A note on the *common support problem* in applied evaluation studies

Michael Lechner^{*}

SIAW

Swiss Institute for International Economics
and Applied Economic Research

University of St.Gallen

First draft: November 2000

Date this version has been printed: 07 December 2000

Preliminary

Comments very welcome

Addresses for correspondence

Michael Lechner

Professor of Econometrics

Swiss Institute for International Economics and Applied Economic Research (SIAW)

University of St. Gallen

Dufourstr. 48, CH-9000 St. Gallen, Switzerland

Michael.Lechner@unisg.ch, www.siaw.unisg.ch/lechner, www.siaw.unisg.edu/lechner

^{*} Financial support from the Swiss National Science Foundation (projects 4043-058311 and 4045-050673) is gratefully acknowledged. I am also affiliated with CEPR, London, ZEW, Mannheim, and IZA, Bonn. I thank Patrick Puhani and particularly Markus Frölich for carefully reading the manuscript and thoughtful comments and suggestions.

Abstract

This paper advocates the use of a nonparametric bounds analysis to check the robustness of the results of applied evaluation studies to the problem of a lack of common support. The typical responses by researchers of either ignoring it, or obtaining estimates only for the subpopulation within the common support, can both be misleading: Ignoring the problem may result in biases because the comparison group may not be comparable. Deleting observations at best yields an estimator that is consistent for the common support. When treatment effects are heterogeneous inside and outside the common support this estimator does is inconsistent. Furthermore, useful information is ignored, because the response of the treated to the treatment can be estimated even outside the common support. This information can be used to derive bounds with width depending on the configuration of the data. The application to an evaluation study of Swiss active labour market policies shows that the relevance of the bounds for changing the interpretation of the results depends very much on the particular data configuration.

Keywords: Bounds, matching, propensity score, treatment effects.

JEL classification: C10, C50, J60, J68.

1 Introduction

Microeconomic evaluation studies, or indeed any econometric study attempting to draw causal inferences, try to answer the question what would happen if observations in one state of the world would be observed in another state instead. Under suitable identification conditions,¹ estimators explicitly or implicitly compare observations observed in the particular state with 'otherwise comparable' observations in the other state. Obviously, such comparisons become problematic, i.e. they require more assumptions, if the probability of observing the 'otherwise comparable' observations in the respective other state is zero. This problem is called the failure of the *common support condition*.

In the field of evaluating the effects of training programmes recent papers by Heckman, Ichimura, Smith, and Todd (1996) and Heckman, Ichimura, and Todd (1997) pointed out that failure to take account of that problem might result in a substantial bias of the so-called matching estimator. The problem is particularly apparent for the matching estimator. Matching is a non-parametric estimator that equates the distribution of selected observable variables for observations in the comparison state to the respective distribution of individuals in the original state (also called treatment state in the following). The simplest type of matching is nearest neighbour matching, which allocates to each observation in the treatment state one observation in the comparison state that has (almost) the same value of selected characteristics. Obviously, if such an observation does not exist, matching could be biased, because a comparison observation would be matched that is not sufficiently similar to the treatment observation it is matched to. A typically used 'quick fix' for that problem is to redefine the effect, i.e. to estimate the effect only for the part where comparison observations are observed.² In many cases this redefinition is however not satisfactory because a parameter that may be of reduced interest is estimated. It is worth mentioning that although this problem comes quite naturally for matching, and has thus attracted most attention for that estimator, it is a potential problem for all approaches relying on non-parametric identification. For example, in instrumental variable estimation, the local average treatment effect (LATE) is derived by comparing different individuals with well-defined differences in observable characteristics (matching compares them with the same characteristics).

¹ See for example Rubin (1974), Heckman and Robb (1986), Imbens and Angrist (1994) and the survey of Heckman, Lalonde, and Smith (1999).

Again, if such individuals cannot appear in the sample because of the violation of the support condition, LATE is not identified. For parametric approaches, the problem is less severe, because the parametric model can be used to predict the expected outcome even in regions of the variable space where no observation can occur. Obviously, one might worry about the ability of any model to predict outside its support.

This paper looks at this problem in more detail and discusses some conditions for the identification of the effects. However, since these conditions do not appear to be attractive from an applied point of view, nonparametric bounds are provided for cases when these conditions do not hold. The bounds are in the spirit of bounds first introduced by Manski (1989, 1990) and Robins (1989) for selection problems and treatment evaluation.³ To focus ideas we concentrate on the set of assumptions entertained in the matching literature, but they could be extended to other identification and estimation strategies as well.

Finally, we show again that the problem of common support could be an important one in practical evaluations by using the suggested approaches to re-examine (a part of) a large Swiss evaluation study by Gerfin and Lechner (2000). This application also points to the value of the additional information that could be extracted from the proposed bounds as a sensitivity check.

The plan of the paper is as follows: The next section discusses identification problems resulting from the lack of common support and provides nonparametric bounds for sensitivity analysis. It also specialises the results to a matching framework. Section 3 applies the bounds and Section 4 concludes. The appendix provides more information on the application.

2 The identification problem

2.1 General case

To focus ideas we concentrate on a binary treatment.⁴ In the following standard notation for that case is introduced. Let Y^1 and Y^0 denote the potential outcomes (1 denotes treatment, 0 non-treatment) assumed to have finite first moments. For participants in the treatment the actual (ob-

² More precisely, where the current sample seems to suggest that observations could be observed.

³ For applications see for example Lechner (1999a).

