HEDG Health, Econometrics and Data Group
THE UNIVERSITY of York

HEDG Working Paper 07/02

Matching estimators of average treatment effects: a review applied to the evaluation of health care programmes

Rodrigo Moreno-Serra

February 2007

ISSN 1751-1976

york.ac.uk/res/herc/hedgwp

Matching estimators of average treatment effects: a review applied to the evaluation of health care programmes

Rodrigo Moreno-Serra*

Department of Economics and Related Studies & Centre for Health Economics

University of York

February 2007

Abstract

The general aim of this paper is to review how matching methods try to solve the evaluation problem – with a particular focus on propensity score matching – and their usefulness for the particular case of health programme evaluation. The "classical" case of matching estimation with a single discrete treatment is presented as a basis for discussing recent developments concerning the application of matching methods for jointly evaluating the impact of multiple treatments and for evaluating the impact of a continuous treatment. For each case, I review the treatment effects parameters of interest, the required identification assumptions, the definition of the main matching estimators and their main theoretical properties and practical features. The relevance of the "classical" matching estimators and of their extensions for the multiple and continuous treatments settings is illustrated using the example of a health programme implemented with different levels of population coverage in different geographic areas.

Keywords: Evaluation methods, treatment effects, matching, propensity score, programme evaluation.

JEL Classification: C14, C21, C33, I10

* Correspondence: Centre for Health Economics, Alcuin "A" Block, University of York, Heslington, York, YO10 5DD, United Kingdom. Tel: +44 1904 321 411, E-mail: rams500@york.ac.uk.

1 Introduction

In broad terms, impact evaluations of health programmes aim to answer the fundamental counterfactual question: how would the health conditions of individuals exposed to the programme have evolved in the absence of the intervention? Or, analogously, how would those who were not exposed to the intervention have fared in the presence of it? Difficulties in answering such a question rise immediately, as at a given point in time individuals are observed in only one situation, either exposed or not exposed to the programme. This feature, inherent to empirical programme evaluation, resembles the classical missing data problem for econometric analysis.

Matching methods have become increasingly popular among applied researchers as tools for solving the so-called evaluation problem, especially within the field of labour economics. Obviously, as it is the case with any other non-experimental approach, the adequacy of matching estimators in a particular evaluation setting depends crucially on the research questions of interest, the characteristics of the analysed intervention and the available data. These factors have been proved extremely relevant for the validity of matching estimates in a given context.

The general aim of this paper is to review how matching methods try to solve the evaluation problem – with a particular focus on one popular variant, namely propensity score matching – and their usefulness for the particular case of health programme evaluation. The well-known case of matching estimation with a single discrete treatment will be presented as a basis for discussing recent developments regarding the application of matching methods for jointly evaluating the impact of multiple treatments and for evaluating the impact of a continuous treatment. These extensions of the traditional matching estimators can be particularly relevant for the investigation of treatment effects in the health sector.

This paper is organised as follows. The next section describes the evaluation problem with a focus on its particularities for the impact assessment of health programmes. Section 3 discusses matching estimation of treatment effects for the classical case of a single discrete treatment, defining the usual treatment effects parameters of interest, presenting the required identification assumptions, defining the main matching estimators (with a special interest on propensity score matching and including the matching with difference-in-differences variant), and discussing their

main theoretical properties and practical features. Sections 4 and 5 follow the same structure as applied to the joint impact evaluation of multiple programmes and a single continuous treatment, respectively. Section 6 presents some concluding remarks.

2 The evaluation problem

Most of the theoretical and empirical literature on programme evaluation relies, at least implicitly, on basic assumptions regarding participation in the programme. For ease of exposition, it is normally assumed that:

i. *Everyone* who is assigned treatment is actually treated (there are not "no shows");
ii. There are only *partial equilibrium effects*, i.e. the health programme does not affect the (pre-treatment) variables X_i taken as exogenous;
iii. *Stable Unit Treatment Value Assumption (SUTVA)*: This assumption contains usually two components. In the first place, all treated individuals are assumed to receive the same active treatment and all comparison individuals are assumed to get the same comparison treatment. The second component is the assumed absence of interference between units, in the sense that the values of treated and untreated outcomes for a given individual are not influenced by the treatment status of other individuals.

