# Nobody Knows...
# How do Different Evaluation Estimators Perform in a Simulated Labour Market Experiment?

Reinhard Hujer, Marco Caliendo and Dubravko Radić*

This draft: June 13, 2001

Discussion Paper

**Abstract**

The need for evaluation of active labour market policies is no longer questioned. Recent overviews of microeconometric evaluation studies for Germany have shown substantial differences regarding the estimated effects of these measures. Whereas most of the differences might be due to different data sets used or different time periods examined, some of the variation is likely to be induced by different evaluation approaches. It is obvious, that nobody knows the real extent of the programme effects, making it hard to assess the quality of different estimation techniques.

The aim of this paper is twofold. In the beginning we present four different evaluation strategies (before-after, cross-section, matching, difference-in-differences estimator) and discuss the methodological concepts associated with them. Whilst presenting the methodological concepts of each estimator, we will also discuss the data needed for their implementation. By doing so, we hope to give advice for the construction of datasets in the future. Then we perform a simulated labour market experiment, which allows us to control and estimate all relevant effects. We are going to show how the different estimators perform under different conditions. The experiment is designed to take into account observable as well as unobservable characteristics in determining the labour market success and the participation decision of individuals. This design enables us to show the (dis-) advantages of each estimation technique and present the environments under which they work best.

**Keywords:** Evaluation, Matching, Difference-in-Differences, Monte Carlo Simulation

**JEL Classification:** C13, C14, C15, H43, J68

*Reinhard Hujer is Professor of Statistics and Econometrics, Marco Caliendo and Dubravko Radić are Research Assistants at the Chair of Statistics and Econometrics. Corresponding Author: Reinhard Hujer, J.W.Goethe-University, Department of Economics, Institute for Statistics and Econometrics, Mertonstr.17, 60054 Frankfurt, Germany, e-mail: hujer@wiwi.uni-frankfurt.de. The authors thank Björn Christensen, Stefan Kokot and Viktor Steiner for valuable comments.

# Contents

# 1. Introduction

Many OECD countries have been plagued by high and persistent unemployment since the early 1970s. Active labour market policies (ALMP) have been seen as one way to fight this unacceptable situation. It has been stated that ALMP are capable to meet efficiency and equity goals at the same time, by providing a more efficient outcome on the labour market, and also equipping individuals with higher skills and therefore lowering the risk of poverty (OECD (1991)). Therefore it is not surprising that public spending on ALMP has grown in the last years, absorbing significant shares of national resources. The German Federal Employment Office spent 45.3 billion DM on ALMP in 1999 which amounts to 1.1 % of the Gross Domestic Product. These spending are than not available for other projects or private consumption. In an era of tight government budgets and a growing disbelief regarding the positive effects of ALMP, evaluation of these policies becomes imperative.

The ideal evaluation process can be viewed as a series of three steps. First, the impacts of the programme on the individual or groups of individuals should be estimated. Second, it should be examined if the estimated impacts are large enough to yield net social gains. Finally, it should be answered if this is the best outcome that could have been achieved for the money spent (Fay (1996)). The focus of our paper is the first step, namely the microeconometric evaluation.

Empirical microeconometric evaluation is conducted with individual data. The main question is if the interesting outcome variable for an individual is affected by the participation in an ALMP programme. Relevant outcome variables could be the future employment probability or the future earnings. We would like to know the difference between the value of the participants outcome in the actual situation and the value of the outcome if he or she had not participated in the programme. The fundamental evaluation problem arises because we 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 is necessary to make a comparison possible. This is not easy because participants in programmes usually differ systematically in more aspects than just participation from the non-participants. Taking simply the difference between their outcomes after training will not reveal the true training impact, i.e. will lead to a biased estimate. The literature on the solution to this problem is dominated by two points of view.

Some analysts like LaLonde (1986) or Ashenfelter and Card (1985) view social experiments as the only valid evaluation method, whereas a second group of researchers like Heckman and Hotz (1989) or Lechner (1998) believe that it is possible to construct a comparison group using non-experimental data and econometric and statistical methods to solve the fundamental evaluation problem. In non-experimental or observational studies, the data are not derived in a process that is completely under the control of the researcher. Instead one has to rely on information about how individuals actually performed after the intervention, that is we observe the outcome with treatment for participants and the outcome without treatment for non-participants. The objective of observational studies is to use this information to restore the comparability of both groups by design. To do so, more or less plausible identification assumptions have to be imposed. There are several approaches

differing with respect to the methods applied for this problem. Some studies control for observables as part of parametric evaluation models, others construct matched samples. Furthermore some authors think that conditioning on observables is not enough and one has to take into account unobservables, too.

Recent overviews of microeconometric evaluation studies for Germany (Hagen and Steiner (2000) or Hujer and Caliendo (2000)) have shown substantial differences regarding the estimated effects of these measures. Whereas most of the differences might be due to different data sets used or different time periods examined, some of the variation is likely to be induced by different evaluation approaches.

Since in an observational evaluation study the "real" extent of the programme effects is not known, it is hard to assess the quality of different estimation techniques. Therefore we perform a simulated labour market experiment, which allows us to control and estimate all relevant effects. We are going to present four different estimation techniques, discuss the methodological concepts associated with them and show how they perform under different conditions. The experiment is designed to take into account observable as well as unobservable characteristics in determining the labour market success of participants and non-participants. This design enables us to investigate the (dis-) advantages of each estimation technique and identiy the environments under which they work best. Whilst presenting the methodological concepts of each estimator, we will also discuss the data requirements needed for their implementation. By doing so, we hope to give advice for the construction of datasets in the future.

The remainder of this paper is organized as follows. Section 2 starts with a discussion of the fundamental evaluation problem in microeconometric studies, and four different estimation techniques are presented. Section 3 describes the design of our Monte Carlo experiment and several scenarios we tested. The results for each estimator in the different scenarios can be found in this section, too. Section 4 concludes and gives an outlook for further research.

# 2. Microeconometric Evaluation

## 2.1. 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 on the labour market, had he or she not received the treatment.[1] The framework serving as a guideline for the empirical analysis of this problem is the potential outcome approach, also known as the Roy-Rubin-model (Roy (1951), Rubin (1974)).

In the basic model there are two potential outcomes (or responses), $Y^T$ and $Y^C$, for each individual, where $Y^T$ indicates a situation with training and $Y^C$ without. In the former case, the individual is in the treatment group and in the latter case it is in the comparison group. To complete the notation we define a binary assignment indicator $D$, indicating whether an individual actually participated in training ($D = 1$) or not ($D = 0$) (Hujer and

[1] This is clearly different from asking whether there is an empirical association between training and the outcome (Lechner (2000)). See Holland (1986) for an extensive discussion of concepts of causality in statistics, econometrics and other fields.

Wellner (2000), Lechner (2000)). The treatment effect for each individual is then defined as the difference between his/her potential outcomes:

$$\Delta = Y^T - Y^C. \tag{1}$$

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

$$Y = D \cdot Y^T + (1 - D) \cdot Y^C. \tag{2}$$

Unfortunately we can never observe $Y^T$ and $Y^C$ for the same individual simultaneously. The unobservable component in (1) is called the counterfactual outcome, so that for individuals who participated in the training measure ($D = 1$), $Y^C$ is the counterfactual outcome, and for those who did not it is $Y^T$.

The 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 $\Delta$ for each person is independent of the treatment of other individuals. In the statistical literature (Rubin (1980)) this is referred to as the stable unit treatment value assumption (SUTVA) and guarantees that average treatment effects can be estimated independently of the size and composition of the treatment population.[2] Note that there will be no opportunity to ever estimate individual effects with confidence. Therefore we have to concentrate on the population average of gains from treatment. The most prominent evaluation parameter is the so-called mean effect of treatment on the treated:

$$E(\Delta \mid D = 1) = E(Y^T \mid D = 1) - E(Y^C \mid D = 1). \tag{3}$$

In the sense that this parameter focuses directly on actual training participants, it determines the realized gross gain from the training programme and can be compared with its costs, helping to decide whether the programme is a success or not (Heckman, Ichimura, Todd (1997, 1998), Heckman, LaLonde, Smith (1999)).

Despite the fact that most evaluation research focuses on average outcomes, partly because most statistical techniques focus on mean effects, there is also a growing interest regarding effects of policy variables on distributional outcomes. Examples where distributional consequences matter for welfare analysis include subsidized training programmes (LaLonde (1995)) or minimum wages (DiNardo, Fortin, and Lemieux (1996)). Koenker and Bilias (2000) show that quantile regression methods can play a constructive role in the analysis of duration (survival) data, too. They describe the link between quantile regression and the transformation model formulation of survival analysis, offering a more flexible analysis than conventional methods.