⁴ The extension to the case of multiple treatments as considered by Imbens (1999) and Lechner (1999c) is conceptionally direct, but would increase the notational burden considerably.

servable) outcome is Y^1 , and it is Y^0 for the non-participants. As a notational convention, capital letters indicate quantities of the population, whereas small letters represent their respective quantities in the sample of size N ($i=1, \dots, N$) or specific values. Additionally, denote variables that are unaffected by treatments⁵ as X . Define a binary assignment indicator S , that determines whether the unit receive the treatment ($S = 1$) or not ($S = 0$). The causal effect is defined as the difference of the two potential outcomes. It cannot be estimated, because the respective counterfactual (Y^1 or Y^0) to the observable outcome ($Y = SY^1 + (1 - S)Y^0$) is unobservable. However, under certain assumptions average causal effects are identified.

To be more specific define the population of interest as denoted by the set Ω , which is some subset from the space defined by participation status ($S = 1$ or $S = 0$) and exogenous characteristics (χ). Let W be a binary random variable equal to one if an observation belongs to Ω . It will be Ω for which identification of the effect is desired for:

$$\theta_0(\Omega) = E[(Y^1 - Y^0) | W = 1]. \quad (1)$$

This definition includes usually considered effects like the average treatment effect on the treated (ATET: $\Omega^{ATE} = \{S = 1\} \times \chi$) as well as the average treatment effect on the population (ATE: $\Omega^{ATE} = \{\chi\}$).

Now consider a partition of Ω into two subsets, one for which the effect of the treatment is identified, denoted as Ω^* with corresponding indicator random variable W^* , and a complementary set for which the effect is not identified, denoted as $\tilde{\Omega}$:⁶⁷

$$\theta_0(\Omega) = \theta_0(\Omega^*) P(W^* = 1 | W = 1) + \theta_0(\tilde{\Omega}) [1 - P(W^* = 1 | W = 1)];$$

$$\theta_0(\Omega^*) = E[(Y^1 - Y^0) | W^* = 1, W = 1]; \quad \theta_0(\tilde{\Omega}) = E[(Y^1 - Y^0) | W^* = 0, W = 1]. \quad (2)$$

⁵ The terms treatment and programme are used as synonyms in the remainder of the paper.

⁶ Strictly speaking the following definitions, etc. can be done without the introducing of indicator variables. However, their use allows to stick to a notation more closely related to what is commonly used in that literature.

⁷ For example $\tilde{\Omega}$ may contain individuals who are not allowed to participate or alternatively who must participate in a specific training programme. Without additional assumptions that allows to extrapolate from the results of people with other characteristics not subject to such restrictive selection regime, the effect of such a programme is not be identified for that group.

$\theta_0(\Omega^*)$ is the parameter estimated when the sample is reduced to the subsample for which identification is achieved. Next, it is shown how far $\theta_0(\Omega^*)$ can diverge from the parameter of interest $\theta_0(\Omega)$. ASSUMPTION 1 states formally the information assumed to be available for that result.

Suppose the large random sample contains independent realisations of y_i, x_i, w_i, w_i^* , thus $P(W^* = 1 | W = 1)$ is identified. Furthermore, assume that $\theta_0(\Omega^*)$ is identified. Hence, $P(W^* = 1 | W = 1)$ and $\theta_0(\Omega^*)$ can be treated as known in the identification analysis. Assume also that the potential outcomes are bounded in the set $\tilde{\Omega}$.⁸ For example, in the case when the outcome variable is a binary indicator variable $\underline{Y}$ equals zero and $\bar{Y}$ equals one. These conditions are summarised in ASSUMPTION 1:

Assumption 1

- a) $P(W^* = 1 | W = 1)$ and $\theta_0(\Omega^*)$ are identified.
- b) Y^1 and Y^0 are bounded in $\tilde{\Omega}$: $P(\underline{Y} \leq Y^1 \leq \bar{Y} | W^* = 0, W = 1) = 1$,
 $P(\underline{Y} \leq Y^0 \leq \bar{Y} | W^* = 0, W = 1) = 1$.

Based on ASSUMPTION 1, the following upper bounds $\bar{\theta}_0(\Omega)$ and lower bounds $\underline{\theta}_0(\Omega)$ on the treatment effect $\theta_0(\Omega)$ are identified.

Bounds I (Minimum information)

$$\theta_0(\Omega) \in [\underline{\theta}_0(\Omega), \bar{\theta}_0(\Omega)];$$

$$\bar{\theta}_0(\Omega) = \theta_0(\Omega^*) P(W^* = 1 | W = 1) + (\bar{Y} - \underline{Y}) [1 - P(W^* = 1 | W = 1)];$$

$$\underline{\theta}_0(\Omega) = \theta_0(\Omega^*) P(W^* = 1 | W = 1) + (\underline{Y} - \bar{Y}) [1 - P(W^* = 1 | W = 1)].$$

⁸ Again for notational convenience (and practical relevance), the bounds for the potential outcomes are assumed to be the same for both potential outcomes.

Note that the effect of any individual treatment cannot be better than $(\bar{Y} - \underline{Y})$ and not be worse than $(\underline{Y} - \bar{Y})$. To derive the (worst case) bounds given in BOUNDS I, the possibilities of extreme effects in the set $\tilde{\Omega}$ have to be weighted by the probability of occurrence of that set. The width of these bounds is therefore $2(\bar{Y} - \underline{Y})[1 - P(W^* = 1 | W = 1)]$. They depend on the width of the bounds of the outcome variables as well as on the share of the relevant population in the sample with the support problem.

For many cases these bounds are unnecessarily conservative, because more information about the distribution of either Y^1 or Y^0 in $\tilde{\Omega}$ can be retrieved from the data without further assumptions. However, the type of information depends on the structure Ω and Ω^* .