Although the validity of all three assumptions might be questioned in specific settings, the third assumption may be particularly unrealistic in the context of public health interventions. Treatment benefits usually positively affect untreated individuals as well, such as in the classical examples of immunisation campaigns and programmes aimed to reduce the prevalence of communicable diseases. These treatment externalities, whose magnitude is likely to depend on the number of actually treated individuals, pose a significant challenge to the assessment of a programme's impact through individually randomised or non-experimental studies, since there is the possibility of non-negligible treatment benefits accruing to the comparison group. This would lead to an underestimation of the programme effect when comparing the average outcomes of treatment and comparison samples, as demonstrated by Miguel and Kremer (2004). However, these authors also demonstrate that it is sometimes

possible to alleviate deviations from SUTVA through design; for example, by considering higher-level randomisation units rather than individuals (schools in their case). Non-experimental evaluation of health programme treatment effects can deal with deviations from SUTVA in a similar way, for instance by considering the *availability* of a health programme in a given geographic area as the treatment variable of interest.

In order to clarify the issues involved, I consider a very common situation in public health policy in which a health authority implements an intervention in some geographic areas selected according to some pre-specified criteria. In the treated areas – those where the health intervention has been implemented – it is not necessarily the case that *all* residents will actually receive the intervention (due to self-selection into the programme, for instance); rather, treatment here is defined as residing in an area where the programme is available. For simplicity, consider only two localities. If the treated locality is far enough from the untreated locality (that where the health intervention has not been implemented) so as to preclude spillovers from occurring between areas, the "no interference between units" assumption is more likely to be valid and it is possible for the estimated average treatment effects to take into account treatment externalities accruing to "individually untreated" people living in the treated area. In other words, this situation can be seen as a particular case of the SUTVA assumption, in which the unit of analysis is the *group* of individuals living in a locality instead of a single individual.

Formally, using the potential outcome notation suggested by authors such as R.A. Fisher and A. Roy, and popularised by Rosenbaum and Rubin (1983), let the programme impact on a particular health outcome Y for an individual i living in the treated area be given by:

$$Y_i^T - Y_i^C \tag{1}$$

where T refers to the health outcome of an individual belonging to the treatment group – i.e. living in the treated area – and C denotes the counterfactual, the health outcome for the same individual had they been living in the untreated area, thus belonging to the comparison group. The above formalisation assumes that each individual i is characterised by a pair of potential outcomes: Y_i^T for the outcome under the active treatment and Y_i^C for the outcome under the comparison treatment (or no treatment).

Since it is impossible to *observe* the individual treatment effect – because we cannot observe both treated and untreated situations for the same individual *i* – we can aim to learn something about the programme impact through its *average* effect in the population. If data on the health outcome(s) of interest are available for a number of individuals living in the treated area and a number of individuals residing in the untreated locality, we can then average these outcomes in both groups and subtract the second from the first in order to obtain a (naïve) estimate of the programme impact:

$$E\left[Y_i^T - Y_i^C\right] \tag{2}$$

In this case, the average health outcome for individuals living in the comparison area is intended to act as a substitute for the unobservable counterfactual. However, individuals exposed to a programme are usually different in a set of observable characteristics – such as education, income and initial health status – from those individuals who are not covered by the programme. This problem will be magnified if individuals self-select into the programme (for instance, through migration towards the treated area), meaning that unobservable factors such as motivation and relative importance attributed to their own health status are key in determining treatment assignment. This makes it difficult to isolate the differences between both groups which are due to already existing distinctions before treatment – the *selection bias* – from those which are due solely to the programme's impact, as it can be seen by extending equation (2):

$$\begin{aligned} E\left[Y_i^T - Y_i^C\right] &= E\left[Y_i^T \mid T\right] - E\left[Y_i^C \mid C\right] \\ &= E\left[Y_i^T \mid T\right] - E\left[Y_i^C \mid T\right] + E\left[Y_i^C \mid T\right] - E\left[Y_i^C \mid C\right] \\ &= E\left[Y_i^T - Y_i^C \mid T\right] + E\left[Y_i^C \mid T\right] - E\left[Y_i^C \mid C\right] \end{aligned} \tag{3}$$

The first term in (3) represents the average treatment effect on the treated, usually the parameter of interest we want to isolate (which will be formally defined below). This parameter will only be identified if the selection bias, represented by the second and third terms above, equals zero; this, in turn, will only happen if there are no systematic differences in the average untreated health outcomes between treatment and comparison groups.