Nevertheless we will focus on the average treatment effect on the treated $E(\Delta \mid D = 1)$ in this paper. The second term on the right side in equation (3) is unobservable as it describes the hypothetical outcome without treatment for those people who received treatment. If the condition

[2] Among other things SUTVA excludes cross-effects or general equilibrium effects. Its validity facilitates a manageable formal setup; nevertheless in practical applications it is frequently questionable whether it holds.

$$E(Y^C \mid D = 1) = E(Y^C \mid D = 0) \tag{4}$$

holds, we can use the non-participants as an adequate control group. In other words we would take the mean outcome of non-participants as a proxy for the counterfactual outcome of participants. This identifying assumption is definitely valid in social experiments. The key concept here is the randomized assignment of individuals into treatment and control groups. Individuals who are eligible to participate in training are randomly assigned to a treatment group which participates in the programme and a control group that does not. This assignment mechanism is a process that is completely beyond the employees' or the administrators' control. If the sample size is sufficiently large, randomization will generate a complete balancing of all relevant observable and unobservable characteristics across treatment and control groups. Therefore the comparability between experimental treatment and control groups is facilitated enormously. On average, the two groups do not systematically differ except for having participated in training. As a result any observed difference in the outcomes of the groups after training is supposed to be solely induced by the programme itself, i.e. the impact of training is isolated and there should be no selection bias. Formally, random assignment ensures, that the potential outcomes are independent of the assignment to the training programme. We write:

$$Y^T, Y^C \amalg D, \tag{5}$$

$\amalg$ denoting independence. When assignment to treatment is completely random it follows that:

$$E(Y^C \mid D = 1) = E(Y^C \mid D = 0),$$

and

$$E(Y^T \mid D = 1) = E(Y^T \mid D = 0). \tag{6}$$

Therefore, treatment assignment becomes ignorable (Rubin (1974)) and we get an unbiased estimate of $E(\Delta)$, i.e. the randomly generated group of non-participants can be used as an adequate control group to consistently estimate the counterfactual term $E(Y^T \mid D = 0)$ and thus the causal training effect $E(\Delta \mid D = 1)$. Although this approach seems to be very appealing in providing a simple solution to the fundamental evaluation problem, there are also some problems associated with it. Besides relatively high costs and ethical issues concerning the use of experiments, in practice, a randomized experiment may suffer from similar problems, that affect behavioural studies. Bijwaard and Ridder (2000) investigate the problem of non-compliance to the assigned intervention, that is, when members of the treatment sample drop out of the programme and members of the control group participate. If the non-compliance is selective, i.e. correlated with the outcome variable, the difference of the average outcomes is a biased estimate of the effect of the intervention, and correction methods have to be applied, too. Further methodological problems might arise, like a

substitution or randomization bias, which make the use of experiments questionable.[3] For an extensive discussion of these topics the interested reader should refer to Burtless (1995), Burtless and Orr (1986) and Heckman and Smith (1995).[4]

More important for practical applications is the fact that in most European countries experiments are, out of several reasons, not conducted and researchers have to work with non-experimental data anyway. In non-experimental data, equation (4) will normally not hold:

$$E(Y^C \mid D = 1) \neq E(Y^C \mid D = 0) \tag{7}$$

The use of the non-participants as a control group might therefore lead to a selection bias. Heckman and Hotz (1989) point out that selection might occur on observable or unobservable characteristics. The aim of any observational evaluation approach is to ensure the comparability of treatment and control group by design, that is through a plausible identifying assumption. Taking account of observable factors might not be sufficient, if unobservable factors invalidate the comparison, e.g. when more motivated workers have a higher employment probability and are also more likely to participate in a training programme (Schmidt (1999)).

In the following sub-sections we will present four different evaluation approaches. Each approach invokes different identifying assumptions to construct the required counterfactual outcome. Therefore each estimator is only consistent in a certain restrictive environment. As Heckman, LaLonde, and Smith (1999) note, all estimators would identify the same parameter only if there is no selection bias at all.

## 2.2. Before-After Estimator

The most obvious and still widely used evaluation strategy is the before-after estimator (BAE). The basic idea is that the observable outcome in the pre-training period $t'$ represents a valid description of the unobservable counterfactual outcome of the participants without training in the post-training period $t$. The central identifying assumption of the before-after estimator can be stated as:

$$E(Y_{t'}^C \mid D = 1) = E(Y_t^C \mid D = 1). \tag{8}$$

Given the identifying assumption in (8), the following estimator of the mean treatment effect on the treated can be derived:

$$\Delta^{BAE} = E[(Y_t^T \mid D = 1) - (Y_{t'}^C \mid D = 1)] \tag{9}$$

[3] A randomization bias occurs when random assignment causes the types of persons participating in a programme to differ from the type that would participate in the programme as it normally operates, leading to an unrepresentative sample. We talk about a substitution bias when members of an experimental control group gain access to close substitutes for the experimental treatment (Heckman and Smith (1995)).

[4] As Smith (2000) notes, social experiments have become the method of choice in America. The most famous among them is the National Job Training Partnership Act which had a major influence regarding the view on non-experimental studies. In Europe however social experiments have not received a similar acceptance, although recently some test experiments have been conducted. The most important one is the RESTART experiment in Britain (see Dolton and O´Neill (1996)).

Heckman, LaLonde, and Smith (1999) note that conditioning on observable characteristics makes it more likely that assumption (8) will hold. If the distribution of $X$ characteristics is different between the treatment and the control group, conditioning on $X$ may eliminate systematic differences in the outcomes.[5] In our Monte Carlo study $X$ will represent the qualification level of an individual, where $X = 1$ indicates a high-skilled individual and $X = 0$ a low-skilled individual. Therefore, conditioning on $X$ results in estimating the treatment effects separately for both skill groups and is intuitively appealing. The first approach in equation (9) can be seen as a 'naive' before-after estimator, because it just compares the results for the whole group of participants before and after the programme took place, whereas the second takes into account observed differences in individual's characteristics like the different skill levels. Conditioning on $X$ gives us a new identifying assumption:

$$E(Y_{t'}^C \mid X, D = 1) = E(Y_t^C \mid X, D = 1). \tag{10}$$

The new estimator can be written as:

$$\Delta^{BAE|X} = E[(Y_t^T \mid X, D = 1) - (Y_{t'}^C \mid X, D = 1)]. \tag{11}$$

The validity of (10) depends on a set of implicit assumptions. First of all, the pre-exposure potential outcome without training should not be affected by training. This may be invalid if individuals have to behave in a certain way in order to get into the programme or behave differently in anticipation of a future training participation. Secondly, no time-variant effects should influence the potential outcomes from one period to the other. If there are changes in the overall state of the economy or changes in the lifecycle position of a cohort of participants, assumption (10) may be violated (Heckman, LaLonde, and Smith (1999)).

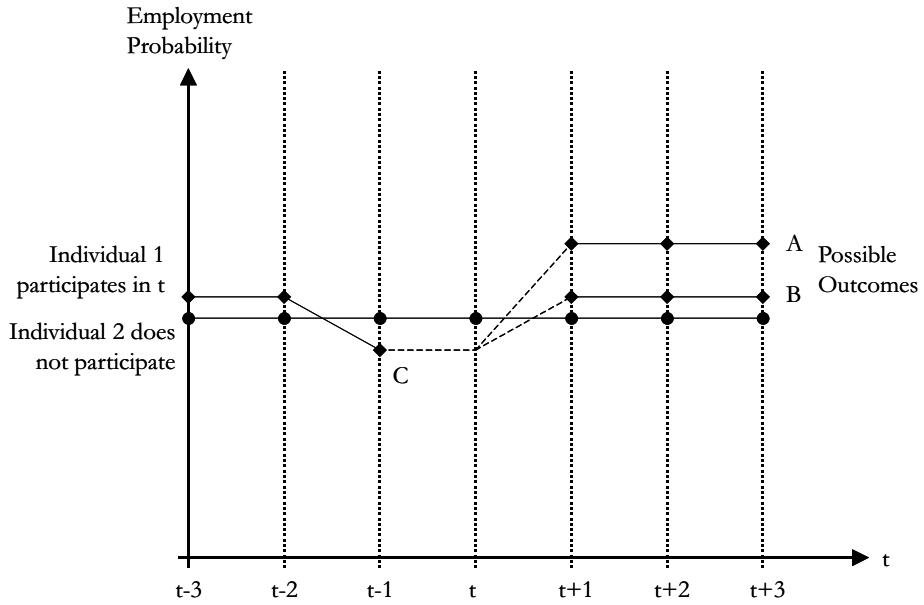
A good example where this might be the case is Ashenfelter's dip that is a situation where shortly before the participation in an ALMP programme the employment situation of the future participants deteriorates. Ashenfelter (1978) found this 'dip' whilst evaluating the effects of treatment on earnings, but later research demonstrated that this dip can be observed on employment probabilities for participants, too. If the dip is transitory and only experienced by participants assumption (10) will not hold. By contrast, permanent dips are not problematic, as they affect the employment probability before and after the treatment in the same way. Figure 1 illustrates Ashenfelter's dip.