2.2 Average treatment effect on the treated

In many applications it is indeed very likely that the data contains a random sample of participants in the treatment. In that case the bounds can be sharpened. The sharper bounds are derived for a special case, namely the average treatment effect on the treated (ATET). Estimation of this parameter is the goal of many econometric evaluation studies.⁹

Redefine the set Ω as $\Omega^{ATET} = \{(S = 1) \cap \chi\}$ with corresponding indicator variable W^{ATET} . Note that the sets Ω^{ATET*} (W^{ATET*}) and the complement set $\tilde{\Omega}^{ATET}$ differ only with respect to χ ($\Omega^{ATET*} = \{(S = 1) \cap \chi^*\}$, $\tilde{\Omega}^{ATET} = \{(S = 1) \cap \tilde{\chi}\}$). Furthermore, assume that the data contains an identifiable random sample from $\tilde{\Omega}^{ATET}$ (i.e. a random sample of participants), so that $E(Y^1 | W^{ATET*} = 0, W^{ATET} = 1)$ is identified (because y_i is identical to y_i^1 in the subset of participants). Note that the total effect is not identified because nothing is known about the identification of $E(Y^0 | W^{ATET*} = 0, W^{ATET} = 1)$ that would be the remaining piece for the identification of $\theta_0(\Omega^{ATET})$. ASSUMPTION 2 collects the conditions necessary to derive the sharpened bounds:

⁹ The information contained in ASSUMPTION 2 below allows to sharpen the bounds of other parameters, like the average treatment effect, as well. However, since this will be a straightforward application of the following results, the derivation is not made explicit in this section.

Assumption 2

- a) $P(W^{ATET*} = 1 | W^{ATET} = 1)$ and $\theta_0(\Omega^{ATET*})$ are identified.
- b) Y^0 is bounded in $\tilde{\Omega}^{ATET} : P(\underline{Y} \leq Y^0 \leq \bar{Y} | W^{ATET*} = 0, W^{ATET} = 1) = 1$.
- c) $\lambda_0^1 = E(Y^1 | W^{ATET*} = 0, W^{ATET} = 1)$ is identified.

With the exception of having to consider the more restricted set related to ATET, part a) is identical to ASSUMPTION 1. Part b) is weaker, because it concerns only potential outcome Y^0 . Part c) formalises the fact that the random sample of participants allows to identify the mean of Y^1 also in the subpopulation without support.

Using ASSUMPTION 2 the bounds given in BOUNDS I shrink:

Bounds II (ATET with random sample of participants)

$$\theta_0(\Omega^{ATET}) \in [\underline{\theta}_0(\Omega^{ATET}), \bar{\theta}_0(\Omega^{ATET})];$$

$$\bar{\theta}_0(\Omega^{ATET}) = \theta_0(\Omega^{ATET*}) P(W^{ATET*} = 1 | W^{ATET} = 1) + (\lambda_0^1 - \underline{Y}) [1 - P(W^{ATET*} = 1 | W^{ATET} = 1)];$$

$$\underline{\theta}_0(\Omega^{ATET}) = \theta_0(\Omega^{ATET*}) P(W^{ATET*} = 1 | W^{ATET} = 1) + (\lambda_0^1 - \bar{Y}) [1 - P(W^{ATET*} = 1 | W^{ATET} = 1)].$$

The width of bounds given by BOUNDS II is reduced to $(\bar{Y} - \underline{Y}) [1 - P(W^{ATET*} = 1 | W^{ATET} = 1)]$. This is not directly comparable to the bounds given by BOUNDS I because of the different conditioning sets for which the parameter is defined ($\theta_0(\Omega^{ATET})$ versus $\theta_0(\Omega)$). However, in cases when $P(W^* = 1 | W = 1)$ and $P(W^{ATET*} = 1 | W^{ATET} = 1)$ are equal, the width of the bounds implied by ASSUMPTION 2 is half the width implied by ASSUMPTION 1.

2.3 Matching

In this section the results are further specialised by considering the case where matching can be used as a consistent estimator of the ATET. For ATET the difficulty of identification comes from the term $E(Y^0 | S = 1)$, because the pair $(y_i^0, s_i = 1)$ is not observable. Matching is a valid estimator for ATET when the assignment S is independent (denoted by $\perp$ in the following) of Y^0 conditional on a set of covariates (Rubin, 1977). This is also called the conditional independence assumption (CIA):¹⁰

$$Y^0 \perp S | X = x, \forall x \in \mathcal{X}. \quad (3)$$

If CIA holds, then $E(Y^0 | S = 1, X = x) = E(Y^0 | S = 0, X = x)$ so that $E(Y^0 | S = 1, W^{ATET} = 1) [= E(Y^0 | W^{ATET} = 1)] = E[E(Y^0 | S = 0, X = x) | W^{ATET} = 1]$ is identified, provided the support condition $P^1(x) < 1$ holds for $\forall x \in \mathcal{X}$.¹¹ $P^1(x)$ denotes the propensity score $[P^1(S = 1 | X = x)]$ defined as the participation probability conditional on X . The condition $P^1(x) < 1$ means that nonparticipants must be observable for all values of x where participants are observable, i.e. for which ATET is desired. This condition breaks down for example when 'perfect predictors' of the treatment status are necessary to make condition (3) hold in any application.¹²

In empirical applications many variables to condition may be needed to justify an assumption like the one given in (3), thus making nonparametric estimation difficult and subject to the curse of dimensionality problem. Rosenbaum and Rubin (1983) show that if CIA is valid, then the estimation problem simplifies. They find that if the two treatments are independent of the assignment conditional on X , then they are also independent conditional on the propensity score:

$$Y^0 \perp S | P^1(x), \forall x \in \mathcal{X} \quad (4)$$

¹⁰ Note that CIA is more restrictive than necessary, because all what is needed to identify mean effects is conditional mean independence. However, the former has the virtue of making the latter valid for all transformations of the outcome variables. Furthermore, in an application it is usually difficult to argue why conditional mean independence should hold and CIA might nevertheless be violated.