Equation (3) guides us also regarding the direction of the bias. If individuals living in the treated area are on average healthier, more health-concerned (for example, have a 'healthier' lifestyle) or more motivated to participate into the

programme than their counterparts in the comparison area to begin with, then the selection bias term will be positive and the programme impact on health outcomes will be overestimated. Conversely, if individuals residing in the comparison area tend to have better health prospects on average than those living in the treated area (perhaps because the programme was aimed to target a health-deprived locality), then the selection bias term will be negative and the estimated treatment effect will underestimate the true programme impact.

For simplicity, no covariates have been included up to this point. Consider now the analysis with covariates and define *common support* as the subspace of individual characteristics that is represented both among treated and comparison groups. One important result due to Heckman et al. (1998) is the decomposition of the selection bias into three components:

i. *Non-overlapping support of the observables:* this is the part of the selection bias due to comparing non-comparable individuals. Using the whole sample of treated and untreated individuals can be a source of bias if the treated and comparison support distributions do not intersect for a given range of covariate values;
ii. *Differences in the distribution of the observables between the two groups over the common support:* this bias component is caused by not adequately weighting comparable individuals when there are differences in the shapes of the covariates distributions between treatment and comparison groups. Even in the region of common support, distributional differences make some untreated individuals to be more comparable to some specific treated individuals in terms of the values of their covariates, and misweighting them – i.e. not reweighting the comparison group's data so as to equate the observed covariates' distribution in the treatment group – can lead to estimation bias;
iii. *Selection on unobservables:* this final component of the selection bias results from selective differences between treatment and comparison groups in terms of unobservable characteristics which are also correlated with their potential health outcomes. These differences may be observable by the individual, but not by the researcher.

The first two components are related to observable characteristics and found to be the most important sources of selection bias by Heckman et al. (1998), although the third component was still a sizeable fraction of the total bias.

An evaluation design in which the selection bias problem tends to disappear is that in which treatment and comparison groups are randomly selected from a large population of potential beneficiaries, such as individuals or localities. In other words, due to randomisation, treatment status does not depend on potential health outcomes, and it may be assured that, on average, those individuals exposed to a given programme are not different from those not exposed to it either regarding observable characteristics (such as income, education and age) or unobservable ones. Consequently, any statistically significant difference in health indicators between both groups can be reliably attributed solely to the programme's impact.

In most real situations, nevertheless, health programmes have been purposively implemented by a central authority – for instance by targeting individuals or areas with worse than average health status – and/or require individuals to self-select into the programme by taking up the benefits. If all the researcher has for evaluating a health intervention is non-experimental data, the explicit treatment of the potential bias caused by omitted variables, either unobserved or intrinsically unobservable, is of crucial importance for the reliability of the estimates of a programme's impact. Using matching methods is one alternative for explicitly addressing and eliminating the first two sources of selection bias, while assuming that selection on unobservables is not a problem in the relevant data under certain assumptions. These assumptions and the formal definition of alternative matching estimators for the single treatment case will be explained in the next section, preceded by a formal introduction to the parameters of interest which most of the empirical literature on programme evaluation attempts to estimate when only one treatment is being considered.

3 Matching estimation of health programme average treatment effects with a single treatment

3.1 Average treatment effects: definition, identification and variance efficiency bounds

3.1.1. Definition of average treatment effects

The empirical literature on programme evaluation has traditionally focused on estimating three main average treatment effects when assessing the impact of a single treatment.[1] The definition and identification conditions of each of these treatment effects will be discussed for the same context of evaluating the impact of the availability of a given health intervention in a number of geographic areas.

