r)}|D=s\right],$$

and estimated as

$$\frac{1}{n_s}\sum_{i=1}^{n_s} Y_{si}\cdot\frac{1-p^{s|rs}(X_{si})}{p^{s|rs}(X_{si})}\frac{n_s}{n_r} = \frac{1}{n_r}\sum_{i=1}^{n_s} Y_{si}\cdot\frac{1-p^{s|rs}(X_{si})}{p^{s|rs}(X_{si})}. \tag{36}$$

Again the conditional independence assumption (4) is not needed to justify using the propensity score for consistent estimation of (25).

3.4 *Properties of Treatment Effect Estimators*

The asymptotic properties of the generalized matching estimator and the re-weighting estimator have been studied in the binary treatment framework ($R=2$) by Hahn (1998), Heckman *et al.* (1998), Hirano *et al.* (2003), Ichimura and Linton (2001) and Abadie and Imbens (2001) under the conditional independence assumption (4). Hahn (1998) derived the *$\sqrt{n}$-semiparametric variance bounds* for nonparametric estimation of the average treatment effect and the average treatment effect on the treated. Adopted to the multiple treatment framework, the variance bound for estimating the expected potential outcome $E[Y^s]$ is

$$\begin{aligned}
&\frac{1}{n}\int\left(\frac{\sigma_s^2(x)}{p^s(x)}+(E[Y^s|X=x]-E[Y^s])^2\right)f_X(x)dx\\
&=\frac{1}{n}\left(\int\frac{\sigma_s^2(x)}{P^s}\frac{f_X^2(x)}{f_s^2(x)}f_s(x)dx+\int(E[Y^s|X=x]-E[Y^s])^2f_X(x)dx\right)\\
&=\frac{1}{n}\left(\frac{1}{P^s}\underset{f_s(x)}{E}\left[\sigma_s^2(x)\frac{f_X^2(x)}{f_s^2(x)}\right]+\underset{f(x)}{\text{var}}\,E[Y^s|X]\right),
\end{aligned}$$

and for estimating the average treatment effect $E[Y^r-Y^s]$ is

$$\frac{1}{n}\int\left(\frac{\sigma_r^2(x)}{p^r(x)}+\frac{\sigma_s^2(x)}{p^s(x)}+(E[Y^r-Y^s|X=x]-E[Y^r-Y^s])^2\right)f_X(x)dx$$
$$=\frac{1}{n}\left(\frac{1}{P^s}\underset{f_s(x)}{E}\left[\sigma_s^2(X)\frac{f_X^2(X)}{f_s^2(X)}\right]+\frac{1}{P^r}\underset{f_r(x)}{E}\left[\sigma_r^2(X)\frac{f_X^2(X)}{f_r^2(X)}\right]+\underset{f(x)}{\text{var}}\,E[Y^r-Y^s|X]\right) \quad (37)$$

where $P^s=P(D=s)$ and $\sigma_s^2(x)=\text{var}(Y^s|X=x)$.

For estimating the mean counterfactual outcome $E[Y^s|D=r]$ and the average treatment effect on the treated $E[Y^r-Y^s|D=r]$ only the observations on the participants in programme r and programme s are used. The observations on the participants in the other programmes are not informative, neither for the estimation of $E[Y^s|X=x, D=s]$ nor for the estimation of the distribution of X in the subpopulation of participants in programme r (programme-r-subpopulation). Hence observations with $D\notin\{r, s\}$ are irrelevant. (See also the discussion below on the value of the propensity score.) Consequently, the normalizing factor for the asymptotic variance is $1/(n_r+n_s)$ instead of $1/n$.

The variance bound for estimating the mean counterfactual outcome $E[Y^s|D=r]$ is

$$\frac{1}{n_r+n_s}\int\left(\frac{\sigma_s^2(x)p^{r|rs^2}(x)}{P^{r|rs^2}p^{s|rs}(x)}+\frac{p^{r|rs}(x)}{P^{r|rs^2}}(E[Y^s|X=x]-E[Y^s|D=r])^2\right)f_{rs}(x)dx$$
$$=\frac{1}{n_r+n_s}\left(\int\frac{\sigma_s^2(x)}{P^{s|rs}}\frac{f_r^2(x)}{f_s(x)}dx+\int\frac{f_r(x)}{P^{r|rs}}(E[Y^s|X=x]-E[Y^s|D=r])^2dx\right)$$
$$=\frac{1}{n_r+n_s}\frac{1}{P^{r|rs}}\left(\frac{P^r}{P^s}\underset{f_s(x)}{E}\left[\sigma_s^2(X)\frac{f_r^2(X)}{f_s^2(X)}\right]+\underset{f_r(x)}{\text{var}}\,E[Y^s|X]\right) \quad (38)$$

