Allgemeines Statistisches Archiv 90, 199–215

The microeconometric estimation of treatment effects - An overview

By Marco Caliendo and Reinhard Hujer*

Summary: The need to evaluate the performance of active labour market policies is not questioned any longer. Even though OECD countries spend significant shares of national resources on these measures, unemployment rates remain high or even increase. We focus on microeconometric evaluation which has to solve the fundamental evaluation problem and overcome the possible occurrence of selection bias. When using non-experimental data, different evaluation approaches can be thought of. The aim of this paper is to review the most relevant estimators, discuss their identifying assumptions and their (dis-)advantages. Thereby we will present estimators based on some form of exogeneity (selection on observables) as well as estimators where selection might also occur on unobservable characteristics. Since the possible occurrence of effect heterogeneity has become a major topic in evaluation research in recent years, we will also assess the ability of each estimator to deal with it. Additionally, we will also discuss some recent extensions of the static evaluation framework to allow for dynamic treatment evaluation.

Keywords: Evaluation, effect heterogeneity, matching, dynamic treatments. JEL C40, H43, J68.

1. Introduction

The need to evaluate the performance of active labour market policies (ALMP) is not questioned any longer. Even though OECD countries spend significant shares of national resources on these measures, unemployment rates remain high or even increase. The ideal evaluation process can be looked at as a series of three steps (Fay, 1996): First, the impacts of the programme on the individual should be estimated (MICROECONOMETRIC EVALUATION). Second, it should be examined if the impacts are large enough to yield net social gains (MACROECONOMIC EVALUATION). Third, it should be answered if this is the best outcome that could have been achieved for the money spent (COST-BENEFIT ANALYSIS). In this paper we focus on the first step. The main question in microeconometric evaluation is if the outcome for an individual is affected by the participation in an ALMP programme or not. We would like to know the difference between the value of the participant's outcome in the actual situation and the value of the outcome if he had not participated in the programme. The fundamental evaluation problem arises because we can never observe both states (participation and non-participation) for the same individual at the same time, i. e. one of the states is counterfactual. Therefore finding an adequate control group and solving the problem of selection bias is necessary to make a comparison possible.

Received: 14.03.2005 / Revised: 26.07.2005

* The authors thank Stephan L. Thomsen, Christopher Zeiss and one anonymous referee for valuable comments. The usual disclaimer applies.

Depending on the data at hand, different evaluation strategies can be thought of. Since in most European countries - unlike in the US - experimental data are not available, researchers have to use non-experimental data. A lot of methodological progress has been made to develop and justify non-experimental evaluation estimators which are based on econometric and statistical methods to solve the fundamental evaluation problem (see e. g. Heckman *et al.*, 1999). The aim of this paper is to give an overview of the most relevant evaluation approaches and provide some guidance on how to choose between them. Thereby we will also discuss the possible occurrence of effect heterogeneity, which has become a major focus of evaluation research in the last years, and the ability of each estimator to deal with it.

Two broad categories of estimators can be distinguished according to the way selection bias is handled. The first category contains approaches that rely on the so-called unconfoundedness or selection on observables assumption. If one believes that the available data is not rich enough to justify this assumption, one has to rely on the second category of estimators which explicitly allows selection on unobservables, too. We will discuss different approaches for both situations in Section 3 where we also present some recent extensions of the static evaluation framework to dynamic concepts. Before we do so, we are going to introduce the evaluation framework in Section 2, where we especially present the potential outcome approach, discuss parameters of interest, selection bias on observable and on unobservable characteristics as well as heterogeneous treatment effects. Finally, Section 4 concludes.

2. The Evaluation Framework

2.1. Potential Outcome Approach and the Fundamental Evaluation Problem. Inference about the impact of a treatment on the outcome of an individual involves speculation about how this individual would have performed in the labour market, if he had not received the treatment. The framework serving as a guideline for the empirical analysis of this problem is the potential outcome approach, also known as the Roy (1951) – Rubin (1974) – model.

The main pillars of this model are individuals, treatment (participating in a programme or not) and potential outcomes, that are also called responses.[1] In the basic model there are two potential outcomes (Y^1, Y^0) for each individual, where Y^1 indicates a situation with treatment and Y^0 without. To complete the notation, we additionally denote variables that are unaffected by treatment by X. Attributes X are exogenous in the sense that their potential values for different treatment states coincide (Holland, 1986). Furthermore we define a binary assignment indicator D, indicating whether an individual actually received treatment ($D = 1$), or not ($D = 0$).

[1] It should be clear, that this framework is not restricted to the evaluation of labour market programmes. It applies for every situation where one group of units, e. g. individuals or firms or other entities, receive some form of treatment and others do not.