Average Treatment Effect on the Treated (ATT): this parameter represents the average health impact of the programme among those who have been exposed to it. Formally, it is defined as:

$$ATT = E\left[Y_i^T - Y_i^C \mid T\right] = E\left[Y_i^T \mid T\right] - E\left[Y_i^C \mid T\right] \tag{4}$$

and its sample analogue is $\widehat{ATT} = \frac{1}{N_T}\sum_{i=1}^{N_T}\left(Y_i^T - Y_i^C \mid T\right)$, where N_T stands for the number of treated individuals in the sample. The second term after the last equality in (4) is the counterfactual to be estimated. The ATT is a measure of the average gain from the programme to a treated individual randomly drawn from the treated population, rather than to *any* member of the population. This is usually a parameter of special interest in the context of narrowly targeted health programmes, a setting in which the likely programme impact on untargeted individuals is not the primary interest of policy-makers. For instance, if a health programme is aimed at extremely unhealthy individuals such as people infected with malaria residing in areas with high incidence of the disease, there will probably be little interest from the heath authority in knowing what the programme effect would be on relatively healthy individuals living in urban areas.

Average Treatment Effect on the Untreated (ATU): this alternative estimand is the expected health programme impact among those who have not been treated. In formal terms:

$$ATU = E\left[Y_i^T - Y_i^C \mid C\right] = E\left[Y_i^T \mid C\right] - E\left[Y_i^C \mid C\right] \tag{5}$$

and its sample analogue is $\widehat{ATU} = \frac{1}{N_C}\sum_{i=1}^{N_C}\left(Y_i^T - Y_i^C \mid C\right)$, where N_C represents the number of individuals in the comparison group. The first term after the last equality in (5) cannot be observed and must be estimated. The ATU parameter recovers the expected health impact of the programme on an individual randomly drawn from the

[1] Some extensions of these estimands have also been introduced in the literature; see for instance the "conditional average treatment effects" developed by Abadie and Imbens (2002).

sub-population of individuals non-exposed to the intervention, and is potentially useful if we would like to assess the impact of a programme expansion to initially untreated individuals.

Average Treatment Effect (ATE): this is the third of the most commonly studied average treatment effects, corresponding to the average health programme effect for the entire population, whether or not a particular individual has been treated. Formally:

$$\begin{aligned} ATE &= ATT * P(T) + ATU * P(C) \\ ATE &= E\left[Y_i^T - Y_i^C \mid T\right] * P(T) + E\left[Y_i^T - Y_i^C \mid C\right] * P(C) \end{aligned} \tag{6}$$

where *P(T)* and *P(C)* are the probabilities of belonging to the treatment and comparison groups, respectively. In the sample these probabilities correspond to the sample frequencies of treated and untreated individuals and hence, as a sample analogue, we have $\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i^T - Y_i^C\right) = \frac{1}{N}\left[\sum_{i=1}^{N_T}\left(Y_i^T - Y_i^C \mid T\right) + \sum_{i=1}^{N_C}\left(Y_i^T - Y_i^C \mid C\right)\right]$.

As can be seen, counterfactuals must now be estimated for both components of the ATE. This parameter is relevant for health interventions that could be universally expanded, addressing the question of what the treatment gain would be to a randomly selected member of the population.

3.1.2. Identification of average treatment effects

Consider the evaluation of a health programme in which each individual is either exposed or not to the single active treatment. Following Imbens (2004), let the observations of each individual be characterised by the triple (D_i, Y_i, X_i), where D_i is an indicator taking the value of 1 if the individual has been exposed to the programme and zero otherwise. X_i represents a set of exogenous covariates and Y_i stands for the realised individual health outcome distributions, defined as:

$$Y_i \equiv Y_i\left(D_i\right) = \begin{cases} Y_i^C \; if \; D_i = 0, \\ Y_i^T \; if \; D_i = 1. \end{cases}$$

Identification of average treatment effects through semi-parametric methods such as matching is based on two fundamental assumptions about treatment assignment:

ASSUMPTION 1 (UNCONFOUNDEDNESS): $\left(Y^T, Y^C\right) \perp D \mid X$.