where $P^{r|rs}=P(D=r|D\in\{r, s\})$ and $f_{rs}(x)$ is the density of X in the union of the programme-r- and programme-s-subpopulations with $f_{rs}(x)=f_r(x)P^{r|rs}+f_s(x)P^{s|rs}$ and $p^{r|rs}(x)=f_r(x)P^{r|rs}/f_{rs}(x)$.[24] The variance bound of the average treatment effect on the treated $E[Y^r-Y^s|D=r]$ is

$$\frac{1}{n_r+n_s}\int\left(\begin{array}{c}\dfrac{\sigma_r^2(x)p^{r|rs}(x)}{P^{r|rs^2}}+\dfrac{\sigma_s^2(x)p^{r|rs^2}(x)}{P^{r|rs^2}p^{s|rs}(x)}\\ \\ +\dfrac{p^{r|rs}(x)}{P^{r|rs^2}}(E[Y^r-Y^s|X=x]-E[Y^r-Y^s|D=r])^2\end{array}\right)f_{rs}(x)dx$$
$$=\frac{1}{n_r+n_s}\frac{1}{P^{r|rs}}\left(\underset{f_r(x)}{E}\left[\sigma_r^2(X)\right]+\frac{P^r}{P^s}\underset{f_s(x)}{E}\left[\sigma_s^2(X)\frac{f_r^2(X)}{f_s^2(X)}\right]+\underset{f_r(x)}{\text{var}}\,E[Y^r-Y^s|X]\right). \quad (39)$$

A remarkable result of Hahn (1998) is that a projection on the propensity score (i.e. matching on the propensity score) does not change the variance bound and that knowledge of the true propensity score is not informative for estimating

average treatment effects. The variance bound (37) is the same regardless of whether the propensity score is known. Hence asymptotically the propensity score does not lead to any reduction in dimensionality. However, the variance bound (39) of the average treatment effect on the treated changes when the true propensity score is known. Hahn (1998) attributes this to the 'dimension reduction' property of the propensity score. In my opinion this interpretation is highly misleading. I rather argue that the only value of knowing the true propensity score is that the observed X values of individuals who participated in other programmes than r can be used to improve the estimation of the density $f_{X|D=r}(x)$ among the programme-r-participants.

If the propensity score would indeed contribute to reducing the dimensionality of the estimation problem, it should also help to estimate potential outcomes $E[Y^s]$ and average treatment effects $E[Y^r - Y^s]$ more precisely. On the other hand, the propensity score provides information about the ratio of the density in the source and the target population and thus allows source observations to identify the density of X in the target population and vice versa. Consider the binary treatment case with $r = 1$ (treated participants) and $s = 0$ (non-participants). The (Y, X) observations of the treated sample are informative for estimating $E[Y^1|X]$, whereas the (Y, X) observations of the non-participant sample are informative for estimating $E[Y^0|X]$. Since the joint distribution of Y^1, Y^0 is not identified, the observations of the treated sample cannot assist in estimating $E[Y^0|X]$ and vice versa. The X observations of both samples are useful for estimating the distribution function of X in the population. With this information the average treatment effect can be estimated by weighting the estimates of $E[Y^1|X]$ and $E[Y^0|X]$ by the distribution of X in the population. Knowledge of the propensity score is of no use. Now consider the estimation of the average treatment effect on the treated $E[Y^1 - Y^0|D = 1]$ or of the counterfactual outcome $E[Y^0|D = 1]$. Again the (Y, X) observations of both samples identify the conditional expectation functions separately. These conditional expectation functions are weighted by the distribution of X among the treated, which can be estimated by the empirical distribution function of X in the treated sample. The non-participant observations are not informative for estimating the distribution of X among the treated. However, if the relationship between the distribution of X among the treated and the distribution of X among the non-participants were known, the X observations of the non-participants would be useful for estimating the distribution of X among the treated. Since the propensity score ratio equals the density ratio times the size ratio of the subpopulations (35), and since the relative size of the treated subpopulation $P(D = 1)$ can be estimated precisely, both the treated and the non-participant observations can be used to estimate $f_{X|D=1}$ if the propensity score is known. Consider a simple example: In the case of random assignment with $p^1(x) = 0.5$ for all x, the distribution of X is the same among the treated and the non-participants, and using only the treated observations to estimate $f_{X|D=1}$ would neglect half of the informative observations. With knowledge of the propensity score the counterfactual outcome $E[Y^0|D = 1]$ is identified as