The treatment effect for each individual i is then defined as the difference between his potential outcomes:

$$\Delta_i = Y_i^1 - Y_i^0. \tag{1}$$

The fundamental problem of evaluating this individual treatment effect arises because the observed outcome for each individual is given by:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0. \tag{2}$$

This means that for those individuals who participated in treatment we observe Y^1 and for those who did not participate we observe Y^0. Unfortunately, we can never observe Y^1 and Y^0 for the same individual simultaneously and therefore we cannot estimate (1) directly. The unobservable component in (1) is called the counterfactual outcome.

Concentration on a single individual requires that the effect of the intervention on each individual is not affected by the participation decision of any other individual, i.e. the treatment effect Δ_i for each person is independent of the treatment of other individuals. In statistical literature this is referred to as the stable unit treatment value assumption (SUTVA)[2] and guarantees that average treatment effects can be estimated independently of the size and composition of the treatment population. In particular, it excludes peer-effects as well as cross-effects and general equilibrium effects (Sianesi, 2004).

2.2. Treatment Effects and Selection Bias. Since there will never be an opportunity to estimate individual effects in (1) with confidence, we have to concentrate on population averages of gains from treatment. Two treatment effects are dominantly used in empirical studies. The first one is the (population) average treatment effect (ATE)

$$\Delta_{ATE} = E(\Delta) = E(Y^1) - E(Y^0), \tag{3}$$

which answers the question which would be the outcome if individuals in the population were randomly assigned to treatment. The most frequently used parameter is the so called average treatment effect on the treated (ATT) and focusses explicitly on the effects on those for whom the programme is actually intended. It is given by

$$\Delta_{ATT} = E(\Delta \mid D = 1) = E(Y^1 \mid D = 1) - E(Y^0 \mid D = 1). \tag{4}$$

In the sense that this parameter focuses directly on participants, it determines the realised gross gain from the programme and can be compared with its costs, helping to decide whether the programme is successful or not (Heckman *et al.*, 1999). Given Equation (4), the problem of selection bias can be straightforwardly seen since the second term on the right hand side

[2] See Holland (1986) for a further discussion of this concept.

is unobservable as it describes the hypothetical outcome without treatment for those individuals who received treatment. Since with non-experimental data the condition $E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0)$ is usually not satisfied, estimating ATT by the difference in sub-population means of participants $E(Y^1 \mid D = 1)$ and non-participants $E(Y^0 \mid D = 0)$ will lead to a selection bias. This bias arises because participants and non-participants are selected groups that would have different outcomes, even in absence of the programme. It might be caused by observable or unobservable factors.

2.3. POTENTIAL OUTCOME FRAMEWORK AND HETEROGENEOUS TREATMENT EFFECTS. For the further discussion it will be helpful to relate the potential outcome framework to familiar econometric notation. To do so, we follow Blundell and Costa Dias (2002) and define the following outcome equations

$$Y_{it}^1 = g_t^1(X_i) + U_{it}^1 \quad \text{and} \quad Y_{it}^0 = g_t^0(X_i) + U_{it}^0, \tag{5}$$

where the subscripts i and t index the individual and the time period, respectively. The functions g^0 and g^1 represent the relationship between potential outcomes and the set of observable characteristics. U^0 and U^1 are error terms which have zero mean and are assumed to be uncorrelated with regressors X. For the familiar case of linear regression, the g functions specialise to $g^1(X) = X\beta_1$, and $g^0(X) = X\beta_0$.

Heckman and Robb (1985) note that the decision to participate in treatment may be determined by a prospective participant, by a programme administrator, or both. Whatever the specific content of the rule, it can be described in terms of an index function framework. Let IN_i be an index of benefits to the relevant decision maker from participating in the programme. It is a function of observed (Z_i) and unobserved (V_i) variables. Therefore

$$IN_i = f(Z_i) + V_i, \tag{6}$$

with enrolment in the programme D_i given by

$$D_i = \begin{cases} 1 & \text{if} \quad IN_i > 0, \\ 0 & \text{otherwise}. \end{cases}$$

Under this specification and the further assumption that treatment takes place in period k, one can define the individual-specific treatment effect for any X_i as

$$\Delta_{it}(X_i) = Y_{it}^1 - Y_{it}^0 = [g_t^1(X_i) - g_t^0(X_i)] + [U_{it}^1 - U_{it}^0] \quad \text{with } t > k. \tag{7}$$

The ATT measured in the post-treatment period $t > k$ is then defined as

$$\Delta_{ATT} = E(\Delta_{it} \mid D_i = 1). \tag{8}$$

The assignment process to treatment is most probably not random. Consequently, the assignment process will lead to non-zero correlation between enrolment (D_i) and the outcome's error term (U^1, U^0). This may occur because of stochastic dependence between (U^1, U^0) and V_i in (6) or because of stochastic dependence between (U^1, U^0) and Z_i. In the former case we have selection on unobservables, whereas in the latter case selection on observables is prevalent (Heckman and Robb, 1985).