This assumption – also called ignorable treatment assignment (Rosenbaum and Rubin, 1983) or conditional independence assumption (Lechner, 2000) – states that treatment assignment of a given individual is independent of her potential health outcomes with and without treatment if the relevant observable covariates (those that influence an individual's treatment assignment) are held constant. In other words, this assumption means that the researcher observes all factors that jointly affect the potential outcomes and exposure to the programme.

Instead of requiring unconfoundedness to hold as defined in Assumption 1, some applied work has assumed a weaker form of unconfoundedness called *conditional mean independence*, $E\left[Y^T, Y^C \mid D, X\right] = E\left[Y^T, Y^C \mid X\right]$. Nevertheless, as argued by Imbens (2004), this weaker version is intrinsically tied to functional-form assumptions (e.g. linearity of outcomes in *D*). Since it would be difficult to argue that conditional mean independence should hold in a setting where unconfoundedness is violated, the stronger assumption is often invoked. Moreover, the stronger unconfoundedness assumption has also the advantage of making conditional mean independence valid for every transformation of the outcome variable.

As it shall be discussed below, the unconfoundedness assumption is crucial for the estimation of average treatment effects by matching methods. It is also a less restrictive assumption than it might seem at first glance: even if two individuals with the same *X* differ in their choices regarding exposition to the health intervention due to unobservables, this does not necessarily invalidate the unconfoundedness assumption if these different choices are due to unobserved factors that are themselves unrelated to the health outcomes of interest. In terms of a health programme with the characteristics already described, the individual-specific gain from living in an area where the programme is available – which is unobserved by the researcher – is allowed to be correlated with treatment participation, provided that this individual gain is not correlated with the individual's potential health outcome Y^C conditional on *X*. Unobserved characteristics will only lead to selection bias if they are correlated both with exposure to the programme and potential health outcomes, e.g. if "more health-concerned" individuals are also more likely to migrate to areas where the

health programme has been implemented in order to obtain access to it, and this selective migration is not observed by the researcher.[2]

ASSUMPTION 2 (OVERLAP): $0 < P\left(D = 1 \mid X\right) < 1.$

The second fundamental assumption states that there are treated and untreated individuals at *all* values of *X*, i.e. there is overlap between treatment and comparison samples. This assumption refers to the joint distribution of the treatment variable and covariates, implying that, conditional on *X*, there must be other variables which affect exposure to the programme, thus preventing *X* from being a perfect predictor of treatment assignment. Importantly, if the unconfoundedness assumption also holds, these unobserved variables are not correlated with the potential health outcomes.

Under assumptions 1 and 2 above[3], average treatment effects can be identified because, conditional on *X*, the potential health outcomes Y^C of untreated individuals have the same distribution of the (counterfactual) potential health outcomes that treated persons would have experienced had they not been treated. Analogous reasoning applies to the potential outcomes Y^T. Therefore, in formal terms and under the two fundamental assumptions we must have:

$$F\left(Y^C \mid X, D = 1\right) = F\left(Y^C \mid X, D = 0\right)$$
$$F\left(Y^T \mid X, D = 1\right) = F\left(Y^T \mid X, D = 0\right)$$

Average treatment effects can hence be identified because the following equalities hold:

$$E\left(Y^C \mid X, D = 1\right) = E\left(Y^C \mid X, D = 0\right) = E\left(Y^C \mid X\right)$$
$$E\left(Y^T \mid X, D = 1\right) = E\left(Y^T \mid X, D = 0\right) = E\left(Y^T \mid X\right)$$

If there is no omitted variable bias (no confounding) once we condition on *X*, systematic differences – such as average or distributional – in the health outcomes of treated and untreated individuals can be attributed solely to programme exposition. This means we can estimate average treatment effects of a health programme in a subpopulation with covariates $X=x$ by using:

[2] The effects of migration selectivity of the form illustrated here were first studied by Rosenzweig and Wolpin (1988) for the general case of public programmes availability.
[3] Together, assumptions 1 and 2 were denominated "strong ignorability" by Rosenbaum and Rubin (1983).

$$
\begin{aligned}
&E\left[Y^T - Y^C \mid X = x\right] = E\left(Y^T \mid X = x\right) - E\left(Y^C \mid X = x\right) \\
&= E\left(Y^T \mid X = x, D = 1\right) - E\left(Y^C \mid X = x, D = 0\right) \\
&= E\left(Y \mid X = x, D = 1\right) - E\left(Y \mid X = x, D = 0\right)
\end{aligned} \tag{7}
$$

As it can be noticed, it is only possible to estimate (7) if we can estimate the expectations $E\left(Y \mid X = x, D = 1\right)$ and $E\left(Y \mid X = x, D = 0\right)$, that is, if we have common support for all values of *D* and *X*; otherwise, we would have either only treated or untreated individuals at some values of the covariates and would be impossible to estimate both expectations.