¹¹ Note that the condition $0 < P^1(x)$ is not relevant because the parameter of interest is ATET which is defined for the population with positive participation probability (i.e. the treated).

¹² Obviously, for $P^1(x) = 1$ a degenerated version of condition (3) is created.

If $P^1(x)$ is smaller than one, then the equality $E[Y^0 | S=1, P^1(X)=P^1(x)] = E[Y^0 | S=0, P^1(X)=P^1(x)]$ can be used for identification. The major advantage of this property is the reduction of the dimension of the estimation problem.¹³

Redefining the previous definitions of the sets to take account of the special structure of matching, we get: $\tilde{\Omega}^{ATET} = \{(S=1) \cap \{x | x \in \mathcal{X}, P(x)=1\}\}$. Thus $\theta_0(\Omega^{ATET*})$ is defined over the set that fulfils the support condition, so $\theta_0(\Omega^{ATET*})$ is identified by CIA. Furthermore, λ_0^1 is identified as well in this case, even if the support condition is violated, because the data contains a random sample of participants. Therefore, ASSUMPTION 2 is applicable and the bounds are given by BOUNDS II. In BOUNDS II $\theta_0(\Omega^{ATET*})$ is the effect ignoring participants without a comparable match. The term $P(W^{ATET*}=1 | W^{ATET}=1)$ denotes the share of such participants relative to the total number of participants.

So far only regions with propensity score equal to one ($P^1(x)=1$) are discussed. The reason is that *holes* in the support, i.e. regions corresponding to values of the propensity score which cannot be attained with the X -process generating the sample, are no problem when estimating the ATET. Whenever there is a realisation of the treated with a probability of less than one, then, if the sample is large enough, there is also a realisation of a non-treated for that particular value of X . If however there cannot be any realisation of the treated, than that particular configuration of X has zero weight anyhow in the computation of the ATET and the non-treated are irrelevant. Hence, the only problem for identification here a regions that have probability one for the treated.

Furthermore, a note on selecting variables relevant in applications is in order. Conditions (3) and (4) require the researcher to find all variables that are jointly correlated with the potential non-treatment outcome and with participation. After having found these variables, the support problem could appear conditional on the specific choice of variables. Obviously, from the point of view of the support condition it could actually be dangerous to include variables that are good predictors of participation but independent of outcomes (these variables could be called instruments, because they satisfy a conditional exclusion restriction with respect to the conditional means of the outcomes). If these variables predict *too good*, i.e. if they are perfect predictors conditional on the other variables, then they could lead to a partial break-down of

¹³ Efficiency issues involved by conditioning on the propensity score instead of X are discussed in detail by Hahn

identification. They would destroy the very randomness that is required for the identification of ATET by matching.¹⁴

Finally consider three types of information that would each allow the identification of $\theta_0(\Omega^{ATET})$:

a) The average effect in the subset with and without support is the same $\theta_0(\Omega^{ATET}) = \theta_0(\Omega^{ATET})$.*

In this case the strategy to concentrate only on the subpopulation with support (Ω^{ATET*}) would identify the effects ($\theta_0(\Omega^{ATET})$).

b) The conditional expectation $E(Y^0 | S = 0, X = x)$ ($= E(Y^0 | S = 1, X = x)$) is the same in the region without support as in the region with support close to the region without support (i.e. for $P^1(x)$ close to one).

The strategy to match comparison observations with the highest value of $P^1(x)$ to treated observations with $P^1(x) = 1$ identifies the effect in that case.

c) There is a parametric model for $E(Y^0 | S = 0, X = x)$ that could be estimated within the support and is able to predict $E(Y^0 | S = 0, X = x)$ in regions without support.

Obviously, in this case the drawback is that the advantage of CIA, namely nonparametric identification, is lost even for the region where it is still feasible.

For particular evaluation studies there might be rare cases where one of those three conditions is plausible and thus can be used to identify the effects.¹⁵ However, in general they would certainly be hard to justify and thus they are no substitute for the bounds analysis.

3 Application

3.1 Introduction

In this section the above considerations are applied to an empirical study about the major programmes of the active labour market policy in Switzerland by Gerfin and Lechner (2000).

(1998), Heckman, Ichimura, and Todd (1998), and Hirano, Imbens, and Ridder (2000).

¹⁴ For more potential problems created by using instruments as matching variables, see Lechner and Smith (2001).

Gerfin and Lechner (2000) focus on the individual success in the labour market that is due to these programmes. The Swiss government made available a novel, very informative and large data base consisting of administrative records from the unemployment insurance system as well as from the social security system. It covers the population of unemployed persons in December 1997. Gerfin and Lechner (2000) claim that in this data all major factors that jointly influence both the selection into the various programmes as well as employment outcomes are observed.

The data from the unemployment registrars cover the period January 1996 to March 1999 for all persons who were registered as unemployed on December 31, 1997. These data provide very detailed information about the unemployment history, ALMP participation and personal characteristics. The social security records cover 1988-1997 for a random subsample of about 25'000 observations.¹⁶ They describe sociodemographics (age, gender, marital status, native language, nationality, type of work permit, language skills); region (town/village and labour office in charge); subjective valuations of case workers (qualifications, chances to find a job), sanctions imposed by the placement office; previous job and desired job (occupation, sector, position, earnings, full- / part-time); a short history of labour market status on a daily basis, and the employment status and earnings on a monthly basis for the last ten years. Gerfin and Lechner (2000) apply a series of sample selection rules to the data. The most important one is to consider only individuals unemployed on Dec 31, 1997 with an unemployment spell of less than 12 months who have not participated in any major programme in 1997 and who are between 25 and 55 years old.