We can use this discussion to highlight the problem of heterogeneous treatments, i. e. situations where the impact of a programme differs across individuals, in a common and intuitive way.[3] If treatment impacts vary across individuals this may come systematically through the observables' component or be part of the unobservables and we can re-write equation (5) as

$$Y_{it} = g_t^0(X_i) + \Delta_t(X_i)D_{it} + [U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)], \tag{9}$$

where

$$\Delta_t(X_i) = E[\Delta_{it}(X_i)] = g_t^1(X_i) - g_t^0(X_i) \tag{10}$$

is the expected treatment effect at time t for individuals characterised by X_i (Blundell and Costa Dias, 2002). Since abandoning the assumption of homogeneous treatment effects and identifying the individuals that benefit from programmes provides some scope to improve their future efficiency, we will assess for each estimation method that we will present in the following its capability to deal with heterogeneous treatment effects.

3. Non-Experimental Evaluation Methods

The discussion in Subsections 2.2 and 2.3 has made clear that the problem of selection bias is a severe one and cannot be solved with more data, since the fundamental evaluation problem will not disappear. We have a distorted representation of a true population in a sample as a consequence of a sampling rule, which is the essence of the selection problem (Heckman, 2001). Hence, we have to use some identifying assumptions to draw inference about the hypothetical population based on the observed population. In the following subsections we will present several evaluation approaches. Each approach invokes different identifying assumptions to construct the required counterfactual outcome. We will start the discussion with two estimators (matching and regression) that are based on the selection on observables assumption.[4] Following that we introduce three estimators that allow for selection on unobservables, too, namely difference-in-differences, instrumental variables and selection models. Finally, we also briefly discuss regression discontinuity models and the estimation of treatment effects in a dynamic framework.

[3] See e. g. the discussion in Smith (2000).

[4] See Imbens (2004) for an extensive overview of estimating average treatment effects under unconfoundedness.

3.1. MATCHING ESTIMATOR. Matching is based on the identifying assumption that conditional on some covariates X, the outcome Y is independent of D.[5] In the notation of Dawid (1979) this is

ASSUMPTION 1 *Unconfoundedness:* $Y^0, Y^1 \amalg D \mid X$,

where $\amalg$ denotes independence. If Assumption 1 is true, then $F(Y^0 \mid X, D = 1) = F(Y^0 \mid X, D = 0)$ and $F(Y^1 \mid X, D = 1) = F(Y^1 \mid X, D = 0)$. This means, that conditionally on X, non-participant outcomes have the same distribution that participants would have experienced if they had not participated in the programme and vice versa (Heckman *et al.*, 1997). Similar to randomisation in a classical experiment, matching balances the distributions of all relevant, pre-treatment characteristics X in the treatment and comparison group.[6] Thus it achieves independence between the potential outcomes and the assignment to treatment.

ASSUMPTION 2 *Overlap:* $0 < P(D = 1 \mid X) < 1$, *for all* X.

This implies that the support of X is equal in both groups, i.e. $S = \text{Support}(X|D = 1) = \text{Support}(X|D = 0)$. Assumption 2 prevents X from being a perfect predictor in the sense that we can find for each participant a counterpart in the non-treated population and vice versa. If there are regions where the support of X does not overlap for the treated and non-treated individuals, matching has to be performed over the common support region only. The estimated effects have then to be redefined as the mean treatment effect for those individuals falling within the common support (Blundell *et al.*, 2004). Rosenbaum and Rubin (1983) call Assumptions 1 and 2 together 'strong ignorability' under which ATT and ATE can be defined for all values of X. If one is interested in ATT only, it is sufficient to assume $Y^0 \amalg D \mid X$ and the weaker overlap Assumption $P(D = 1 \mid X) < 1$. The mean impact of treatment on the treated can be written as

$$\Delta_{ATT}^{MAT} = E(Y^1|X, D = 1) - E_X[E(Y^0|X, D = 0)|D = 1], \qquad (11)$$

where the first term can be estimated from the treatment group and the second term from the mean outcomes of the matched comparison group. The outer expectation is taken over the distribution of X in the treated population. The method of matching can also be used to estimate ATT at some points $X = x$, where x is a particular realisation of X. Two things have to be mentioned: First, it should be clear that conditioning on all

[5] These are the covariates which also appear in Z as defined in Equation (6).

[6] If we say relevant we mean all those covariates that influence the assignment to treatment as well as the potential outcomes.

relevant covariates is limited in case of a high dimensional vector X. For that case Rosenbaum and Rubin (1983) suggest the use of so-called balancing scores to overcome this dimensionality problem.[7] Second, there are several different matching algorithms suggested in the literature, e. g. kernel or nearest-neighbour matching, and the choice between them is not trivial since it involves a trade-off between bias and variance (see Smith and Todd, 2005, for an overview).