We assume that individual 1 participates in a training programme in period $t$ and experiences a transitory dip in the pre-training period $t-1$, e.g. the employment probability is lowered as the individual is not actively seeking work, because it knows that it will participate in a programme in the next period. There are no economy-wide time-varying effects that influence the employment probability of the individual (as can be seen by the development of individual 2, who has a fairly similar employment probability as individual 1 but does not participate in the programme). After the programme took place, the employment probability might take many values. For the sake of simplicity we consider two cases. Case $A$ assumes that there has been a positive treatment effect on the employment probability

[5] On the other hand, if the difference in the treatment and the control group are due to unobservable characteristics, conditioning may accentuate rather than eliminate the differences in the no-programme state between the both groups (Heckman, LaLonde, and Smith (1999)).

(vertical difference between $A$ and $B$). As the BAE compares the employment probability of the individual in period $t+1$ and $t-1$ (vertical difference between point $A$ and $C$) it would overestimate the true treatment effect, because it would attribute the restoring of the transitory dip completely to the programme. In a second example, we assume, that there has been no treatment effect at all ($B$). In period $t+1$ the employment probability is restored to its original value before the dip took place in $t-1$. Again the BAE attributes this restoring completely to the programme and would estimate a positive treatment effect (vertical difference between point $B$ and $C$). Clearly the problem of Ashenfelter's dip could be avoided, if a time period is chosen as a reference level, before the dip took place, e.g. period $t-2$. But as we will not know a-priori in an empirical application when the dip starts the question arises, which period to choose.

Figure 1: **Ashenfelter's Dip**




A major advantage of the before-after estimator is, that it does not require information on non-participants. All that is needed is longitudinal data on outcomes on participants before and after the programme took place.[6] As the employment status of the participants is known in nearly all of the programmes in Germany, the BAE does not impose any major problems regarding the data availability, which might explain why it is still widely used.[7]

## 2.3. Cross-Section Estimator

Instead of comparing participants at two different time periods, the cross-section estimator compares participants and non-participants at the same time (after the programme took

[6] The BAE might also work with repeated cross-sectional data from the same population, not necessarily containing information on the same individuals. See Heckman and Robb (1985) or Heckman, LaLonde, and Smith (1999) for details.

[7] Note that being unemployed is one of the entry conditions for many ALMP programmes in Germany.

place), i.e. the population average of the observed outcome of non-participants replaces the population average of the unobservable outcome of participants. This is useful if no longitudinal information on participants is available or macroeconomic conditions shift substantially over time (Schmidt (1999)). The identifying assumption of the cross-section estimator can be stated formally as:

$$E(Y_t^C \mid D = 1) = E(Y_t^C \mid D = 0), \tag{12}$$

that is persons who participate in the programme have on average the same no-treatment outcome as those who do not participate. If this assumption is valid, the following estimator of the mean true treatment effect can be derived:

$$\Delta^{CSE} = E[(Y_t^T \mid D = 1) - (Y_t^C \mid D = 0)]. \tag{13}$$

Again, conditioning on observable characteristics $X$ leads to assumption

$$E(Y_t^C \mid X, D = 1) = E(Y_t^C \mid X, D = 0), \tag{14}$$

and the corresponding estimator:

$$\Delta^{CSE|X} = E[(Y_t^T \mid X, D = 1) - (Y_t^C \mid X, D = 0)]. \tag{15}$$

Schmidt (1999) notes, that for assumption (14) to be valid, selection into treatment has to be statistically independent of its effects given $X$ (exogenous selection), that is, no unobservable factor should lead individual workers to participate. A good example where this is violated might be the case if motivation plays a role in determining the desire to participate and the no-treatment outcomes. Then we have even in the absence of any treatment effect, a higher average outcome in the participating group compared to the non-participating group. Ashenfelter's dip is not problematic for the cross-section estimator, as we compare only participants and non-participants after the programme took place. Moreover, as long as economy-wide shocks and individuals lifecycle patterns operate identically for the treatment and the control group, the cross-section estimator is not vulnerable to the problems that plague the BAE (Heckman, LaLonde, and Smith (1999)).

## 2.4. Matching Estimator

The matching approach originated in the statistical literature and shows a close link to the experimental context.[8] The basic idea underlying the matching approach is to find in a large group of non-participants those individuals who are similar to the participants in all relevant pre-training characteristics. That being done, the differences in the outcomes between the well selected and thus adequate control group and the trainees can then be attributed to the programme. Matching does not need to rely on functional form or distributional assumptions, as its nature is non-parametric (Augurzky (2000)).

Of course matching is first of all plagued by the same problem as all non-experimental estimators, which means that assumption (4) cannot be expected to hold when treatment

[8] See Rubin (1974), (1977), (1979), Rosenbaum and Rubin (1983), (1985a), (1985b) or Lechner (1998))

assignment is not random. However, following Rubin (1977) treatment assignment may be random given a set of covariates. The construction of a valid control group via matching is based on the identifying assumption that conditional on all relevant pre-training covariates $Z$, the potential outcomes $(Y^T, Y^C)$ are independent of the assignment to training.[9] This so called conditional independence assumption (CIA) can be written formally as:

$$Y^T, Y^C \amalg D \mid Z.^{10} \tag{16}$$

If assumption (16) is fulfilled we get:

$$E(Y^C \mid Z, D = 1) = E(Y^C \mid Z, D = 0) = E(Y^C \mid Z). \tag{17}$$

Similar to randomization in a classical experiment, the role of matching is to balance the distributions of all relevant pre-treatment characteristics $Z$ in the treatment and control group, and thus to achieve independence between potential outcomes and the assignment to treatment, resulting in an unbiased estimate.The exact matching estimator can be written as:

$$\Delta^{MAT} = E[(Y_t^T \mid Z, D = 1) - (Y_t^C \mid Z, D = 0). \tag{18}$$

Conditioning on all relevant covariates is, however, limited in case of a high dimensional vector $Z$. For instance, if $Z$ contains $n$ covariates which are all dichotomous, the number of possible matches will be $2^n$. In this case cell matching, that is exact matching on $Z$, is in practice not possible, because an increase in the number of variables increases the number of matching cells exponentially (Hujer and Wellner (2000)). To deal with this dimensionality problem, Rosenbaum and Rubin (1983) suggest the use of balancing scores $b(Z)$, i.e. functions of the relevant observed covariates $Z$ such that the conditional distribution of $Z$ given $b(Z)$ is independent of the assignment to treatment, that is $Z \amalg D \mid b(Z)$ holds.

For trainees and non-trainees with the same balancing score, the distributions of the covariates $Z$ are the same, i.e. they are balanced across the groups. Moreover Rosenbaum and Rubin (1983) show that if the treatment assignment is strongly ignorable[11] when $Z$ is given, it is also strongly ignorable given any balancing score. The propensity score $P(Z)$, i.e. the probability of participating in a programme is one possible balancing score. It summarizes the information of the observed covariates $Z$ into a single index function. Rosenbaum and Rubin (1983) show how the conditional independence assumption extends to the use of the propensity score so that

$$Y^C \amalg D \mid P(Z). \tag{19}$$

[9] If we say relevant we mean all those covariates that influence the assignment to treatment as well as the potential outcomes. In contrast to the cross-section estimator, the matching procedure can also use information from the pre-treatment period, like the employment status or other time-varying covariates. To make this difference clear, we denote the covariates by $(Z)$.

[10] For the purpose of estimating the mean effect of treatment on the treated the assumption of conditional independence of $Y^C$ is sufficient, because we like to infer estimates of $Y^C$ for persons with $D = 1$ from data on persons with $D = 0$ (Heckman, Ichimura, and Todd (1997)).

[11] Strongly ignorable means that assumption (16) holds and: $0 < P(Z) \equiv P(D = 1 \mid Z) < 1$. The latter ensures, that there are no characteristics in $Z$ for which the propensity score is zero or one. Proofs go beyond the scope of this work and can be found in Rosenbaum and Rubin (1983).

Therefore we get:

$$E(Y^C \mid P(Z), D = 1) = E(Y^C \mid P(Z), D = 0) = E(Y^C \mid P(Z)), \tag{20}$$

which allows us to rewrite the crucial term in the average treatment effect (3) as:

$$E(Y^C \mid D = 1) = E_{P(Z)}[E(Y^C \mid P(Z), D = 0) \mid D = 1]. \tag{21}$$

Hujer and Wellner (2000) note that the outer expectation is taken over the distribution of the propensity score in the treated population. The major advantage of the identifying assumption (19) is that it transforms the estimation problem into a much easier task since one has to condition on an univariate scale, i.e. on the propensity score, only. When $P(Z)$ is known the problem of dimensionality can be eliminated. The evaluation of the counterfactual term via matching on the basis of the group of non-participants then only requires to pair participants with non-participants which have the same propensity score. This insures a balanced distribution of $Z$ across both groups. Unfortunately $P(Z)$ will not be known a-priori so it has to be replaced by an estimate. This can be achieved by any number of standard probability models, e.g. a probit model. The empirical power of matching to reduce the problem of selection bias relies crucially on the quality of the estimate of the propensity score on the one hand and on the existence of comparison persons that have equal propensity scores as the treated persons. If the latter is not ensured we face the risk of incomplete matching with biased estimates.[12] Several procedures for matching on the propensity score have been suggested and will be discussed briefly, a good overview can be found in Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2000).[13] To introduce them a more general notation is needed: We estimate the effect of treatment for each observation $i$ in the treatment group, by contrasting his/her outcome with treatment with a weighted average of control group observations in the following way:

$$Y_i^T - \sum_{j \in \{D=0\}} W_{N_0 N_1}(i, j) Y_j^C, \tag{22}$$

where $N_0$ is the number of observations in the control group and $N_1$ is the number of observations in the treatment group. Matching estimators differ in the weights attached to the members of the comparison group (Heckman, Ichimura, Smith, and Todd (1998).