The active labour market programmes (ALMP) can be grouped into 3 broad categories: a) training courses, b) employment programmes, and c) temporary employment with wage subsidy (TEMPORARY WAGE SUBSIDY). The former two groups are fairly standard encompassing a variety of different programmes. The last type of programme is rather unique, however. The difference between b) and c) is that employment programmes take place outside the “regular” labour market. By contrast TEMPORARY WAGE SUBSIDY refers to a regular job.

For this illustration the focus is only on a pair-wise comparison of two specific programmes, namely LANGUAGE COURSES (LAC) and TEMPORARY WAGE SUBSIDY (TEMP). Note that the validity of CIA allows us to analyse the effects of these programmes on the subsample of partici-

¹⁵ Note that obviously these conditions can be combined for example by assuming a parametric model to be valid only the neighbourhood of the no-support region.

¹⁶ The variables and corresponding descriptive statistics can be obtained from Gerfin and Lechner (2000), which is available on the internet (www.siaw.unisg.ch/lechner).

pants in the respective programmes without incurring any selectivity bias problem from ignoring individuals in other programmes not considered here (see Lechner, 1999c).

Table 1: Number of observations and selected characteristics

Group		obs.	duration of programme	unem- ployment before...	qualifi- cation	foreign nationality	employed 1 year after begin
		(persons)	(mean days)	(mean days)	(mean)	(share in %)	(share in %)
LANGUAGE COURSE	(LAC)	1719	71	225	2.2	72	30
TEMPORARY WAGE SUBSIDY	(TEMP)	4390	114	228	1.7	46	52

Note: Qualification is measured as skilled (1), semiskilled (2), and unskilled (3).

Table 1 shows the number of observations as well as some descriptive statistics for the subsample composed of participants in LAC and TEMP. The mean duration of the programmes is a little more than two months for LAC and almost 4 months for TEMP. The table shows that important variables like qualification, duration of unemployment, and particularly nationality, also vary substantially.¹⁷ In particular these differences led Gerfin and Lechner (2000) to argue that their estimated effects for LAN courses are subject to additional uncertainty, because this group appears to be very specific. The final column shows the employment rate one year after the individual begin of the programme. It shows a marked advantage of 52% for TEMP compared to 30% for LAN. However, it is already obvious from the few characteristics shown in Table 1 that this is not indicative for programme success, because participants in TEMP are expected to do better in the labour market anyway.

3.2 Estimation

In the following the interest is to estimate the effect of TEMP compared to LAC for participants in TEMP. Following the arguments by Gerfin and Lechner (2000) about the information available in the sample leads to the independence assumption between the decision for participating in TEMP compared to LAN and the outcomes (conditional on the covariates shown in Table A.1 below and conditional on being in either of the two states). Hence, a simple binary treatment evaluation strategy is available to estimate ATET. Here we employ a simple nearest neighbour matching estimator. In the first step a binary probit model is estimated to explain the choices

¹⁷ The share of foreigners among Swiss unemployment is about 40%.

between the states. The functional form of the conditional participation probability is checked by standard specification tests as suggested in Lechner (1999b, 2000). In the second step a comparison group to TEMP is formed by nearest neighbour matching. Matching is however not directly on the propensity score, but on its linear index. The latter should have the advantage to generate better matches in regions when probabilities are very close to zero and one. Matching is implemented with replacement (i.e. by allowing to use one comparison observation many times), because the treatment group is much larger than the number of observations in the comparison group.

The results of the binary probit are generally in line with the findings reported by Gerfin and Lechner (2000). It appears again that variables that measure region, language skills and nationality, skills, as well as a couple of other characteristics, play an important role. The detailed results are contained in the Appendix.

3.3 Support

In an empirical analysis of the ATET under CIA the breakdown of the common support requirement can result from two sources:

- a) There are variables that are perfect predictors, i.e. they lead to an estimated participation probability of 1 for some observations.
- b) The estimated values for the propensity score are such that there are no matches in the sample.

Note that the mere fact that there is no match in the sample does not cause a consistency problem, because the fact that probabilities below 1 exist for that particular data configuration suggests that when the sample grows, at some sample size there will be a match and consistency is no problem.

So Case a) seems to be clear. It is not in practise. Suppose for example that the propensity score is such that a parametric model like logit or probit with a linear index structure is estimated. In that case every given set of finite coefficients generates probabilities strictly bounded away from zero and one. One will usually detect binary variables that are perfect predictors because their estimated coefficients converge to infinity during estimation.¹⁸ This is however not the case for

¹⁸ Although some software tries to prevent them to do so – which is in fact a very dangerous thing to do, because it obscures a potential identification problem.

categorical variables if not all (but one) of the categories are perfect predictors (or if in fact interaction terms omitted from the estimation would generate such regions).

What could be the approximate solution to that problem in an empirical study? First, one might want to use a nonparametric estimator for the propensity score or condition directly on the X 's. In that case one needs to get an estimate of the regions with probability one. However, the curse of dimensionality would be prohibitive for high dimensional X . Hence, this is not an option for the current and many other studies.

One of the possible alternatives is to first pre-screen important variables (in combination with institutional knowledge) to spot perfect predictors. Then run the smooth probability model and check the distribution. If there is a lack of overlap appearing in the distribution of the propensity score in the subsample of participants and potential comparisons (it should happen in the upper region, i.e. when $P^1(x)$ is close to one), than one might either ignore this and appeal to asymptotic arguments (and also hope the probability model chosen is the correct one) or state the inability to identify the effects for this group with the sample at hand. The latter is a different argument compared to the types of arguments used in Section 2, because it is not about asymptotic identification. It is more about small sample properties (or a misspecified probability model). The current application follows that line of thinking as do for example Gerfin and Lechner (2000) and Smith and Todd (2001).