3.2. LINEAR REGRESSION APPROACH. Even though regression and matching both rely on the unconfoundedness assumption, there are some key differences between both approaches which are worth to be discussed. One key difference is that matching, due to its non-parametric nature, avoids functional form assumptions implicit in linear regression models. The potential outcomes in a linear regression framework can be written as $Y^1 = X\beta_1 + U^1$ and $Y^0 = X\beta_0 + U^0$ and ATT under regression is given by[8]:

$$\Delta_{ATT}^{Reg} = E(Y^1 - Y^0|X, D = 1) = X(\beta_1 - \beta_0) + E(U^1 - U^0|X, D = 1). \quad (12)$$

The identifying assumption needed to justify regression under unconfoundedness is analogue to Assumption 1 and can be re-written as:

ASSUMPTION 3 *Unconfoundedness in Regression:* $U^0, U^1 \amalg D \mid X$.

In the matching framework, the goal is to set the bias $B(X) = 0$ which basically only requires that the mean of the error terms in the treatment group given a covariate cell X equals the corresponding mean in the control group, that is $B(X) = E(U^1|X, D = 1) - E(U^0|X, D = 0) = 0$. This means that it is possible to match on variables that are correlated with the error term in the outcome equation (Hui and Smith, 2002). In the regression framework, however, we need to eliminate the dependence between (U^0, U^1) and X, that is $E(U^1|X, D = 1) = E(U^0|X, D = 0) = 0$ (Heckman *et al.*, 1998). Of course, as Smith (2000) notes, the difference between both approaches fades with the inclusion of a sufficient number of higher-order and interaction terms in the regression. However, not only is such an inclusion not very common in practice, it is also not straightforward to choose these terms. Moreover, whereas matching estimators do rely on the common support assumption, regression estimators do not and will produce estimates even in the absence of similar comparison units, since the linear functional form assumption fills in for the missing data (Smith, 2004). Another key difference

[7] One possible balancing score is the propensity score. See Rosenbaum (2002) or Caliendo and Kopeinig (2005) for an introduction into propensity score matching estimators and some guidance for their implementation.

[8] For notational convenience we drop individual subscript i and time subscript t.

between regression and matching is the way both approaches handle heterogeneous treatment effects. As Lechner (2002) notes, the non-parametric matching approach leaves the individual causal effect unrestricted and allows individual effect heterogeneity in the population. This is not true for the regression approach which will not recover ATT, although, at times it might provide a close approximation as shown by Angrist (1998) and Blundell *et al.* (2004).

3.3. INSTRUMENTAL VARIABLES ESTIMATOR. Let us now turn to estimators that account for selection on unobservables, too. We will start with the method of instrumental variables (IV). Its underlying identification strategy is to find a variable which determines treatment participation but does not influence the outcome equation. The instrumental variable affects the observed outcome only indirectly through the participation decision and hence causal effects can be identified through a variation in this instrumental variable. IV methods are extensively discussed in Imbens and Angrist (1994) and Angrist *et al.* (1996) among others. In terms of the discussion in Subsection 2.3, IV requires the existence of at least one regressor to the decision rule, Z^*, that satisfies the following three conditions (Blundell and Costa Dias, 2000):

- Z^* determines programme participation. For that to be true, it has to have a non-zero coefficient in the decision rule in Equation (6).
- We can find a transformation, s, such that $s(Z^*)$ is uncorrelated with the error terms (U^1, V) and (U^0, V), given the exogenous variables X.
- Z^* is not completely determined by X.

The variable Z^* is then called the instrument. In providing variation that is correlated with the participation decision but does not affect potential outcomes from treatment directly, it can be used as a source of exogenous variation to approximate randomised trials (Blundell and Costa Dias, 2000).

Clearly, a major problem with this estimator is to find a good instrument. In the treatment evaluation problem it is hard to think of variables that satisfy all three above mentioned assumptions. The difficulty lies mainly in the simultaneous requirement that the variable has to predict participation but does not influence the outcome equation. As pointed out by Blundell and Costa Dias (2000), a second drawback arises when considering the heterogeneous treatment framework. Recall that the error term from Equation (9) in Subsection 2.3 is given by $[U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)]$. Even if Z^* is uncorrelated with U_{it}, the same cannot be true by definition for $U_{it}^0 + D_{it}(U_{it}^1 - U_{it}^0)$ since Z^* determines D_i by assumption. The violation of this assumption invalidates the application of IV methodology in a heterogeneous framework (Blundell and Costa Dias, 2000). However, in this situation it might still be possible to provide a potentially interesting parameter of the IV estimation - called local average treatment effect (LATE) by Imbens and Angrist (1994). This estimator identifies the treatment effect for those individuals

(with characteristics X) who are induced to change behaviour because of a change in the instrument.[9] It should be clear that each instrument implies its own LATE, and LATEs for two different instruments may differ substantially depending on the impacts realised by the persons each instrument induces to participate (Hui and Smith, 2002).