$$E\left[Y^0|D=1\right] = \int E\left[Y^0|X=x, D=0\right] \cdot f_{X|D=1}(x)dx$$
$$= \frac{1}{P(D=1)} \int E\left[Y^0|X=x, D=0\right] p^1(x) \cdot f_X(x)dx \quad (40)$$

and could be estimated by the empirical moment estimator

$$\frac{\sum_{D_i \in \{0,1\}} \hat{m}_0(X_i) p^1(X_i)}{\sum_{D_i \in \{0,1\}} p^1(X_i),}$$

which uses the X observations of both the treated *and* the non-participants. This estimator is suggested by Hahn (1998, Proposition 7) and achieves the variance bound for known propensity score.

The value of knowing the propensity score for estimating the distribution function $f_{X|D=1}$ becomes even more obvious when rewriting (40) as

$$E\left[Y^0|D=1\right] = \int E\left[Y^0|X=x, D=0\right] \cdot f_{X|D=1}(x)dx$$
$$= \frac{P(D=0)}{P(D=1)} \int E\left[Y^0|X=x, D=0\right] \frac{p^1(x)}{1-p^1(x)} \cdot f_{X|D=0}(x)dx,$$

if $p^1(x) \neq 1 \, \forall x$. This suggests the empirical moment estimator

$$\frac{\sum_{D_i=0} \hat{m}_0(X_i) \frac{p^1(X_i)}{1-p^1(X_i)}}{\sum_{D_i=0} \frac{p^1(X_i)}{1-p^1(X_i)}},$$

which uses *only* non-participant observations ($D_i = 0$) to estimate the counterfactual outcome for the treated. Hence with knowledge of the propensity score the counterfactual outcome $E[Y^0|D=1]$ *for the treated* could be estimated nonparametrically even without a single treated observation!

In the case of multiple treatment evaluation there are a variety of propensity scores. Knowledge of $p^{r|rs}$ would allow using the X observations of the s sample to improve the precision of estimating the distribution of X in the r subpopulation. Knowledge of $p^{r|rt}$ would allow using the X observations of a t sample for estimating $f_{X|D=r}(x)$. Knowledge of p^r would allow using all X observations to improve upon the estimation of $f_{X|D=r}(x)$. Hence, in the multiple treatment setting, the variance bound for the average treatment effect on the treated depends on which and how many propensity scores are known.

Besides deriving the efficiency bounds, Hahn (1998) further gives general conditions under which a generalized matching estimator based on a particular nonparametric series regression estimator attains both variance bounds (37) and (39).

Abadie and Imbens (2001) analyzed the asymptotic efficiency of κ-nearest-neighbours matching estimators in estimating average treatment effects when κ is fixed, i.e. when the number of neighbours is fixed and does not grow with increasing sample size.[25] This includes the standard pair-matching estimator ($\kappa = 1$). They consider matching with respect to the X variables and show that (1) these estimators do not attain the variance bound (37) and, hence, are inefficient. (2) The bias term of the estimator is of order $O(n^{-2/c})$ where c is the number of continuous covariates. Consequently, if the number of continuous covariates is 4, the estimator is asymptotically biased. If the number of continuous covariates is even larger, the estimator does no longer converge at rate $\sqrt{n}$. (3) The bias term can be removed through re-centering. However, since re-centring leaves the variance term unchanged, the modified estimator is still inefficient.