If the interest lies on the ATT, both a weaker unconfoundedness assumption and a weaker overlap assumption can be invoked.

ASSUMPTION 3 (WEAK UNCONFOUNDEDNESS): $Y^C \perp D \mid X$.

ASSUMPTION 4 (WEAK OVERLAP): $P\left(D = 1 \mid X\right) < 1$.

These assumptions suffice for identifying the ATT because the moments of the distribution of Y^T for the treated are directly measurable; only assumptions about the potential outcomes of comparison individuals are needed for estimating the counterfactual in the ATT formula. Analogously, only $Y^T \perp D \mid X$ and $0 < P\left(D = 1 \mid X\right)$ are required if the ATU is to be estimated.

Identification of average treatment effects by matching methods is thus based on a basic assumption, unconfoundedness, which may or may not be plausible depending on the particular context, and which is inherently untestable due to the impossibility of actually observing the counterfactual. Therefore, tests for assessing the validity of this assumption in the data can only be indirectly made. Imbens (2004) discusses two approaches. One alternative is to use only comparison groups – such as individuals who live in two localities where the health programme is not available – for estimating the average treatment effect of interest, considering one of these groups as the "treated" sample.[4] Although not a conclusive evidence, non-rejection of the null hypothesis of no treatment effect makes more plausible that unconfoundedness holds in the data, whilst rejecting the null points to the invalidity of at least one of the comparison groups.

Another approach for proxy-testing the unconfoundedness assumption would be to use the model for estimating the treatment effect on a variable determined before

[4] Estimation methods are discussed in the next section.

the intervention was launched and thus not affected by exposure to the programme, for instance a lagged value of the health outcome of interest (or a number of them, provided they do not affect future treatment status). If a statistically significant treatment effect is found, there is evidence that treatment and comparison groups are systematically different, and so are their outcome distributions. Conversely, non-rejecting the null gives some credibility to the unconfoundedness assumption in the analysed setting.[5]

UNCONFOUNDEDNESS GIVEN THE PROPENSITY SCORE. A striking result due to Rosenbaum and Rubin (1983) is that, if unconfoundedness holds by conditioning on *X*, all biases due to observable characteristics are also removed by conditioning solely on a scalar representing the individuals' conditional probability of receiving treatment given the set of observable pre-treatment characteristics *X* – known as the *propensity score* – and hence the unconfoundedness assumption remains valid. Formally, define the propensity score *p(X)* as the conditional probability of receiving treatment:

$$p(X) \equiv P(D=1 \mid X=x) = E[D \mid X=x] \tag{8}$$

Then, it can be shown that the following must be true:

$$(Y^T, Y^C) \perp D \mid X \Rightarrow (Y^T, Y^C) \perp D \mid p(X) \tag{9}$$

The proof of the above result is given in Rosenbaum and Rubin (1983). It implies that the important results of unconfoundedness given covariates also hold when conditioning solely on the propensity score: if by conditioning on *X* we get rid of the correlation between *D* and *X*, the same occurs if we condition on the propensity score instead. In this case, for instance, estimation of the ATT can be based on:

$$E[Y^C \mid p(X)=p] = E[Y^C \mid p(X)=p, D=0] = E[Y^C \mid p(X)=p, D=1]$$

and hence $E[Y^C \mid D=1] = E\left[E[Y^C \mid p(X)=p, D=0] \mid D=1\right]$ can be used for estimation purposes. It should be noted that, whilst the first part of (9) (independence of treatment assignment and potential outcomes given observables) represents an