Figure 1 shows a histogram of the distribution of the propensity score and its linear index in the two subsamples. Since identification as well as matching is conditional on the propensity score, it is the correct measure to evaluate any support problem. Considering the probabilities first, it appears that whereas they are more or less equally distributed in the LAC subsample, they are heavily concentrated close to one for the TEMP sample (note that to the very right of Figure 1a, the line related to TEMP *leaves* the top of the chart due to the heavy concentration of observations in this region). Considering the index it becomes very clear that there is a considerable number of TEMP observations on the right of Figure 1b with very high scores, for which no comparable LAC observations can be found in this sample. It also becomes clear that there appear to be no *holes* in the support, i.e. zero densities somewhere in the middle of the propensity score. Hence, we conclude that there is a support problem for very high values of the propensity score.

Figure 1: Distribution of the propensity score and its linear index in the two subsamples

Fig. 1a: Propensity score

Fig. 1b: Linear index

Note: Dashed line relates to the TEMP sample. x denotes values for variables included in the probit estimation, $\hat{\beta}_N$ is the respective estimate, and $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution. The distribution is estimated by a histogram estimator based on 50 equally sized bins.

Given that the support problem can be spotted directly by inspection of Figure 1b, there is no need to implement a complicated formal estimator for the support as for example proposed by Heckman, Ichimura, and Todd (1997) and Smith and Todd (2001). Instead a very simple one is used that defines the support as the region where at least c potential comparison observations are available. To avoid mismatches to the very right, we set c to 15, which corresponds to about 0.8% of the observations in the comparison sample.¹⁹ On the other hand, 17% (626) of TEMP participants are located to the right of the 15th largest control, that has a value of the score of 97.3%. Figure 2 gives the distribution of the score and the linear index after removing the 626 TEMP observations.

¹⁹ The results are not sensitive to small variations of c .

Figure 2: Distribution of the propensity score and its linear index in the subsamples after deleting 15 LAC observation with the highest scores

Fig. 2a: Propensity score

Fig. 2b: Linear index

Note: Dashed line relates to the TEMP sample. x denotes values for variables included in the probit estimation, $\hat{\beta}_N$ is the respective estimate, and $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution. The distribution is estimated by a histogram estimator based on 50 equally sized bin's.

Table 2 gives a gives a brief idea of the distribution of characteristics in the subsample of TEMP participants considered to have no support by the above rule. Not surprisingly, this group displays a much higher share of Swiss nationals. Inspecting more differentiated statistics, it appears also that all the deleted foreigners have a permanent work permit for Switzerland, thus they are indeed not very likely to appear in a language course (because this type of permit typically can only be obtained after a stay of more than five years).

Table 2: Number of observations and selected characteristics for TEMP participants

Group	obs.	duration of programme	unem- ployment before...	qualifi- cation	foreign	employed 1 year after begin
	(persons)	(mean days)	(mean days)	(mean)	(share in %)	(share in %)
Full sample	4390	114	228	1.7	46	52
Deleted sample	626	104	235	1.6	22	52
Remaining sample	3764	116	226	1.7	51	51

Note: Qualification is measured as skilled (1), semiskilled (2), and unskilled (3).

It is clear that the refinements of the sample based on the estimated score are indeed part of the estimation process and should be part of the estimated standard errors of the estimator. Here, this

is taken care of by integrating this step, as well as the estimation of the probit, into the bootstrap procedure used to guide the inference below.

Due to the large number of observations in the comparison sample and the use of nearest neighbour matching with replacement the match quality for the part of the sample with common support as defined here is good for all variables that are important factors in determining the participation decision.

3.4 Results

The measure for the success of the programme is employment in the regular labour market at any given time after the beginning of the programme. Hence the outcome variable is binary 0-1. The time during participation in the programme is not considered to be regular employment. Due to data limitations the potential period of observing programme effects cannot be longer than 15 months, because the latest observation dates from March, 31, 1999. In that sense the analysis is restricted to the short run effects.

Table 3 displays the mean effects of TEMP compared to LAC for participants in TEMP 9 months after the individual programme participation starts. A positive number indicates that the effect of TEMP compared to LAC is on average a higher rate of employment for those who participated in TEMP.

Table 3 contains estimates of all elements necessary to compute the parameters discussed so far. First note the raw difference between the employment probability nine months after the beginning of the program is more than 20%-points in favour of TEMP. However, for the population within the common support, this is reduced by matching to 8.7%-points ($\hat{\theta}_N(\Omega^{ATET*})$). Because the outcome variable is binary, the width of the bounds corresponds to the probability of not being in the common support ($\hat{P}_N(W^{ATET*} = 0)$). The latter is about 14% points after 270 days.²⁰ The location of the bounds depends on the mean of Y^1 for participants outside the common support ($\hat{\lambda}_N^1$), as well as on the effect inside the support. Here the bounds range from 0 to 13.6% points. Taken into account that the bounds are estimated with error, it would not be possible to reject the hypothesis that the true effect for all treated of TEMP compared to LAC is

²⁰ Note that not everyone is still observed after 9 months depending on the day of entering the programme. This type of attrition is assumed to be independent of outcomes.

zero. Hence, the additional uncertainty due to the support problem may change the conclusion from the empirical study.