3.4. SELECTION MODEL. This method is also known as the Heckman selection estimator (Heckman, 1978). It is more robust than the IV method but also more demanding in the sense that it imposes more assumptions about the structure of the model. Two main assumptions are required (Blundell and Costa Dias, 2000):

- There has to be one additional regressor in the decision rule which has a non-zero coefficient and which is independent of the error term V.
- Additionally, the joint density of the distribution of the errors U_{it} and V_i has to be known or can be estimated.

The basic idea of this estimator is to control directly for the part of the error term in the outcome equation that is correlated with the participation dummy variable. It can be seen as a two-step-procedure. First, the part of the error term U_{it} that is correlated with D_i is estimated. Second, this term is then included in the outcome equation and the effect of the programme is estimated. By construction, the remains of the error term in the outcome equation are not correlated with the participation decision any more (Blundell and Costa Dias, 2000).[10]

The Heckman selection estimator is not without critique, which rests mainly on the following point (see e. g. Puhani, 2000): If there are no exclusion restrictions, the models are identified only by assumptions about functional form and error distributions. This may lead to large standard errors and results that are very sensitive to the particular distributional assumptions invoked. This point of criticism is very closely related to the problem of finding a good instrument as described for the IV method. In fact, in a recent paper Vytlacil (2002) shows that the identifying assumptions for the selection model are equivalent to those invoked by Imbens and Angrist (1994) in the linear instrumental variables context.

3.5. DIFFERENCE-IN-DIFFERENCES ESTIMATOR. The difference-in-differences (DID) estimator requires access to longitudinal data and forms simple averages over the group of participants and non-participants between pre-treatment period t' and post-treatment period t, that is, changes in the

[9] Additionally to those assumptions already made, we further have to assume that the instrument has the same directional effect on all those whose behaviour it changes. This assumption rules out the co-existence of defiers and compliers and is known as 'monotonicity assumption' (Imbens and Angrist, 1994).

[10] Blundell and Costa Dias (2000) also show that this approach is capable of identifying ATT if effects are assumed to be heterogeneous.

outcome variable Y for treated individuals are contrasted with the corresponding changes for non-treated individuals (Heckman *et al.*, 1998):

$$\Delta^{DID} = [Y_t^1 - Y_{t'}^0 \mid D = 1] - [Y_t^0 - Y_{t'}^0 \mid D = 0]. \tag{13}$$

The identifying assumption of this method is

$$E(Y_t^0 - Y_{t'}^0 \mid D = 1) = E(Y_t^0 - Y_{t'}^0 \mid D = 0). \tag{14}$$

The DID estimator is based on the assumption of time-invariant linear selection effects, so that differencing the differences between participants and non-participants eliminates the bias (Heckman *et al.*, 1998). To make this point clear, we can re-write the outcome for an individual i at time t as $Y_{it} = \pi_{it} + D_{it} \cdot Y_{it}^1 + (1 - D_{it}) \cdot Y_{it}^0$, where π_{it} captures the effects of selection on unobservables. The validity of the DID estimator then relies on the assumption $\pi_{it} = \pi_{it'}$, where it is not required that the bias vanishes completely, but that it remains constant (Heckman *et al.*, 1998). One problem when using DID is Ashenfelter's dip, i. e. a situation where shortly before participation in an ALMP programme the employment situation of future participants deteriorates (Ashenfelter, 1978). If the 'dip' is transitory and the dip is eventually restored even in the absence of participation in the programme, the bias will not average out. To allow a more detailed discussion, Blundell and Costa Dias (2002) further decompose π_{it} in three parts: an individual-specific fixed effect, a common macroeconomic effect and a temporary individual-specific effect. Clearly, for the DID to be unbiased it is sufficient that selection into treatment is independent from the temporary individual-specific effect, since the other two effects vanish in the sequential differences. They also discuss the case where the macroeconomic effect has a differential impact across the group of participants and non-participants. This may happen when both groups differ on unobserved characteristics which make them react differently to macroeconomic shocks. To overcome this problem they propose a differential trend adjusted DID estimator (Blundell and Costa Dias, 2002). Heckman *et al.* (1998) combine the DID approach with the already presented matching estimator by comparing the before-after outcome of participants with those of matched non-participants. Smith and Todd (2005) show that this 'conditional DID estimator' is more robust than traditional cross-section matching estimators, as it allows for selection on observables as well as time-invariant selection on unobservables.

3.6. REGRESSION DISCONTINUITY MODEL. The regression discontinuity model (RDM) can be seen as a particular type of instrumental variable identification strategy. It uses discontinuities in the selection process to identify causal effects. In this model, treatment depends on some observed variable, Z, according to a known, deterministic rule, such as $D = 1$ if $Z > \overline{Z}$ and $D = 0$ otherwise (Heckman *et al.*, 1999). The variable Z has direct impact on Y as well as an indirect impact on Y through D. This indirect

impact is the causal effect we would like to identify. Frölich (2002) notes that this effect is identified if the direct and indirect impacts of Z on Y can be separated.