Heckman *et al.* (1998) analyzed local polynomial matching for the estimation of average treatment effects on the treated. They prove $\sqrt{n}$-consistency and asymptotic normality when matching with respect to X, with respect to the known propensity score or with respect to the estimated propensity score. The asymptotic distribution consists of a bias term and a variance term. The variance term equals (39) when matching with respect to X. When matching with respect to the known propensity score the variance term corresponds to (39) with X replaced by the propensity score and the density functions $f(x)$ replaced by density functions with respect to the propensity score. Heckman *et al.* (1998) show that this variance term can be either larger or smaller than the variance when matching on X and conclude that neither matching on X nor matching on the propensity score necessarily dominates the other. (However, they ignore in their discussion the different bias terms.) This ambiguity holds also when the propensity score is estimated since the variance contribution due to estimating the propensity score may be small. This variance contribution of estimated-propensity-score matching is derived for a propensity score estimated parametrically or non-parametrically by local polynomial regression with a suitably chosen bandwidth value.

Hirano *et al.* (2003) analyzed the efficiency of the re-weighting estimator for estimating average treatment effects and average treatment effects on the treated. They show that re-weighting using a propensity score estimated by a particular series estimator attains the variance bounds (37) and (39).

Ichimura and Linton (2001) derived higher-order expansions for the re-weighting estimator. Including second-order terms in the analysis is relevant since the first-order approximations do not depend on the smoothing or bandwidth parameters used in the nonparametric first step, such that optimal bandwidth choice cannot be discussed with first-order asymptotics. They consider estimation of the propensity score by local linear regression methods and show that the optimal bandwidth is of order $O(n^{-1/3})$ and $O(n^{-2/5})$ for a bias corrected version.

The analysis of the asymptotic properties of the evaluation estimators implied no firm recommendations on which estimator to use in practice. Generalized matching estimators as well as re-weighting estimators with estimated propensity scores can be efficient. Yet, these considerations may be of limited use for choosing an estimator in a particular application with a given dataset. For

example, although from an asymptotic perspective matching on the propensity score implies no reduction in dimensionality and there are no reasons why matching should not proceed with respect to X, propensity score matching can often be quite useful since 'in practice inference for average treatment effects is often less sensitive to misspecification of the propensity score than to specifications of the conditional expectation of the potential outcomes' (Imbens, 2000).

To examine the properties of these various evaluation estimators in finite samples, Frölich (2001) investigated the mean squared error of different matching and re-weighting estimators in samples of size 40, 200 and 1000, respectively. Matching on an observed covariate as well as matching on an estimated propensity score were analyzed. Regarding the re-weighting estimator, it turned out that it is rather sensitive to the choice of a trimming rule and it performed very poorly without trimming.[26]

The generalized matching estimators, on the other hand, appeared more promising, apart from the inconsistent (global) least squares matching estimator, which performed poorly. The pair-matching estimator, as the benchmark estimator, was compared to three different local polynomial matching estimators: Kernel regression matching, local linear matching, and a ridging local linear matching variant. Ridge regression was proposed by Seifert and Gasser (1996, 2000), among others, to overcome the unbounded variance problem of local linear regression.[27]

The local polynomial matching estimators require the choice of a bandwidth parameter h, which governs the size of the local smoothing neighbourhood. Since their behaviour depends on this bandwidth value, sensitivity to the bandwidth choice is an important issue. As a first result it turned out that although bandwidth choice by cross-validation is inconsistent for choosing h in the covariate-distribution adjustment setting, it nevertheless performed remarkably well in finite samples. Particularly the local linear ridge-regression variant of Seifert and Gasser (1996, 2000) (SG matching) turned out to be very insensitive to the bandwidth choice. Kernel matching was also quite robust to bandwidth choice, although to a much lesser extent. Local linear matching, however, proved less reliable. It appeared sensitive to the bandwidth choice and quite often performed worse than pair-matching.

The relative ordering of the various estimators with respect to mean squared error was remarkably stable across sample sizes and simulation schemes (known and unknown optimal bandwidth, known and unknown propensity score). SG matching, followed by kernel matching, were the most reliable estimators in finite samples. The MSE of SG matching (with cross validation bandwidth-selection) was on average about 25% smaller than the MSE of pair-matching, when matching on an observed covariate. On the other hand, when matching on an estimated propensity score, the reduction in MSE is about 40%.[28] The reason for this difference is that pair-matching becomes less precise (relative to all other estimators) when matching proceeds on estimated covariates, because it compares each target sample observation with only *one* source sample observation. Although the observations within each matched pair are supposed to have identical characteristics, they might indeed be rather different if the covariates are imprecisely

estimated. Hence matching each target sample observation to many source sample observations (as in local polynomial matching) reduces not only the susceptibility of the estimate with respect to the variability in the outcome Y but also with respect to the variance of the *estimated* covariates.