Nearest neighbour (NN) matching sets

$$C(P_i) = \min_j \|P_i - P_j\|, j \in N_0. \tag{23}$$

Doing so, the non-participant with the value of $P_j$ that is closest to $P_i$ is selected as the

[12] Matching was much discussed in the recent econometric literature. Heckman and his colleagues reconsidered and further developed the identifying assumptions of matching stated by Rubin (1977) and Rosenbaum and Rubin (1983). It turns out that the new identifying assumptions are weaker compared to the original statements which brings along some advantages. Presenting these ideas goes beyond the scope of this work. The interested reader should refer to Heckman, Ichimura, Smith and Todd (1996, 1998), Heckman, Ichimura and Todd (1997,1998) and Heckman and Smith (1995).

[13] At this point of the paper we have not been able to take into account different suggested matching procedures. This should be done in further studies. For our Monte Carlo study we will rely on the exact matching approach.

match, therefore $W_{N_0N_1}(i,j) = 1$ for this unit and $W_{N_0N_1}(i,j) = 0$ otherwise.[14] Several variants of NN matching are proposed, e.g. NN matching 'with' and 'without replacement'. In the former case a non-participating individual can be used more than once as a match, whereas in the latter case it is considered only once. It is also suggested to use more then one nearest neighbour ('oversampling'). NN matching faces the risk of bad matches, if the closest neighbour is far away.

This can be avoided by imposing a tolerance on the maximum distance $\|P_i - P_j\|$ allowed. This form of matching, caliper matching (Cochrane and Rubin (1973)), imposes the condition:

$$\|P_i - P_j\| < \epsilon, j \in N_0, \tag{24}$$

where $\epsilon$ is a pre-specified level of tolerance.

Kernel matching (KM) is a nonparametric matching estimator that uses all units in the control group to construct a match for each programme participant. KM defines:

$$W_{N_0}(i,j) = \frac{K_{ij}}{\sum_{k \in \{D=0\}} K_{ik}}, \tag{25}$$

where $K_{ik} = K((P_i - P_k)/h)$ is a kernel that downweights distant observations from $P_i$ and $h$ is a bandwidth parameter (Heckman, Ichimura, Smith, and Todd (1998)). A generalized version of KM is local linear (LL) matching, that has some advantages like a faster rate of convergence near boundary points and greater robustness to different data design densities (Heckman, Ichimura, and Todd (1997)).

The matching estimator is very data demanding in the sense, that we need information for participants and non-participants before and after the programme took place. In the case of exact matching a 'rich' dataset is needed to ensure that we find comparable individuals in the control group for every combination of observable characteristics. Even if we do not use exact matching but matching over the propensity score, a rich dataset is needed. In that case the quality of the score depends on our ability to account for all relevant covariates that determine the participation decision.

## 2.5. Difference-in-Differences and Conditional Difference-in-Differences

It has been claimed, that controlling for selection on observables may not be sufficient since remaining unobservable differences may still lead to a biased estimation of treatment effects. These differences may arise from differences in the benefits which individuals expect from participation in a treatment that might influence their decision to participate. Furthermore some groups might exhibit bad labour market prospects or differences in motivation. These features are unobservable to a researcher and might cause a selection bias.

To account for selection on unobservables, Heckman, LaLonde, and Smith (1999) suggest econometric selection models and difference-in-differences (DID) estimators. The DID estimator requires access to longitudinal data and can be seen as an extension to the classical before-after estimator. Whereas the BAE compares the outcomes of participants after they

[14] Exact Matching imposes an even stronger condition, that is only non-participants with exactly the same propensity score or the same realization of characteristics $X$ are considered as matches.

participate in the programme with their outcomes before they participate, the DID estimator eliminates common time trends by subtracting the before-after change in non-participant outcomes from the before-after change for participant outcomes. The simplest application of the method does not condition on $X$ and forms simple averages over the group of participants and non-participants, that is changes in the outcome variable $Y$ for the treated individuals are contrasted with the corresponding changes for non-treated individuals (Heckman, Ichimura, Smith, and Todd (1998)):

$$\Delta^{DID} = [Y_t^T - Y_{t'}^C \mid D = 1] - [Y_t^C - Y_{t'}^C \mid D = 0]. \tag{26}$$

The DID estimator is based on the assumption of time-invariant linear selection effects. The critical identifying assumption of this method is, that the biases are the same on average in different time periods before and after the period of participation in the programme, so that differencing the differences between participants and non-participants eliminates the bias (Heckman, Ichimura, Smith, and Todd (1998)). To make this point clear, we give an example by denoting the outcome for an individual $i$ at time $t$ as:

$$Y_{it} = \alpha_{it} + D_{it} \cdot Y_{it}^T + (1 - D_{it}) \cdot Y_{it}^C, \tag{27}$$

where $\alpha_{it}$ captures the effects of selection on unobservables. The validity of the DID estimator relies crucially on the assumption:

$$\alpha_{it} = \alpha_{it'}. \tag{28}$$

Only if the selection effect is time-invariant it will be cancelled out and an unbiased estimate results. The differencing leads to:

$$\begin{aligned} Y_{it} - Y_{it'} \quad = \quad & [D_{it} \cdot Y_{it}^T + (1 - D_{it}) \cdot Y_{it}^C] - \\ & [D_{it'} \cdot Y_{it'}^T + (1 - D_{it'}) \cdot Y_{it'}^C] + \\ & [\alpha_{it} - \alpha_{it'}]. \end{aligned} \tag{29}$$

If (28) is fulfilled the last term in the expression can be cancelled out, leading to an unbiased estimate. Compared to the method of matching the DID approach does not require that the bias vanishes for any 'matched' individuals, but only that it remains constant (Heckman, Ichimura, Smith, and Todd (1998)).

If we condition the DID approach on observable characteristics $X$ the new estimator is given by:

$$\Delta^{DID|X} = [Y_t^T - Y_{t'}^C \mid X, D = 1] - [Y_t^C - Y_{t'}^C \mid X, D = 0]. \tag{30}$$

The identifying assumption of this method is:

$$E(Y_t^T - Y_{t'}^C \mid X, D = 1) = E(Y_t^C - Y_{t'}^C \mid X, D = 0). \tag{31}$$

Ashenfelters's dip is definitely a problem for the DID estimator. 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. Therefore Bergemann, Fitzenberger, Schultz, and Speckesser (2000) apply a combination of the matching- and the DID estimator in a recent paper, by implementing a 'conditional difference-in-differences estimator', where conditional means, that treatment and control groups are already partly comparable regarding their observable characteristics. Kluve, Lehmann, and Schmidt (1999) suggest to use the pre-treatment (labour market) histories of the individuals as an important variable in the matching process, so that only individuals with identical pre-treatment histories are compared.

The estimator in equation (30) can easily be extended, by taking additionally the pre-treatment employment history ($Y_{t-1}$, $Y_{t-2}$) of the individuals into account:

$$\Delta^{DID|Z} = [Y_t^T - Y_t^{C\prime} \mid Z, D = 1] - [Y_t^C - Y_t^{C\prime} \mid Z, D = 0].^{15} \tag{32}$$

Basically the conditional DID estimator extends the one suggested from Heckman, Ichimura, Smith, and Todd (1998) by adding a longitudinal dimension. The basic question in this case is, the determination of the appropriate time span to be considered.

# 3. Monte Carlo Study

## 3.1. Study Design

In the following we will describe the design of our Monte Carlo study. We will examine four time periods, where the third period is our treatment period. We assume that the outcome variable $Y_{it}$ measures the employment status of an individual $i$ at a certain time period $t$. $Y_{it}$ is assumed to be distributed according to a Bernoulli distribution, which can either take the value of 0 (individual is unemployed) or 1 (individual is employed) with a given probability. This employment probability will be influenced by a set of covariates.