There are several things to note about RDM (see e. g. Heckman *et al.*, 1999). First, it is assumed that selection is on observable characteristics only. Second, it should be clear that there is no common support for participants and non-participants making matching impossible. Hence, RDM takes over when there is selection on observables (here: the deterministic rule) but the overlapping support condition required for matching breaks down (with a certain Z you either belong to the participant or the non-participant group). Finally, the selection rule is assumed to be deterministic and known and that variation in the relevant variable Z is exogenous.

3.7. Dynamic Evaluation Concepts.

3.7.1. Sequential Matching Estimators. What we have discussed so far is basically a static evaluation framework where an individual can participate in one programme (or not). A recent extension of this framework for matching estimators considers the case, where individuals can participate in subsequent treatments. Lechner and Miquel (2002) discuss identifying assumptions for so-called sequential matching estimators. These estimators mimic the matching estimators described above but allow to estimate effects in a dynamic causal model. Their framework can be made clear in a three-periods-two-treatments model. We follow the discussion in Lechner (2004) and present the needed additional notation in the following. First, we introduce a time index $t \in \{0,1,2\}$ and extend the treatment indicator D by this time index, that is $D = (D_0, D_1, D_2)$. It is further assumed that in period 0 everybody is in the same treatment state $D_0 = 0$, whereas from the second period on D_t can take two values. Realisations of D_t are denoted by $d_t \in \{0,1\}$. So in period 1 an individual is observed in exactly one of these two treatments $(0,1)$, whereas in period 2 an individual participates in one of four possible treatment sequences $\{(0,0),(1,0),(0,1),(1,1)\}$. Additionally, the history of variables up to period t are denoted by a bar below a variable, e. g. $\underline{d}_2 = (d_1, d_2)$. The potential outcomes are indexed by treatments and the time period, i. e. $Y^{\underline{s}_t} = (Y_0^{\underline{d}_t}, Y_1^{\underline{d}_t}, Y_2^{\underline{d}_t})$. The observed outcomes are given by the following equation

$$Y_t = D_1 Y_t^1 + (1 - D_1) Y_t^0 = D_1 D_2 Y_t^{1,1} + D_1 (1 - D_2) Y_t^{1,0} + (1 - D_1) D_2 Y_t^{0,1} + (1 - D_1)(1 - D_2) Y_t^{0,0} . \quad (15)$$

As in the static model, variables that influence treatment selection and potential outcomes are called attributes and are denoted by X. An important distinction has to be made regarding the exogeneity of these variables. Whereas in the static model exogeneity is assumed, in the dynamic model the X-variables in later periods can be influenced by treatment realisations. Hence, there are potential values of these variables as well: $X^{d_t} =$

$(X_0^{d_t}, X_1^{d_t}, X_2^{d_t})$, where e. g. $X_1^{d_1}$ may contain $Y_1^{d_1}$ or functions of it. The sequential matching framework is a powerful tool and is applicable for situations where individuals can participate more than once in a programme and where it is possible to identify treatment sequences.

3.7.2. DURATION MODELS. Another methodology for modelling dynamically assigned treatments is the application of duration models (Abbring and van den Berg, 2003). In these models not only the information if an individual participates in a programme is considered, but also the timing of the treatment within the unemployment spell. To introduce the notation we normalise the point in time when an individual enters unemployment to zero, denote the duration until the individual enters regular employment with T_e and the duration until the individual enters a programme with T_p (realisations are denoted by t_u and t_p, respectively). Both durations are assumed to vary with observable characteristics x and unobservable characteristics v_e and v_p. Abbring and van den Berg (2003) assume that the realisation t_p affects the distribution of T_e in a deterministic way from t_p onwards. For the specification of the hazard rates a mixed proportional hazard model is used. Basic feature of this model is that the duration dependence, observable covariates and unobservable components enter the hazard rate multiplicatively:

$$\theta_e(t|t_p, x, v_e) = \lambda_e(t) \exp[x'\beta_e + \mu(t - t_p)I(t > t_p) + v_e]\,. \tag{16}$$

The hazard rate for the transition into regular employment θ_e consists of the baseline hazard $\lambda_e(t)$ that determines the duration dependence, the systematic part $\exp(x'\beta_e)$ and the unobserved heterogeneity term $\exp(v_e)$. The treatment effect $\exp[\mu(t - t_p)I(t > t_p)]$ with $I(t > t_p)$ as an indicator function taking the value 1 if $t > t_p$ is specified as a function of the difference $t - t_p$. In general, the treatment effect is allowed to vary over time after the treatment has started and can be interpreted as a shift of the hazard rate by $\exp(\mu(t - t_p))$. The transition rate from unemployment into programmes θ_p is analogously specified as a mixed proportional hazard model:

$$\theta_p(t|x, v_p) = \lambda_p(t) \exp[x'\beta_p + v_p]\,. \tag{17}$$

Identifying the treatment effect requires to consider selectivity which is present if individuals with a relatively high transition rate into employment also have a relatively high transition into programme participation (Abbring and van den Berg, 2003). In this case we obviously would observe a positive correlation between v_e and v_p and the joint distribution $G(v_e, v_p)$ has to be specified. Abbring and van den Berg (2003) show that the bivariate model (16) and (17) and especially the treatment effect is nonparametrically identified, since no parametric assumptions with respect to the baseline hazard and the unobserved heterogeneity distribution are required.