A reduction in MSE of about 40% means that pair-matching needs almost 70% more observations to achieve the same precision as SG matching. If the source sample is larger than the target sample ($n_s > n_r$), the precision gains of local polynomial matching vis-a-vis pair-matching are even larger.

4. Conclusions

In this paper various aspects of programme evaluation have been reviewed. Particular emphasis has been laid on the evaluation of policies consisting of multiple programmes, as they are often found in real world applications. First, different nonparametric strategies to solve the selection bias problem and to identify average treatment effects have been inspected. Crucial issues such as the time structure of the outcome variables, the multiplicity of policy goals and the selection of conditioning variables have been examined and illustrated in the context of the evaluation of active labour market policies. Second, in Section 3, nonparametric estimation of average treatment effects has been studied, including a discussion of the dimension-reducing property of the propensity score and of the asymptotic and finite-sample properties of these estimators.

The evaluation of multiple treatments is in many respects similar to the evaluation of a single treatment (which has been the focus in most evaluation studies). However, many more different treatment effects can be defined in the evaluation of multiple treatments and the analysis becomes more complex with the number of treatments. Furthermore, some of the identification strategies that are fruitful in the evaluation of a single programme, are less appealing in the evaluation of multiple treatments. Particularly, the difference-in-difference and the instrumental variable approach often identify only the treatment effect: participation versus non-participation, and do not permit a comparison *between* the different treatments.

Acknowledgements

I am grateful to Michael Lechner and an anonymous referee for helpful comments and suggestions. This research was supported by the Swiss National Science Foundation, project NSF 4043-058311. Address for correspondence: Markus Frölich, University of St.Gallen, SIAW, Dufourstrasse 48, 9000 St.Gallen, Switzerland; markus.froelich@unisg.ch, www.siaw.unisg.ch/froelich

Notes

1. For example with respect to active labour market programmes, welfare-to-work programmes, vocational training programmes, entrepreneurship promotion schemes, educational programmes, tuition subsidies, sickness rehabilitation programmes or disease prevention programmes.

2. For the evaluation of sequential programmes see Lechner and Miquel (2002).
3. For an introduction to causal reasoning see Holland (1986) and, particularly, Pearl (2000).
4. Considering the potential outcomes Y_i^r as deterministic (non-random) values is only for convenience. The analysis would not change if Y_i^r were random variables.
5. A further discussion about these effects and their appropriateness in different circumstances is found in Heckman *et al.* (1999).
6. These can be assigned to the different programmes and to the 'non-participation' treatment (randomized-out).
7. Even if a proper experiment is conducted, it might still occur by chance that the treatment groups differ substantially in their characteristics particularly if the sample sizes are small. Although the differences in sample means provide unbiased estimates of average treatment effects, adjusting for the differences in the covariates, as discussed in Section 3, can reduce the variance of the estimates (Rubin, 1974).
8. And to delete the 'non-participant' observations for which the assigned start date implies an inconsistency. For example, if unemployment is a basic eligibility condition for participation in an active labour market programme, individuals with an assigned start date *after* the termination of their unemployment spell are discarded (because participation could not have been possible at that date).
9. The control-for-confounding-variables evaluation strategy is widely applied in the evaluation of active labour market programmes, see for instance Heckman *et al.* (1997) and Dehejia and Wahba (1999) for the USA, Lechner (1999) for Eastern Germany, Larsson (2000) for Sweden, Brodaty *et al.* (2001) for France, Gerfin and Lechner (2002) for Switzerland, and Jalan and Ravallion (2002) for Argentina.
10. Provided a random element exists that guarantees that each individual could be assigned to each of the programmes (with non-zero probability). This is the common support condition discussed below.
11. The definition $S^r = \{x : p^r(x) > 0\}$ means $S^r = \{x : p^r(x) > 0$ and $p^r(x)$ is defined$\}$ and, thus, excludes all x where the density $f_X(x)$ in the population is zero.
12. If all pair-wise treatment effects $E[Y^r - Y^s]$ $\forall r, s$ are of interest, Lechner (2002b) suggests to define the effects with respect to the *joint common support* $\bar{S} = \cap_{r=0}^{R-1} S^r$ such that all effects are defined for the same subpopulation and can easier be compared with each other.
13. Estimates of $E[Y_t^r - Y_t^s | D = r]$ for $s \neq 0$ or of $E[Y_t^s]$ for $s \neq 0$ generally cannot be obtained with this approach, since the pre-programme outcomes are only informative about the potential 'non-participation' outcome Y_t^0.
14. For an application of nonparametric difference-in-difference estimation to the evaluation of active labour market programmes in East Germany, see Eichler and Lechner (2002) or Bergemann *et al.* (2000, 2001).
15. For instance, if $B_t = 0$ and $B_\tau \neq 0$ and, erroneously, bias stability (11) is assumed.
16. Although the guideline for the control-for-confounding-variables approach is rather vague, it still gives some indication which variables are relevant and which are not.
17. The confounding variables X are all variables that affect Z *and* $D_{i,Z_i=0}$, $D_{i,Z_i=1}$ or Z *and* $Y_{i,Z_i}^{D_i=0}$, $Y_{i,Z_i}^{D_i=1}$.
18. Further applications of the regression-discontinuity approach include the effects of unemployment benefits on recidivism rates of prisoners (Berk and Rauma, 1983), the effects of classroom size on students' test scores (Angrist and Lavy, 1999), or parents' willingness to pay for higher quality schooling for their children (Black, 1999), among others.