Table 3: Effects for TEMP participants measured as difference in employment rates 9 months after start of TEMP in %-points

	$\hat{E}_N(Y^1 W^{ATET*} S = 1)$	$\hat{E}_N(Y^0 S = 0)$	$\hat{E}_N(Y^0 W^{ATET*} S = 1)$	$\underline{Y}$	$\bar{Y}$
Estimate	47.6	26.4	38.9	0	1
Bootstrap: Mean	47.3	26.4	36.0	0	1
Median	47.2	26.4	35.8	0	1
2.5% quantile	45.6	24.4	29.4	0	1
97.5% quantile	49.1	28.6	43.4	0	1
	$\hat{\theta}_N(\Omega^{ATET*})$	$\hat{\theta}_N(\Omega^{ATET})$	$\hat{\theta}_N(\Omega^{ATET})$	$\hat{P}_N(W^{ATET*} = 1)$	$\hat{\lambda}_N^1$
Estimate	8.7	13.6	0.0	13.6	44.6
Bootstrap: Mean	11.3	17.0	1.0	15.2	46.7
Median	11.4	16.9	1.5	15.2	46.5
2.5% quantile	3.7	-	-5.5	10.3	41.5
97.5% quantile	18.2	22.8	-	20.7	51.1

Note: All quantiles obtained from the bootstrap distribution with 1000 replications.

Figure 3: Average effects for TEMP participants ($\theta_0(\Omega^{ATET})$) measured as difference in employment rates one year after start of the start of programme in %-points*

Note: Results are based on matched samples. Mean and 5% as well as 95% quantiles given. All quantiles based on bootstrap estimates.

Figure 3 shows the dynamic development of the estimated treatment effect $\hat{\theta}_N(\Omega^{ATET*})$ over time, as well as associated 5% and 95% confidence bounds obtained by 1000 bootstrap replications. It appears that TEMP has a significantly positive effect almost everywhere, perhaps with the exception of the very end of the period under consideration.

Figure 4 shows the same dynamic development but includes the bounds as sensitivity check. Again taking into account sampling uncertainty, the conclusions are changed and the hypothesis that the true effect is zero cannot be rejected for any particular point in time.

Figure 4: Bounds for $\theta_0(\Omega^{ATET})$

Note: Results are based on matched samples. Dashed line between the bounds refers to $\hat{\theta}_N(\Omega^{ATET*})$. Mean and 5% quantile of lower bound as well as 95% quantile of upper bound given. All quantiles are based on bootstrap estimates.

Figure 5 gives the bounds for the effects of LAC compared to TEMP for participants in LAC (and not in TEMP as considered before). In this case the width of the bounds is only about 7%-points and the magnitude of $\hat{\theta}_N(\Omega^{ATET*})$ is sufficiently large to not change the conclusions that $\theta_0(\Omega^{ATET})$ is positive.

Figure 5: Bounds for $\theta_0(\Omega^{ATET})$: LAC compared to TEMP for participants in LAC

Note: Results are based on matched samples. Dashed line between the bounds refers to $\hat{\theta}_N(\Omega^{ATET*})$. Mean and 5% quantile of lower bound as well as 95% quantile of upper bound given. All quantiles are based on bootstrap estimates.

In conclusion the bounds analysis appears as a very useful tool to assess the sensitivity of the estimated effects with respect to the common support problem and its impact on the inference drawn from estimates based on regions with common support. Furthermore, if the researcher has strong prior beliefs about the plausible range of the expectation of the nontreatment outcome variable outside the support, then the bounds may appear to be too conservative. In that case, they could be tightened further. Finally, the bounds allow an assessment about what values of the expectation of nontreatment outcomes outside the support are compatible with an positive or negative effect of $\theta_0(\Omega^{ATET})$ of a particular magnitude. Then, the prior beliefs of the researcher about the plausibility of these magnitudes can be used to draw conclusion about likely sensitivity to the common support problem.

6 Conclusion

This paper advocated to use a nonparametric bounds analysis to check the robustness of the results of applied evaluation studies to the problem of a lack of common support. The typical responses to this problem, i.e. either just ignoring it, or obtaining estimates only for the subpopulation within the common support, can both be misleading. First, ignoring the problem may

result in serious biases because a comparison group may not be comparable at all. Deleting the sample is not prone to this criticism. Nevertheless, although the estimate is consistent for the region of common support, it may be problematic as well. When treatment effects are heterogeneous inside and outside the common support, then the estimated effect does no longer correspond to the original parameter of interest. This is particularly damaging when the common support is hard to characterise in economic terms, thus obscuring the substantive meaning of the estimated effects. Furthermore, useful information is ignored by throwing away all observations outside the common support, because the response of the treated to the treatment can be still estimated outside the common support. This information can be used to derive bounds that are more or less narrow depending on the data configuration in the particular application.

In conclusion it is proposed to routinely compute bounds in order to be able to extract all available information about the parameter of interest from the data. At least, this exercise gives clear hints about the sensitivity of conclusions drawn from the (often implicitly) redefined parameter, obtained by deleting the problematic part of the sample, to the common support problem.

References

- Gerfin, M., and M. Lechner (2000): "A Microeconometric Evaluation of the Active Labour Market Policy in Switzerland", discussion paper 2000-10, University of St. Gallen.
- Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66, 315-331.
- Heckman, J. J., and R. Robb (1986): "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes", in H. Wainer (ed.), *Drawing Inferences from Self-selected Samples*, New York: Springer, 1986, 63-107.
- Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program", *The Review of Economic Studies*, 64, 605-654.
- Heckman, J. J., H. Ichimura, and P. E. Todd (1998): "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.
- Heckman, J. J., H. Ichimura, J. Smith, and Petra Todd (1996): "Sources of selection bias in evaluation social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method", *Proceedings of the National Academy of Science, USA*, 13416-13420, November 1996.
- Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999): "The Economics and Econometrics of Active Labour Market Programs", in O. Ashenfelter and D. Card (eds.): *Handbook of Labor Economics*, Vol. III A, chapter 31, 1865-2097.
- Hirano, K., G. W. Imbens, and G. Ridder (2000): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *NBER Technical Working Papers*, 251.
- Imbens, G.W., and J.D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 446-475.
- Lechner, M. (1999a): "Nonparametric Bounds on Employment and Income Effects of Continuous Vocational Training in East Germany," *The Econometrics Journal*, 2, 1-28.