Furthermore the identification does not require exclusion restrictions on x which are often hardly to justify from a theoretical point of view.[11]

3.7.3. MATCHING WITH TIME-VARYING TREATMENT INDICATORS. An alternative concept of modelling dynamic treatment effects is presented by Fredriksson and Johansson (2004) and Sianesi (2004). They introduce a non-parametric matching estimator that takes the timing of events into account but does not rely on proportionality assumptions. An important topic in this framework is the choice of an appropriate control group. Instead of defining control individuals as those who never participate, Sianesi (2004) defines control individuals as those who did not participate until a certain time period. Fredriksson and Johansson (2004) formalise her approach and argue that the standard way of defining a control group, i.e. those individuals who never participated in a given time interval, might lead to biased results, because the unconfoundedness assumption might be violated as the treatment indicator itself is defined conditional on future outcomes. Following Sianesi (2004), the key choice faced by the unemployed in this framework is not whether to participate at all, but whether to participate in a programme or not now. In the latter case, the individual searches longer in open unemployment. The corresponding parameter of interest in this setting is then defined as the effect of joining a programme now in contrast to waiting longer. The population of interest at time u are those still openly unemployed after u months. Treatment receipt in u is denoted by $D^{(u)} = 1$. The comparison group consists of all persons who do not join at least up to u, denoted by $D^{(u)} = 0$. The outcome of interest is defined over time t and is given by $Y_t^{(u)}$. The potential outcome if an individual joins in u is denoted by $Y_t^{1(u)}$ and if he does not join at least up to u by $Y_t^{0(u)}$. For each point of elapsed unemployment duration the parameter of interest is

$$\Delta_u^t = E(Y_t^{1(u)} - Y_t^{0(u)} | D^{(u)} = 1) = E(Y_t^{1(u)} | D^{(u)} = 1) \qquad (18)$$
$$- E(Y_t^{0(u)} | D^{(u)} = 1), \quad \text{for} \quad t = u, u+1, \ldots, T.$$

This is the average impact at time t, for those joining a programme in their u^{th} month of unemployment compared to waiting longer in open unemployment. Sianesi (2004) notes that the treatment effects are based on a comparison of individuals who have reached the same elapsed duration of unemployment. Measurement starts at time u, the start of the programme and therefore possible locking-in effects might encounter. The second term on the right hand side of (18) is not identified and the CIA needed in that case is given by

$$Y_t^{0(u)} \amalg D^{(u)} | X = x \quad \text{for} \quad t = u, u+1, \ldots, T, \qquad (19)$$

which means that given a set of observed characteristics X, the counterfactual distribution of $Y_t^{0(u)}$ for individuals joining in u is the same as for

[11] It should be noted that anticipatory programme effects are ruled out in the above mentioned specification (Abbring and van den Berg, 2003).

those not joining in u and waiting longer. The estimated treatment effect is then the effect for those who participate in a programme at some time in their unemployment spell instead of waiting longer. Even though this is not a standard evaluation parameter of interest, it still shows whether a programme was effective or not.

4. Summary - Which Estimator to Choose?

We have presented several different evaluation strategies in this paper. The final question to be answered is: Which strategy to choose when evaluating labour market programmes? Unfortunately, there is no 'one' answer to this question because there is no 'magic bullet' that will solve the evaluation problem in any case. As described above, different strategies invoke different identifying assumptions and also require different kinds of data for their implementation. When those assumptions hold, a given estimator will provide consistent estimates of certain parameters of interest (Smith, 2004). The literature provides a lot of guidance for making the right choice, based either on experimental datasets to benchmark the performance of alternative evaluation estimators or Monte-Carlo simulations.

The different estimators can be classified with respect to two dimensions. The first dimension is the required data for their implementation. Except the DID estimator, the presented methods for the static evaluation framework require only cross-sectional information for the group of participants and non-participants. However, longitudinal information might help to justify the unconfoundedness assumption, enables the researcher to combine e. g. matching with DID estimators and allows an extension to dynamic concepts of treatment evaluation. The second dimension concerns the handling of selection bias. We have presented three estimators that are based on the unconfoundedness assumption. Clearly, the most crucial point for these estimators is that the identifying assumption is in general a very strong one and they are only as good as the used control variables X (Blundell *et al.*, 2004). If the assumption holds, both, matching and regression, can be used. Since regression analysis ignores the common support problem, imposes a functional form for the outcome equation, and is not as capable as matching of handling effect heterogeneity, matching might be preferred. If there is no common support at all, regression discontinuity models can be applied. For the situation where there is selection on unobservables, too, we have presented three strategies. Whereas selection models try to model the selection process completely, IV methods focus on searching a source of independent variation affecting the participation decision (but not the outcome) and DID methods erase a time-invariant selection effect by differencing outcomes of participants and non-participants before and after treatment took place. The crucial assumption for the latter approach is that the selection bias is time invariant. Finding a suitable and credible instrument and heterogeneous treatment effects are possible drawbacks for the IV method. The