19. Obviously, this includes the sharp design as a special case.
20. Estimation of the common support is discussed in Heckman *et al.* (1998) and Lechner (2002b).
21. This is a mechanical result because $D \in \{r, s\}$ is binary and thus $P(D = s|X, p^{s|rs}(X), D \in \{r, s\}) = p^{s|rs}(X)$.
22. Conditioning on $(p^s(x), p^r(x))$ is 'finer' than conditioning on $p^{s|rs}(x)$, since $p^{s|rs} = p^s/(p^s + p^r)$.
23. Proof: By Bayes' theorem $P(D = r|X) = f_{X|D=r}(X)P(D = r)/f_X(X)$. Hence $p^{s|rs}(x) = \frac{p^s(x)}{p^r(x)+p^s(x)} = \frac{f_{X|D=s}(X)P(D=s)}{f_{X|D=r}(X)P(D=r)+f_{X|D=s}(X)P(D=s)}$ and $\frac{p^{s|rs}(x)}{1-p^{s|rs}(x)} = \frac{f_{X|D=s}(X)P(D=s)}{f_{X|D=r}(X)P(D=r)}$.
24. Because $\frac{f_r(x)P^{r|rs}}{f_{rs}(x)} = \frac{f_r(x)P(D=r)/(P(D=r)+P(D=s))}{f_r(x)P^{r|rs}+f_s(x)P^{s|rs}} = \frac{f_r(x)P(D=r)}{f_r(x)P(D=r)+f_s(x)P(D=s)} = \frac{f_r(x)P(D=r)}{f_r(x)P(D=r)+f_s(x)P(D=s)}$ $= \frac{p^r(x)}{p^r(x)+p^s(x)} = p^{r|rs}(x)$.
25. Consistent estimation of $E[Y^r|X, D = r]$ would require $\kappa \to \infty$ as $n \to \infty$.
26. Further use of the re-weighting estimator would thus require the development of an optimal trimming rule.
27. In the study the implementation proposed by Seifert and Gasser (1996, 2000) was used.
28. The corresponding values for kernel matching are 15% and 30%, respectively. If the source sample is larger than the target sample ($n_s > n_r$), which is often the case in binary treatment evaluation with a large control sample, the precision gains of local polynomial matching *vis-à-vis* pair-matching are even larger.

References

Abadie, A. and Imbens, G. (2001) Simple and bias-corrected matching estimators for average treatment effects, mimeo. Harvard University.

Angrist, J. (1990) Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review*, 80, 313–36.

——— (1998) Estimating labour market impact of voluntary military service using social security data. *Econometrica*, 66, 249–88.

Angrist, J., Imbens, G. and Rubin, D. (1996) Identification of causal effects using instrumental variables. *Journal of American Statistical Association*, 91, 444–72 (with discussion).

Angrist, J. and Krueger, A. (1991) Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106, 979–1014.