For simplicity we assume that the covariates have a linear additive effect on the employment probability and also include an intercept term. The probabilities used for simulating the outcome variable for the various time periods are as follows (see also figure 2):

Beginning with the first time period, the employment probability for $t = 1$ is assumed to be a function of observable and unobservable characteristics of an individual:

$$P(Y_{i1} = 1) = \alpha + \alpha_X X_i + \alpha_M M_i. \tag{33}$$

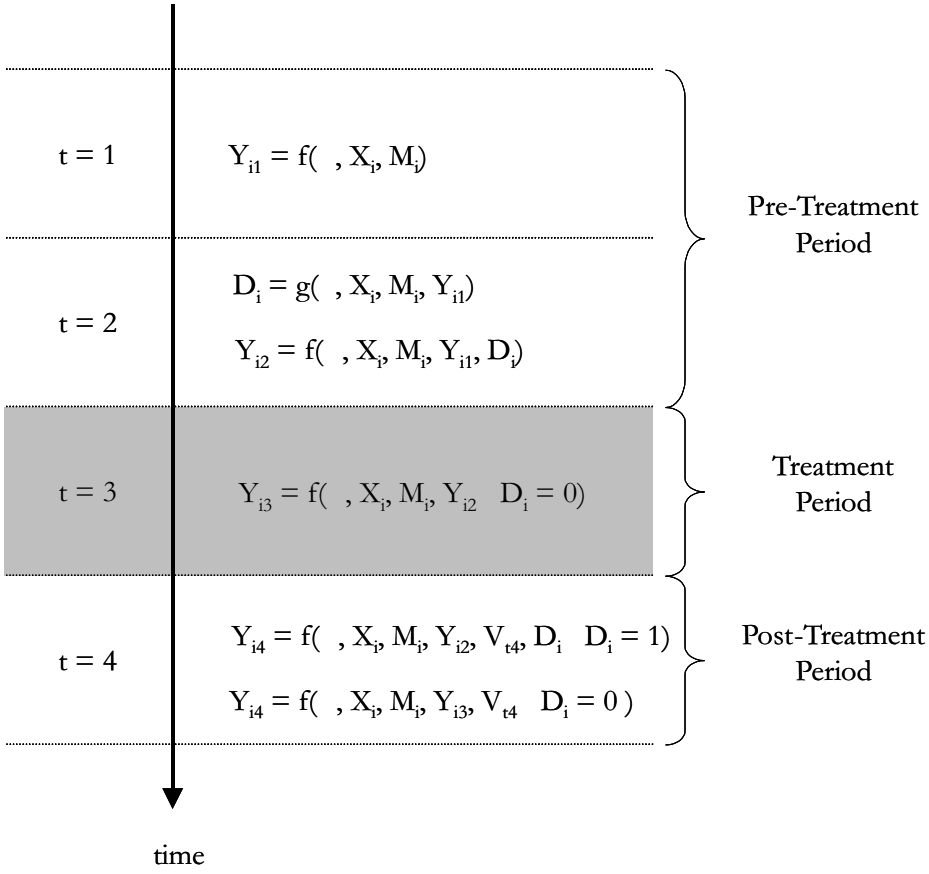
$X_i$ measures observable individual characteristics that influence the employment probability, e.g. the formal educational level. In the following we will consider two levels of education indicating high- ($X_i = 1$) and low-skilled employees ($X_i = 0$). A value for $\alpha_X > 0$ will thus indicate that high-skilled employees have a higher employment probability compared with low-skilled employees. Setting $\alpha_X$ equal to zero would indicate that the employment probability is not influenced by formal qualification.

$M_i$ measures unobservable factors like the motivation of an individual. A value of $\alpha_M > 0$ thus means that high-motivated individuals have a higher employment probability ceteris

[15] In our study design conditioning the DID on the previous labour market outcome of the individuals, $Y_{t-1}$, corresponds to the exact matching estimator. Therefore we conditioned only on $X$ (see table 3 in the appendix for details).

paribus whereas setting $\alpha_M = 0$ would imply the absence of such effects. It is assumed that the qualification and motivation levels are not correlated.[16]

Figure 2: **Time Schedule of the Monte Carlo Experiment**




At the beginning of our experiment we drew for 1.000 individuals values for $X_i$ and $M_i$. These were obtained by drawing separately 1.000 observations from a Bernoulli distribution. The probability to obtain a high-skilled individual was set for simplicity equal to 0.5. The same holds true for the probability to get an high-motivated individual.

After doing so, we generated for each of the 1.000 individuals the outcome variable for the first time period, $Y_{i1}$, by plugging the characteristics into (33).

In the next time period $t = 2$, the outcome variable is not only influenced by the observable and unobservable characteristics, but also by the previous employment status, thus allowing for persistency effects:

$$P(Y_{i2} = 1) = \alpha + \alpha_X X_i + \alpha_M M_i + \alpha_Y Y_{i1} + \alpha_{dip} D_i. \tag{34}$$

Our discussion in section 2.2 has shown, that Ashenfelter's dip might cause problems for some of the estimators. To examine the effects of such a transitory dip, we assume that

[16] By doing this we implicitly assume that formal education has only a modest signaling effect for the employer.

the participation decision for a programme is made at the beginning of period two, even though the programme itself takes place in period three. Including the participation decision in the outcome equation for the second period, enables us to model a dip and examine its consequences. A value for $\alpha_{dip}$ less then zero will indicate the presence of Ashenfelter's dip whereas setting $\alpha_{dip}$ equal to zero will imply the absence of such an effect.

The participation decision itself is influenced by a variety of variables leading to the following equation for the participation probability:

$$P(D = 1) = \beta + \beta_X X_i + \beta_M M_i + \beta_Y Y_{i1}. \tag{35}$$

We allow the participation decision to be influenced by observable and unobservable characteristics. This will offer us the opportunity to analyse how the presence of selection on observables and unobservables will influence the various estimators.

Setting e.g. $\alpha_X$ in the outcome equation and $\beta_X$ in the participation equation equal to zero will imply the absence of selection effects on observables whereas setting $\alpha_M$ and $\beta_M$ equal to zero will exclude the presence of selection on unobservables.

Using the simulated participation probabilities and again the Bernoulli distribution, we derive the participation decisions for our 1.000 individuals.

Our treatment individuals are assumed to be absent from the labour market as long as they receive the treatment, so we will not compute the outcome variable for the third period if $D_i = 1$. For those individuals who do not participate in the programme, the employment probability is again assumed to depend on observable and unobservable characteristics as well as on the lagged employment status:

$$P(Y_{i3} = 1|D_i = 0) = \alpha + \alpha_X X_i + \alpha_M M_i + \alpha_Y Y_{i2}. \tag{36}$$

The simulated employment status in $t = 4$ will in addition to all other variables also be influenced by the fact if an individual has participated in a programme in $t = 3$ or not:

$$\begin{aligned} P(Y_{i4} = 1) &= \alpha + \alpha_X X_i + \alpha_M M_i + \alpha_Y Y_{i2} D_i \\ &+ \alpha_Y Y_{i3}(1 - D_i) + \Delta D_i + \alpha_V V_t. \end{aligned} \tag{37}$$

$\Delta$ captures the effect of the programme on the outcome variable, i.e. $\Delta$ is the true treatment effect on the treated, the parameter of primary interest for our study.[17] By including $Y_{i2}$ (for participants) and $Y_{i3}$ (for non-participants) into the equation we again allow for persistency effects. Furthermore we also include a time dummy variable $V_t$ which takes into account effects common to all individuals, e.g. an economic upswing ($\alpha_V > 0$) or downturn ($\alpha_V < 0$), which of course will also have an impact on the estimators. Setting $\alpha_V$ equal to zero would indicate the absence of such time varying effects.

In the following sections we will present different scenarios which assume different sets of parameters and examine how the estimators introduced previously perform under these circumstances.

[17] Please note that the estimated true treatment effect may differ from $\Delta$ due to the randomness of the Monte Carlo study.

## 3.2. Description of the Different Scenarios

We decided to introduce six different scenarios, whereby scenario A will be our reference scenario. In this scenario all identifying assumptions of the different estimators will be met, whereas in each of the scenarios B, C and D, one of those assumptions will be violated. The last two scenarios, E and F, should reflect more realistic economic conditions. Every scenario is defined by a different set of parameters (see table 1 in the appendix), and for every scenario 1.000 random samples each consisting of 1.000 individuals were drawn.[18] The scenarios were the following (see also table 2 in the appendix):

**Scenario A:** In the basis scenario we assume that there are no selection effects on observable or unobservable characteristics. We also exclude the possibility that the previous employment status may affect the participation decision. Every individual has the same probability to participate in the programme which was set equal to 0.5 in order to get a randomization of the participation decision. Additionally we assume that there is neither Ashenfelter's dip nor changing economic conditions. With respect to the outcome variable high-skilled employees have a higher employment probability while motivation as well as the previous employment status play no role. As already mentioned, the high-skilled and motivated employees are assumed to be distributed equally in the population. The true treatment effect on the treated is set equal to 0.05.

**Scenario B:** In scenario B we will assume that there is an economic downturn in the fourth period by setting $\alpha_V = -0.04$. All other parameters remain the same as in the basis scenario.

**Scenario C:** This scenario allows for Ashenfelter's dip by setting the parameter $\alpha_{dip}$ equal to $-0.02$ whereas all other parameters remain the same as in the basis scenario.

**Scenario D:** The only difference here compared to the basis scenario is the introduction of selection on unobservables by allowing the employment probability as well as the probability to participate in a programme to depend on the motivation of an individual. By setting $\alpha_M = 0.02$ and $\beta_M = 0.02$, high-motivated individuals are more likely to participate in a programme and also to be employed in a certain period.