- Lechner, M. (1999b): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification", *Journal of Business & Economic Statistics*, 17, 74-90.
- Lechner, M. (1999c): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption", *Discussion paper 9908, University of St. Gallen*, forthcoming in M. Lechner and F. Pfeiffer (eds., 2001), *Econometric Evaluations of Active Labour Market Policies in Europe*, Physica.
- Lechner, M. (2000), "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany", *The Journal of Human Resources*, 347-375.
- Lechner, M., and J. Smith, (2001): "Some Exogenous Information Should Not Be Used in Evaluation Studies", *mimeo*.
- Manski, C. F. (1989): "The Anatomy of the Selection Problem", *Journal of Human Resources*, 24, 343-360.
- Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects", *American Economic Review, Papers and Proceedings*, 80, 319-323.
- Robins, J. M. (1989): "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies", Sechrest, L., H. Freeman, A. Mulley (eds.), *Health Service Research Methodology: A Focus on Aids*, 113-159, Washington, D.C.: Public Health Service, National Center for Health Services Research.
- Rosenbaum, P.R., and D.B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-50.
- Rubin, D.B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D.B. (1977): "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, 1-26.
- Smith, J., and P. Todd (2001): "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators", *mimeo*.

Appendix: Estimates of the binary probit model

Table A.1: Estimated coefficients of the binary probit model for participation in TEMP vs. LAC

Variable	Coefficients	t-values
Younger than 30	0.12	2.6
Older than 50	0.11	1.5
Female	0.14	1.4
Marital status married	-0.11	1.3
Married women	-0.19	-2.1
Number of persons to support	-0.03	2.0
<i>Mother tongue</i>		
French	0.31	2.6
French in French speaking canton	-0.91	-4.4
Italian	-0.29	2.4
Not German/French/Italian	-1.34	11.4
GF/I, but not canton language	-0.76	-6.4
GF/I, not canton language and married	0.24	2.4
GF/I, not canton language in Tecino	0.70	3.4
<i>Foreign Languages</i>		
Other Swiss language	-0.11	-1.8
English, Spanish, Portuguese	-0.34	-4.4
<i>Unemployment-status</i>		
In part-time employment	0.83	4.5
<i>Nationality</i>		
Foreign with permanent permit	-0.14	-1.5
Foreign with yearly permit	-0.29	-3.1
<i>Monthly earnings in last job</i>		
4000-5000	0.09	1.8
More than 5000	-0.01	-0.3
<i>Chances to find a job</i>		
No information	0.09	1.0
Very easy	0.08	0.8
Easy	0.18	2.9
Difficult	-0.27	-4.8
Special case	-0.33	-2.3
<i>Qualification</i>		
Skilled	0.26	3.0
Unskilled	-0.01	-0.2
Skilled and living in canton Waadt	-0.65	-3.6
Skilled and one of Swiss languages	-0.25	-2.5
Skilled and Italian	-0.44	-2.4
Skilled women	0.24	2.7
Skilled * duration to start * 100	-0.13	-3.0
Unskilled and none of Swiss languages	-0.27	-2.6
Unskilled and French	0.31	3.1
<i>Previous industry sector</i>		
Public services	-0.27	-3.3
machinery and equipment	-0.42	-1.6
electrical machinery/optics,	-0.39	-2.7

Table A.1 to be continued

Table A.1 continued

	Coefficients	t-values
Variable		
<i>Job position function</i>		
Self-employed	-0.66	-2.6
<i>Previous occupation</i>		
Food, Tobacco	0.33	1.5
Textiles	-0.42	2.7
Metals	0.27	3.1
Construction	0.23	2.9
Transportation	0.48	3.9
Restaurants	0.17	2.9
Entrepreneurs, senior officials, justice	-0.53	3.7
Office and computer	-0.25	-2.1
Retail trade	-0.19	-2.1
Security, cleaning, clerical, social work	0.25	2.4
News and communication	-0.61	-2.4
Other	-0.22	-2.1
<i>Region (reference category: Zurich)</i>		
Central	0.23	3.4
South-west	0.84	7.5
North-west	0.43	4.4
<i>Additional regional effects by canton</i>		
Aargau	-0.64	-5.3
Schaffhausen	0.48	2.1
Ticino	0.20	1.0
Neuenburg	0.91	5.2
Geneva	-0.16	-1.3
<i>Size of town where worked before</i>		
>100'000	0.14	2.0
<20'000	0.18	3.1
<i>Region of placement office</i>		
Rural	0.14	2.5
<i>Long-term unemployment in regional placement office</i>		
No information	0.2	3.8
<i>Previous programme participation</i>		
Begin of programme / 100	0.91	11.4
(Begin of programme / 100) squared	-0.21	-8.0
(Begin of programme / 100) * female	-0.16	-3.5
<i>Duration of unemployment spell at beginning of programme</i>		
Duration (days)	-0.09	-0.25
Less than 90 days	0.09	1.3
Less than 270 days	-0.19	-2.4
<i>Employment history from social security data</i>		
Never unemployed before	-0.21	-4.6
Month of entry into social security system	-0.50	-5.2
Number of employment spells	0.03	3.2
Proportion of time employed, in %	0.58	4.4

Note: N=6109. Log of Likelihood at maximum: -2622. t-values based on QML estimates of covariance matrix. If not stated otherwise, all information in the variables relates to the last day of December 1997.