——— (1999) Empirical strategies in labor economics. In O. Ashenfelter, and D. Card, (eds), *The Handbook of Labor Economics,* (pp. 1277–1366). New York: North-Holland.

Angrist, J. and Lavy, V. (1999) Using maimonides rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114, 533–75.

Ashenfelter, O. (1978) Estimating the effect of training programms on earnings. *Review of Economics and Statistics*, 6, 47–57.

Barnow, B., Cain, G. and Goldberger, A. (1981) Selection on observables. *Evaluation Studies Review Annual*, 5, 43–59.

Bergemann, A., Fitzenberger, B., Schultz, B. and Speckesser, S. (2000) Multiple active labor market policy participation in East Germany: an assessment of outcomes. *Konjunkturpolitik*, 51, 195–244.

Bergemann, A., Fitzenberger, B. and Speckesser, S. (2001) Evaluating the employment effects of public sector sponsored training in East Germany: conditional difference-in-differences and Ashenfelter's dip, mimeo. University of Mannheim.

Berk, R. and Rauma, D. (1983) Capitalizing on nonrandom assignment to treatments: a regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, 78, 21–27.

Black, S. (1999) Do 'better' schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics*, 114, 577–99.

Blanchard, O. and Diamond, P. (1989) The Beveridge curve. *Brookings Papers on Economic Activity*, 1, 1–60.

——— (1990) The aggregate matching function. In P. Diamond, (ed.) *Growth, Productivity, Unemployment, Essays to Celebrate Bob Solow's Birthday* (pp. 159–201). Cambridge: MIT Press.

Brodaty, T., Crépon, B. and Fougère, D. (2001) Using matching estimators to evaluate alternative youth employment programmes: evidence from France, 1986–1988. In M. Lechner, and F. Pfeiffer, (eds), *Econometric Evaluation of Labour Market Policies* (pp. 85–124). Heidelberg: Physica/Springer.

Card, D. (1995) Using Geographic Variation in College Proximity to estimate the return to schooling. In L. Christofides, E. Grant R. Swidinsky, (eds), *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp,* (pp. 201–22). Toronto: University of Toronto Press.

Cox, D. (1958) *Planning of Experiments*. New York: Wiley.

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

Dehejia, R. and Wahba, S. (1999) Causal effects in non-experimental studies: reevaluating the evaluation of training programmes. *Journal of the American Statistical Association*, 94, 1053–62.

Eichler, M. and Lechner, M. (2002) An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics*, 9, 143–86.

Fisher, R. (1935) *Design of Experiments*. Edinburgh: Oliver and Boyd.

Frölich, M. (2001) Nonparametric covariate adjustment: pair-matching versus local polynomial matching. University of St. Gallen Economics Discussion Paper Series, 2000–17.

——— (2002) Nonparametric IV estimation of local average treatment effects with covariates. IZA Discussion Paper, 588.

Frölich, M., Heshmati, A. and Lechner, M. (2000) A microeconometric evaluation of rehabilitation of long-term sickness in Sweden. *Journal of Applied Econometric*, 19.

Gerfin, M. and Lechner, M. (2002) Microeconometric evaluation of the active labour market policy in Switzerland. *Economic Journal*, 112, 854–93.

Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–31.

Hahn, J., Todd, P. and van der Klaauw, W. (1999) Evaluating the effect of an anti-discrimination law using a regression-discontinuity design. NBER working paper, 7131.

——— (2001) Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69, 201–9.

Hearst, N., Newman, T. and Hulley, S. (1986) Delayed effects of the military draft on mortality: a randomized natural experiment. *New England Journal of Medicine*, 314, 620–4.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998) Characterizing selection bias using experimental data. *Econometrica*, 66, 1017–98.

Heckman, J., Ichimura, H. and Todd, P. (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies*, 64, 605–54.

——— (1998) Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65, 261–94.

Heckman, J., LaLonde, R. and Smith, J. (1999) The economics and econometrics of active labour market programs. In O. Ashenfelter, and D. Card, eds, *The Handbook of Labor Economics,* (pp. 1865–2097). New York: North-Holland.

Heckman, J. and Robb, R. (1985) Alternative methods for evaluating the impact of interventions. In J. Heckman, and B. Singer, (eds), *Longitudinal Analysis of Labour Market Data*, Cambridge: Cambridge University Press.