**Scenario E:** Scenario E allows the employment probabilities in different time periods to depend on more than one factor. By setting $\alpha_X$ equal to 0.25 and $\alpha_M$ equal to 0.15 the employment probability for high-skilled as well as for high-motivated is higher than for the reference group of low-skilled and low-motivated individuals. Setting $\alpha_Y = 0.3$ allows for persistency effects, i.e. an already employed individual is more likely to be employed in the future as well. Furthermore we assume an economic downturn in period $t = 4$ by setting $\alpha_V$ equal to -0.1. Additionally, we assume that individuals participating in the programme decrease their pre-programme efforts to get employed by setting $\alpha_{dip}$ = -0.01. Regarding the

[18] By considering the question how many trials are needed in a single Monte Carlo experiment, we followed Mooney (1997), who recommends that the best practical advice on how many trials are needed for a given experiment is lots!

participation decision we assume that selection effects on observable characteristics prevail by setting $\beta_X = 0.4$, and $\beta_M$ as well as $\beta_Y$ to 0.05.

**Scenario E:** This scenario differs from scenario E only with regard to the participation decision. While in scenario E selection effects on observables prevail, here selection on unobservable characteristics is dominant. This is done by setting $\beta_X$ and $\beta_Y$ equal to 0.05 and $\beta_M$ to 0.4.

### 3.3. Evaluating the Evaluation Estimators

In the following we will summarize how the various estimators performed in the different scenarios. Table 3 presents the implemented estimators. Table 4 contains the means, standard deviations and mean squared errors (MSE) of the different estimators. Additionally, we plotted for each estimator so called Kernel density estimates in Figures 3-8.[19] We will conclude this section with an overall assessment of the different estimators with regard to their relative (dis-)advantages and their data requirements.

**Scenario A:** Since we have excluded in this scenario any selection effects and also other effects that may violate the identifying assumptions of the various estimators, every estimator should yield an unbiased estimate of the true underlying treatment effect. Our Monte Carlo study confirms this a-priori expectation. All estimators have means around 0.05. The only difference is the standard deviation which for the unconditional and the conditional difference-in-differences estimator is slightly higher. Scenario A is the case mentioned by Heckman, LaLonde, and Smith (1999) where all estimators identify the same parameter since there is no selection bias at all.

**Scenario B:** In this scenario an economic downturn in the fourth period has been assumed. Since only the before-after estimator relies on the assumption that there are no changing economic conditions we expect this estimator to be biased downwards. Our Monte Carlo experiment confirms this a-priori expectation. The before-after approach estimates an average effect around 0.01, whereas all other estimators have means around 0.05. Again the variances for the different estimators are nearly the same except for the DID and the conditional DID which are markedly higher.[20]

**Scenario C:** In the Ashenfelter's dip scenario it is again the before-after estimator which suffers most and exhibits an upward bias. The difference-in-differences estimator also yields

[19] See Silverman (1986) for details on Kernel density estimation. We used Gaussian kernels and set the bandwidth according to:

$$h = \frac{0.9 min(SD, IQR/1.34)}{n^{1/5}} \tag{38}$$

where SD is the standard deviation of the datapoints and IQR the interquartile range of the data points (Silverman (1986), eq. (3.31)).

[20] One explanation could be, that all estimators except for the DID and the conditional DID are linear functions of only two random variables their variances will only consist of three parts, i.e. the variances of the two random variables and their covariance. On the other side, the DID and the conditional DID estimator are functions of four random variables. The variances of these estimators therefore are functions of four variances and additionally the 6 covariances between them. The variances of the DID and the conditional DID estimators will thus likely exceed the variance of the other estimators.

an upward bias which can not be eliminated even by conditioning on observable characteristics using the conditional DID. All other estimators which only use the cross-section information have no problem with this scenario and provide unbiased estimates of the true treatment effect.

**Scenario D:** The correlation between the participation decision and the outcome variable via the motivation of the individuals in this scenario, violates the assumption of no selection effects on unobservables. Our Monte Carlo experiment reveals that especially those estimators which do not account for selection effects on unobservables, i.e. the CSE and the matching estimator, yield biased estimates. The only estimators which do account for a possible selection effect on unobservables are the DID and the conditional DID estimators. Thus these are also the only estimators which are able to provide a more reliable estimate of the true treatment effect in this situation. Since only the assumption of no selection effects on unobservable characteristics is violated in this scenario, the BAE also provides an unbiased estimate of the true treatment effect with an even lower standard deviation.

**Scenario E:** Since we allowed more than one of the identifying assumptions to be violated here, we can not make a definite a-priori prediction how the different estimators will perform. However, since the selection effects on observables prevail the participation decision, we expect estimators which take account of this issue to perform best. Again this expectations are confirmed by the experiment. The cross-section estimator and the matching estimator perform best, followed by the conditional DID and the DID. The BAE is the worst estimator as it neither accounts for selection effects nor Ashenfelter's dip or the changing economic conditions.

**Scenario F:** In this scenario the selection on unobservables dominates. The results show that the conditional and the unconditional DID perform best under these conditions, since they are the only ones that take this effect into account. The matching and the cross-section estimators as well as the before-after estimator are not able to come close to the true treatment effect.[21]

To conclude this section we give an overall assessment of each estimator. Starting with the before-after estimator, it is the one that performs worst in nearly all the scenarios. It cannot take into account changing economic conditions, Ashenfelter's dip or selection effects. It is, however, also the estimator with the lowest data requirements, since it only requires information about the participants in the post-treatment period and their employment status in the pre-treatment period. This information requirements do not impose any problems and this might be one of the reasons why the BAE is still widely used.

All the other estimators require additional information about non-participants. In this group the cross-section estimator is the least demanding, since it requires only information about one (post-treatment) period for participants and non-participants. Nevertheless it can take into account several effects. Changing economic conditions or Ashenfelter's dip

[21] The obvious bimodal distribution of the CSE and the DID stems from the fact that they condition on different values for X.

do not impose problems for the CSE and also selection on observable characteristics can be captured, as long as the selection is not caused by pre-treatment information, an assumption which seems somehow unrealistic, however.

The latter imposes no problem for the matching estimator, which is able to take selection effects on pre-treatment observable characteristics into account. Therefore it performs slightly better as the CSE in some scenarios. This advantage comes with a higher data requirement as information for at least two periods (one before and one after the treatment took place) for the group of participants and non-participants is needed.

The DID estimator can be seen as an extension to the BAE by taking into account time-invariant linear selection effects on unobservable characteristics. If the unobservables dominate the participation decision, the DID estimator performs good. However, it neglects selection effects on observables. This is tackled by the conditional DID which combines the matching and the DID approach. This estimator performs well in nearly all of the scenarios, even though it is not always the best. This performance comes with a high data requirement, which is comparable to the matching estimator and additional with a larger variability of the estimates.

*< to be completed >*

### 3.4. The Ideal Evaluation Dataset

*< to be completed >*

## 4. Conclusion

After presenting the methodological aspects of different evaluation strategies, we conducted a Monte Carlo study that was general enough to take account of all effects that might influence the various estimators. Selection on observable and unobservable characteristics has been modeled as well as changing economic conditions and Ashenfelter's dip. Therefore we were able to show how the different estimators react to violations of their underlying identifying assumptions. We have seen that the estimators demanding most data, i.e. the matching and the conditional DID estimator, perform well in nearly all of the scenarios. However, there is a trade-off between the data requirement and the performance.

Clearly for scientific research it would be desirable if future datasets would contain as much information as possible not only for the participants but also for the non-participants. The datasets should be rich enough to allow for the controlling of selection on observable characteristics and have a longitudinal dimension to allow a better matching quality.

Our aim for further studies is first of all to include different matching strategies in our evaluation process. Concerning the design, it might be worth to introduce more explanatory variables and allow them to change over time. Doing so would undermine the advantages of the matching and the conditional DID estimator. And furthermore sensitivity tests should be introduced, allowing to show how the estimators react on changes in the selection process. Finally, the data requirements should be described more explicitly to give clear advice for the creation of datasets.

# References


Ashenfelter, O. (1978): "Estimating the Effects of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47–57.

Ashenfelter, O., and D. Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programmes," *Review of Economics and Statistics*, 66, 648–660.

Augurzky, B. (2000): "Optimal Full Matching - An Application Using the NLSY79," Discussion Paper No.310, University of Heidelberg, Department of Economics.

Bergemann, A., B. Fitzenberger, B. Schultz, and S. Speckesser (2000): "Multiple Active Labor Market Policy Participation in East Germany: An Assesment of Outcomes," Working Paper, Institute for Economic Research Halle, University of Mannheim.

Bijwaard, G., and G. Ridder (2000): "Correcting for Selective Compliance in a Re-employment Bonus Experiment," Working Paper, John Hopkins University, Baltimore.

Burtless, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 63–84.

Burtless, G., and L. Orr (1986): "Are Classical Experiments Needed for Manpower Policy?," *The Journal of Human Resources*, 21, 606–640.

Cochrane, W., and D. Rubin (1973): "Controlling Bias in Observational Studies," *Sankyha*, 35, 417–446.

DiNardo, J., N. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001–1045.

Dolton, P., and D. O´Neill (1996): "Unemployment Duration and the Restart Effect: Some Experimental Evidence," *Economic Journal*, 106, 387–400.