latter point is not a problem for selection models, even though this flexibility comes at a price, because a full specification of the assignment rule and stronger assumptions are required. Hence, if the common effect assumption is plausible in a given context, the IV estimator might be preferred (Smith, 2004). Finally, we have also presented some recent extensions of the static evaluation framework to analyse dynamic treatment effects, e. g. to allow for subsequent treatments and to take the timing of events into account.

References

Abbring, J. H., van den Berg, G. J. (2003). The non-parametric identification of treatment effects in duration models. *Econometrica* **71** 1491–1517.

Angrist, J. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica* **66** 249–288.

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91** 444–472.

Ashenfelter, O. (1978). Estimating the effects of training programs on earnings. *Review of Economics and Statistics* **60** 47–57.

Blundell, R., Costa Dias, M. (2000). Evaluation methods for non-experimental data. *Fiscal Studies* **21** 427–468.

Blundell, R., Costa Dias, M. (2002). Alternative approaches to evaluation in empirical microeconomics. *Portuguese Economic Journal* **1** 91–115.

Blundell, R., Dearden, L., Sianesi, B. (2004). Evaluating the impact of education on earnings in the UK: Models, methods and results from the NCDS. Working Paper No. 03/20, The Institute of Fiscal Studies, London.

Caliendo, M., Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. Discussion Paper No. 1588, IZA, Bonn.

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

Fay, R. (1996). Enhancing the effectiveness of active labor market policies: Evidence from programme evaluations in OECD countries. *Labour Market and Social Policy Occasional Papers*, OECD, Paris.

Frederiksson, P., Johansson, P. (2004). Dynamic treatment assignment - The consequences for evaluations using observational data. Discussion Paper No. 1062, IZA, Bonn.

Frölich, M. (2002). *Programme Evaluation and Treatment Choice*. Lecture Notes in Economics and Mathematical Systems, Springer, Berlin.

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica* **46** 931–959.

Heckman, J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy* **109** 673–748.

Heckman, J., Ichimura, H., Smith, J., Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica* **66** 1017–1098.

HECKMAN, J., ICHIMURA, H., TODD, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* **64** 605–654.

HECKMAN, J., LALONDE, R., SMITH, J. (1999). The economics and econometrics of active labor market programs. In *Handbook of Labor Economics Vol. III* (O. Ashenfelter, D. Card, eds.), 1865–2097. Elsevier, Amsterdam.

HECKMAN, J., ROBB, R. (1985). Alternative methods for evaluating the impact of interventions - An overview. *Journal of Econometrics* **30** 239–267.

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

HUI, S., SMITH, J. (2002). The labor market impacts of adult education and training in Canada. Report prepared for the Human Resources Development Canada (HRDC), Quebec.

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

IMBENS, G., ANGRIST, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

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

LECHNER, M. (2004). Sequential matching estimation of dynamic causal models. Discussion Paper No. 1042, IZA, Bonn.

LECHNER, M., MIQUEL, R. (2002). Identification of effects of dynamic treatments by sequential conditional independence assumptions. Working Paper, SIAW, University St. Gallen.

PUHANI, P. A. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* **14** 53–68.

ROSENBAUM, P. R. (2002). *Observational Studies*. Springer, New York.

ROSENBAUM, P., RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–50.

ROY, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers* **3** 135–145.

RUBIN, D. (1974). Estimating causal effects to treatments in randomised and nonrandomised studies. *Journal of Educational Psychology* **66** 688–701.

SIANESI, B. (2004). An evaluation of the active labour market programmes in Sweden. *The Review of Economics and Statistics* **86** 133–155.

SMITH, J. (2000). A critical survey of empirical methods for evaluating active labour market policies. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* **136** 1–22.

SMITH, J. (2004). Evaluating local development policies: Theory and practice. Working Paper, University of Maryland.

SMITH, J., TODD, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* **125** 305–353.

VYTLACIL, E. (2002). Independence, monotonicity and latent index models: An equivalence result. *Econometrica* **70** 331–341.

Marco Caliendo
DIW Berlin
Abteilung Staat
Königin-Luise Str. 5, 14195 Berlin
Germany
mcaliendo@diw.de

Reinhard Hujer
Institut für Statistik und Ökonometrie
J.W.Goethe Universität
Mertonstr. 17, 60054 Frankfurt
Germany
hujer@wiwi.uni-frankfurt.de