Heckman, J. and Smith, J. (1995) Assessing the case for social experiments. *Journal of Economic Perspectives*, 9, 85–110.

Heckman, J. and Vytlacil, E. (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings National Academic Sciences USA, Economic Sciences*, 96, 4730–34.

Hirano, K., Imbens, G. and Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71, 1161–89.

Holland, P. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–70.

Horvitz, D. and Thompson, D. (1952) A generalization of sampling without replacement from a finite population. *Journal of American Statistical Association*, 47, 663–85.

Ichimura, H. and Linton, O. (2004) Asymptotic expansions for some semiparametric program evaluation estimators. Forthcoming in Festschrift of Tom Rothenberg.

Imbens, G. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika*, 87, 706–10.

——— (2001) Some remarks on instrumental variables. In M. Lechner and F. Pfeiffer, (eds), *Econometric Evaluation of Labour Market Policies* (pp. 17–42). Heidelberg: Physica/Springer.

Imbens, G. and Angrist, J. (1994) Identification and estimation of local average treatment effects. *Econometrica*, 62, 467–75.

Imbens, G. and van der Klaauw, W. (1995) Evaluating the cost of conscription in the Netherlands. *Journal of Business and Economic Statistics*, 13, 207–15.

Jalan, J. and Ravallion, M. (2003) Estimating the benefit incidence of an antipoverty program by propensity-score matching. *Journal of Business and Economic Statistics*, 21, 19–30.

Larsson, L. (2003) Evaluation of Swedish youth labour market programmes. *Journal of Human Resources*, 38, 891–927.

Lechner, M. (1999) Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business and Economic Statistics*, 17, 74–90.

——— (2001a) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner, and F. Pfeiffer (eds), *Econometric Evaluation of Labour Market Policies* (pp. 43–58). Heidelberg: Physica/Springer.

——— (2001b) A note on the common support problem in applied evaluation studies. University of St. Gallen Economics Discussion Paper Series, 2001–01.

——— (2002a) Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies. *Review of Economics and Statistics*, 84, 205–20.

——— (2002b) Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society, Series A*, 165, 59–82.

Lechner, M. and Miquel, R. (2002) Identification of effects of dynamic treatments by sequential conditional independence assumptions. University of St. Gallen Economics Discussion Paper Series, 2001–07.

Manski, C. (1989) Anatomy of the selection problem. *Journal of Human Resources*, 24, 343–60.

——— (1990) Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80, 319–23.
——— (1993) The selection problem in econometrics and statistics. In G. Maddala, C. RaoH. Vinod (eds), *Handbook of Statistics* Elsevier Science.
——— (1997) Monotone treatment response. *Econometrica*, 65, 1311–34.
Manski, C. and Pepper, J. (2000) Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*, 68, 997–1010.
Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. *Statistical Science, reprint*, 5, 463–80.
Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
Puhani, P. (1999) *Evaluating Active Labour Market Policies: Empirical Evidence for Poland during Transition*. Heidelberg: Physica.
Rosenbaum, P. and Rubin, D. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
Rubin, D. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.
——— (1977) Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2, 1–26.
——— (1980) Comment on 'Randomization analysis of experimental data: the Fisher randomization test' by D. Basu. *Journal of the American Statistical Association*, 75, 591–93.
Seifert, B. and Gasser, T. (1996) Finite-sample variance of local polynomials: analysis and solutions. *Journal of American Statistical Association*, 91, 267–75.
Seifert, B. and Gasser, T. (2000) Data adaptive ridging in local polynomial regression. *Journal of Computational and Graphical Statistics*, 9, 338–60.
Smith, J. and Todd, P. (2004) Does matching address LaLonde's critique of nonexperimental estimators? Forthcoming in *Journal of Econometrics*.
Stone, C. (1980) Optimal rates of convergence of nonparametric estimators. *Annals of Statistics*, 8, 1348–60.
Thistlethwaite, D. and Campbell, D. (1960) Regression-discontinuity analysis: an alternative to the *ex post facto* experiment. *Journal of Educational Psychology*, 51, 309–17.
Todd, P. (1999) Matching and local linear approaches to program evaluation using a semiparametric propensity score, mimeo. University of Pennsylvania.
Trochim, W. (1984) *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Beverly Hills: Sage.
van der Klaauw, W. (2002) Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. *International Economic Review*, 43, 1249–87.