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.

Hagen, T., and V. Steiner (2000): *Von der Finanzierung der Arbeitslosigkeit zur Förderung von Arbeit - Analysen und Empfehlungen zur Arbeitsmarktpolitik in Deutschland*. Nomos Verlagsgesellschaft, Baden-Baden.

Heckman, J., and J. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–880.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1996): "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy od Sciences*, 93, 13416–13420.

——— (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

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

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Heckman, J., R. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics Vol.III*, ed. by O. Ashenfelter, and D. Card, pp. 1865–2097. Elsevier, Amsterdam.

Heckman, J., and R. Robb (1985): "Alternative Models for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. Heckman, and B. Singer, pp. 156–245. Cambridge University Press.

Heckman, J., and J. Smith (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85–110.

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

Hujer, R., and M. Caliendo (2000): "Evaluation of Active Labour Market Policy: Methodological Concepts and Empirical Estimates," Discussion Paper No.236, IZA, Bonn.

Hujer, R., and M. Wellner (2000): "The Effects of Public Sector Sponsored Training on Individual Employment Performance in East Germany," Discussion Paper No.141, IZA.

Kluve, J., H. Lehmann, and C. Schmidt (1999): "Active Labour Market Policies in Poland: Human Capital Enhancement, Stigmatization, or Benefit Churning?," *Journal of Comparative Economics*, 27, 61–89.

Koenker, R., and Y. Bilias (2000): "Quantile Regression for Duration Data:A Reappraisal of the Pennsylvania Reemployment Bonus Experiments," Working Paper, University of Illinois.

LaLonde, R. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 76, 604–620.

——— (1995): "The Promise of Public-Sector Sponsored Training Programs," *Journal of Economic Perspectives*, 9, 149–168.

Lechner, M. (1998): "Mikroökonometrische Evaluationsstudien: Anmerkungen zu Theorie und Praxis," in *Qualifikation, Weiterbildung und Arbeitsmarkterfolg. ZEW-Wirtschaftsanalysen Band 31*, ed. by F. Pfeiffer, and W. Pohlmeier. Nomos-Verlag.

——— (2000): "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany," *Journal of Human Resources*, Spring, 347–375.

Mooney, C. (1997): *Monte Carlo Simulation*. Sage Publications, Thousand Oaks, London, New Dehli.

OECD (1991): *Employment Outlook*. Paris.

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

——— (1985a): "The Bias due to Incomplete Matching," *Bioemetrics*, 41, 103–116.

——— (1985b): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistican*, 39, 33–38.

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.

——— (1977): "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Studies*, 2, 1–26.

——— (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.

——— (1980): "Comment on Basu, D. - Randomization Analysis of Experimental Data: The Fisher Randomization Test," *Journal of the American Statistical Association*, 75, 591–593.

SCHMIDT, C. (1999): "Knowing What Works - The Case for Rigorous Program Evaluation," Discussion Paper No.77, IZA.

SMITH, J. (2000): "Evaluating Active Labour Market Policies: Lessons from North America," *Mitteilungen aus der Arbeitsmarkt und Berufsforschung, Schwerpunktheft: Erfolgskontrolle aktiver Arbeitsmarktpolitik*, 3, 345–356.

SMITH, J., AND P. TODD (2000): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estiamtors," Working Paper.

# A Tables

Table 1: **Parameters of the Different Scenarios**

| Scen. | $\alpha$ | $\alpha_X$ | $\alpha_M$ | $\alpha_Y$ | $\alpha_V$ | $\alpha_D$ | $\alpha_{dip}$ | $\beta$ | $\beta_X$ | $\beta_M$ | $\beta_Y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.7 | 0.2 | 0 | 0 | 0 | 0.05 | 0 | 0.5 | 0 | 0 | 0 |
| B | 0.7 | 0.2 | 0 | 0 | -0.04 | 0.05 | 0 | 0.5 | 0 | 0 | 0 |
| C | 0.7 | 0.2 | 0 | 0 | 0 | 0.05 | -0.02 | 0.5 | 0 | 0 | 0 |
| D | 0.5 | 0.2 | 0.2 | 0 | 0 | 0.05 | 0 | 0.5 | 0 | 0.2 | 0 |
| E | 0.15 | 0.25 | 0.15 | 0.3 | -0.1 | 0.05 | -0.01 | 0.05 | 0.4 | 0.05 | 0.05 |
| F | 0.15 | 0.25 | 0.15 | 0.3 | -0.1 | 0.05 | -0.01 | 0.05 | 0.05 | 0.4 | 0.05 |

Table 2: **Descriptions of the Different Scenarios**

| Scen. | Description |
|---|---|
| A | Basis scenario, no selection effects for the participation decision, high skilled employees have a higher employment probability while motivation plays no role, uniform treatment effect |
| B | Like A, except for economic downturn in $t = 4$ |
| C | Like A, except for Ashenfelter´s dip |
| D | Like A, except for selection on unobservables |
| E | High skilled and high motivated have a higher employment probability, persistence effects in the employment status, economic downswing and Ashenfelter´s dip, selection effect on observable characteristics prevails, uniform treatment effect |
| F | Like E, except for selection effect on unobservable characteristics prevails |

Table 3: **The Implemented Estimators**

| Estimator | Formula |
|---|---|
| $\Delta^{TTE}$ | $E[(Y_t^T \mid D = 1) - (Y_t^C \mid D = 1)]$ |
| $\Delta^{BAE\|X}$ | $E[(Y_t^T \mid X, D = 1) - (Y_{t'}^T \mid X, D = 1)]$ |
| $\Delta^{CSE\|X}$ | $E[(Y_t^T \mid X, D = 1) - (Y_t^C \mid X, D = 0)]$ |
| $\Delta^{MAT}$ | $E[(Y_t^T \mid X, Y_{t-1}, D = 1) - (Y_t^C \mid X, Y_{t-1}, D = 0)$ |
| $\Delta^{DID}$ | $E[Y_t^T - Y_{t'}^C \mid D = 1] - [Y_t^C - Y_{t'}^C \mid D = 0]$ |
| $\Delta^{DID\|X}$ | $E[Y_t^T - Y_{t'}^C \mid X, D = 1] - [Y_t^C - Y_{t'}^C \mid X, D = 0]$ |

Table 4: **Results of the Different Scenarios**

| Estimation Approach | Scen.A | Scen.B | Scen.C | Scen.D | Scen.E | Scen.F |
|---|---|---|---|---|---|---|
| *True Treatment Effect* | 0.0498 | 0.0506 | 0.0507 | 0.0511 | 0.0495 | 0.0507 |
| *Cross-Section Estimator* | 0.0492 | 0.0496 | 0.0503 | 0.0912 | 0.0515 | 0.1362 |
| Standard deviation | (0.0233) | (0.0250) | (0.0222) | (0.0275) | (0.0383) | (0.0328) |
| MSE | 0.0005 | 0.0006 | 0.0005 | 0.0024 | 0.0015 | 0.0084 |
| *Before-After Estimator* | 0.0509 | 0.0107 | 0.0696 | 0.0515 | -0.0128 | -0.0162 |
| Standard deviation | (0.0233) | (0.0231) | (0.0234) | (0.0238) | (0.0320) | (0.031) |
| MSE | 0.0005 | 0.0021 | 0.0009 | 0.0006 | 0.0049 | 0.0054 |
| *Difference-in-Differences* | 0.0500 | 0.0503 | 0.0687 | 0.0493 | 0.0756 | 0.0697 |
| Standard deviation | (0.0334) | (0.0341) | (0.0340) | (0.0393) | (0.0399) | (0.0384) |
| MSE | 0.0011 | 0.0012 | 0.0015 | 0.0015 | 0.0023 | 0.0018 |
| *Conditional DiD* | 0.0499 | 0.0503 | 0.0686 | 0.0493 | 0.0629 | 0.0683 |
| Standard deviation | (0.0333) | (0.0339) | (0.0341) | (0.0393) | (0.0486) | (0.0386) |
| MSE | 0.0011 | 0.0011 | 0.0015 | 0.0015 | 0.0025 | 0.0018 |
| *Exact Matching Estimator* | 0.0492 | 0.0495 | 0.0503 | 0.0894 | 0.0492 | 0.1273 |
| Standard deviation | (0.0233) | (0.0251) | (0.0224) | (0.0277) | (0.0364) | (0.0317) |
| MSE | 0.0005 | 0.0006 | 0.0005 | 0.0022 | 0.0013 | 0.0069 |
| Number of simulations | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| Number of individuals | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |

# B Figures

Figure 3: **Basis Scenario (A)**

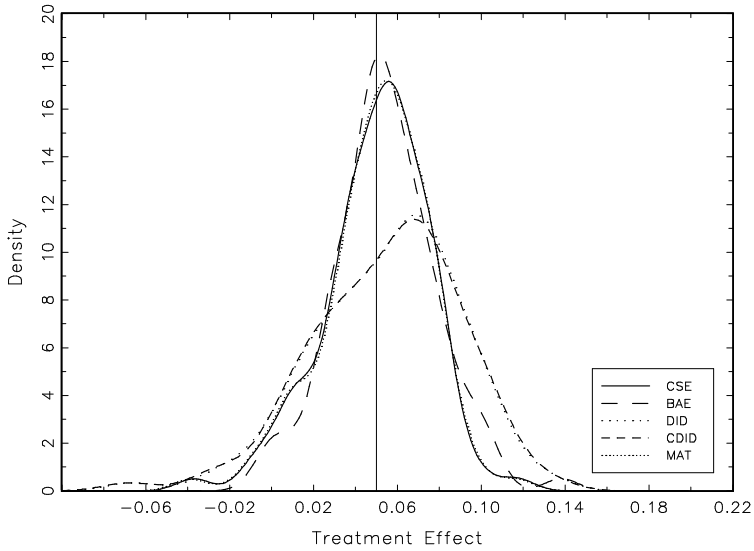



Figure 4: **Economic Downturn (B)**

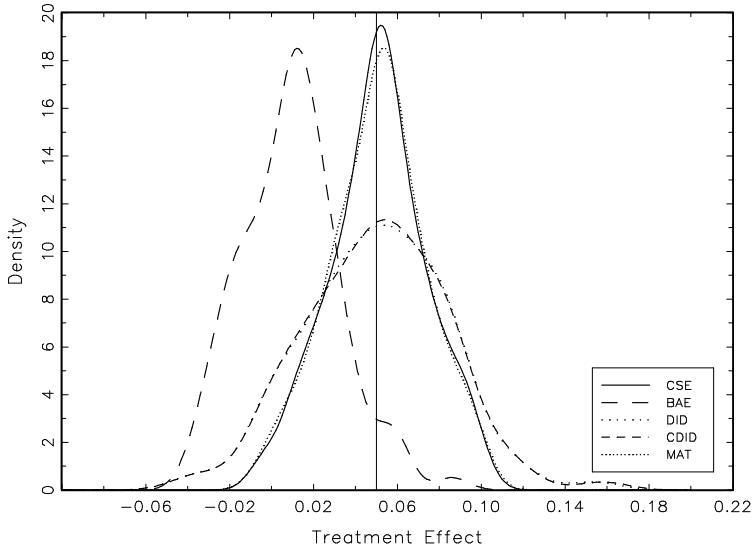


Figure 5: **Ashenfelter's Dip (C)**

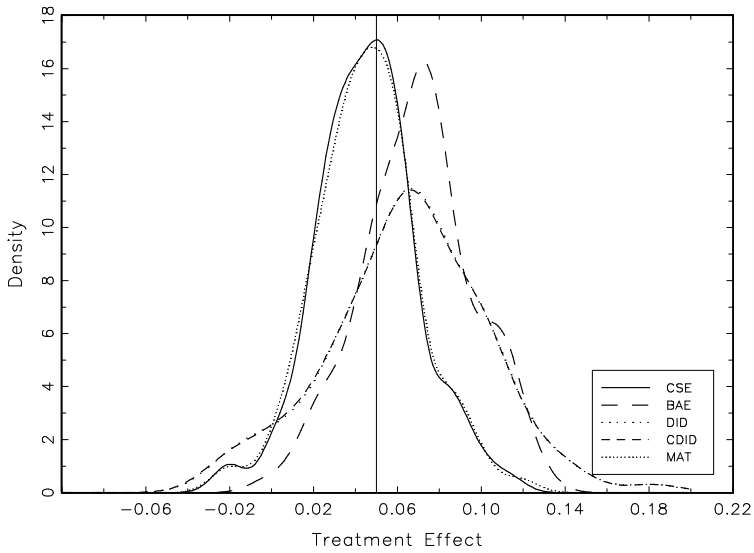



Figure 6: **Selection on Unobservables (D)**

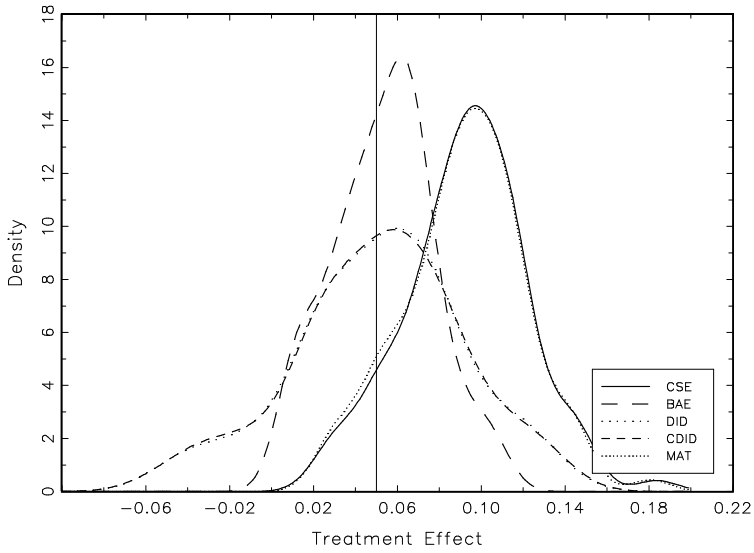


Figure 7: **Mixed Scenario, Selection on Observables Prevails (E)**

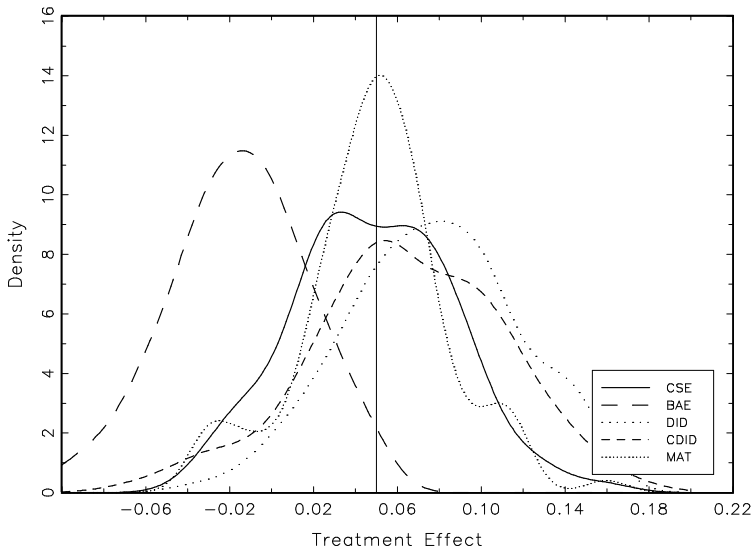
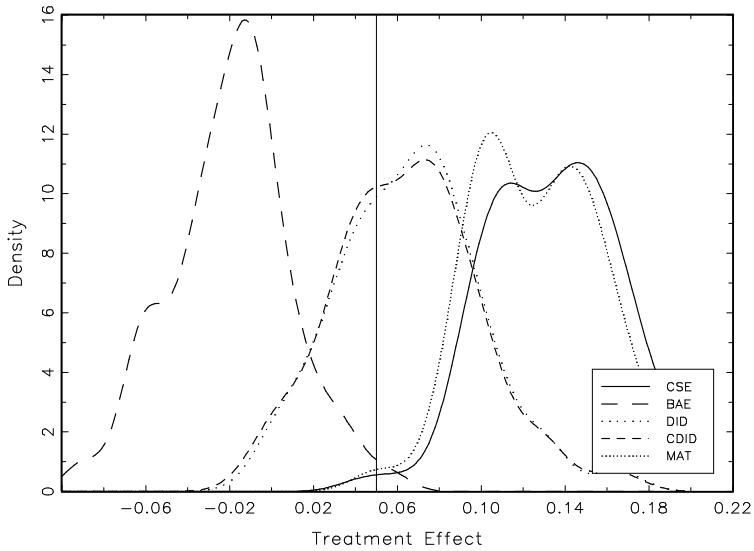



Figure 8: **Mixed Scenario, Selection on Unobservables Prevails (F)**



Kernel matching (KM) is a matching estimator that uses all units in the control group to construct a match for each programme participant. KM defines:

$$W_{N_0}(i,j) = \frac{K_{ij}}{\sum_{k\in\{D=0\}} K_{ik}} \tag{39}$$

where $K_{ik} = K((P_i - P_k)/h)$ is a kernel that downweights distant observations from $P_i$ and $h$ is a bandwidth parameter (Heckman, Ichimura, Smith, and Todd (1998)). Kernel matching can be thought of as running the following weighted least squares regression: $Y_{tj}^C/K_{ij} = \alpha/K_{ij} + u_j/K_{ij}$ with $Y_{tj}^C$ being the outcome of the $j$-th control individuals at time $t$ and $K_{ij}$ as the weights assigned to the various pairs of treatment and control individuals. The estimator of $\hat{\alpha}$ is then used as an estimate for the otherwise unobservable counterfactual outcome for the $i$-th individual.

A generalized version of KM is local linear (LL) matching, which amounts to running a weighted least squares not only on an intercept term but also a linear term in $K_{ij}$. Its advantage is a faster rate of convergence near boundary points and greater robustness to different data design densities (See Heckman, Ichimura, and Todd (1997)).