[5] Frolich (2004) discusses a somewhat weaker test, based on additional untestable assumptions, which consists in testing the equality of mean pre-programme conditional outcomes between treatment and comparison groups, $E\left[Y_{t-1}^T \mid X\right] = E\left[Y_{t-1}^C \mid X\right] = E\left[Y_{t-1} \mid X\right]$. Systematic differences between the two mean outcomes would cast doubt on the plausibility of the unconfoundedness assumption in that particular setting. It must be noted that, if the interest lies on the programme effects over more than one health outcome, the plausibility of the unconfoundedness assumption should be analysed on a case-by-case basis.

assumption and is hence inherently untestable, independence between the treatment variable and covariates once conditioning on the propensity score, $X \perp D \mid p(X)$, is a key condition for obtaining reliable estimates in a propensity score matching estimation context and can be tested with the observed data, as it will be discussed in Section 3.2.2.

3.1.3. Asymptotic variances and efficiency bounds for average treatment effects

As expected, estimators proposed in the programme evaluation literature for recovering the average treatment effects described above have been judged not only according to their unbiasedness and consistency results, but also according to their ability of achieving the so-called *efficiency bound* – the lower bound for the asymptotic variance of a root-N consistent estimator. Efficiency bounds have been derived for estimators of the main average treatment effects; here, I will focus on the ATE and ATT parameters.

Following Hahn (1998), let $\sigma_C^2(X) = Var\left(Y^C \mid X\right)$ and $\sigma_T^2(X) = Var\left(Y^T \mid X\right)$ be the conditional variances of the potential outcomes, $E\left[p(X)\right]$ the unconditional treatment probability, $\tau(X) = E\left(Y^T - Y^C \mid X\right)$ the conditional ATE, and $\tau = E\left[\tau(X)\right]$ the unconditional ATE; analogously, let the unconditional ATT be $\tau_1 = E\left[\tau(X) \mid D = 1\right]$. Hahn (1998) shows that estimators of the ATE must have asymptotic variances such as:

$$\hat{\sigma}_{ATE}^2 \geq E\left[\frac{\sigma_T^2(X)}{p(X)} + \frac{\sigma_C^2(X)}{1-p(X)} + \left(\tau(X) - \tau\right)^2\right] \tag{10}$$

Hahn (1998) also shows that knowing the propensity score does not affect the variance lower bound (10) for estimating the ATE, but it does change (reduces) the lower bound for estimating the ATT. The lower bound for the asymptotic variance of a root-N consistent ATT estimator when the propensity score is known must be:

$$\hat{\sigma}_{ATT}^2 \geq E\left[\frac{p(X)\sigma_T^2(X)}{E\left[p(X)\right]^2} + \frac{\left[p(X)\right]^2 \sigma_C^2(X)}{E\left[p(X)\right]^2\left[1-p(X)\right]} + \frac{\left(\tau(X) - \tau_1\right)^2\left[p(X)\right]^2}{E\left[p(X)\right]^2}\right] \tag{11}$$

whilst without knowledge of the propensity score the lower bound will be:

$$\hat{\sigma}^2_{ATT} \geq E\left[\frac{p(X)\sigma^2_T(X)}{E\left[p(X)\right]^2}+\frac{\left[p(X)\right]^2\sigma^2_C(X)}{E\left[p(X)\right]^2\left[1-p(X)\right]}+\frac{\left(\tau(X)-\tau_1\right)^2 p(X)}{E\left[p(X)\right]^2}\right] \quad (12)$$

which is larger than (11). Intuitively, this is because the ATT is a weighted average of the treatment effect conditional on the covariates, with weights given by the product of the density of the covariates and the propensity score; if the latter is known, there is no need to estimate the weighting function and precision is improved, leading to the reduced lower bound (12) (Imbens, 2004).

By inspecting the formulae above, it is clear that estimating the variances of ATE and ATT estimators is a difficult task. This requires the estimation of at least one unknown regression function and conditional variance and usually of the propensity score as well, as can be seen by rewriting the variance lower bound for the ATE case:

$$\hat{\sigma}^2_{ATE} \geq E\left[\frac{\sigma^2_T(X)}{p(X)}+\frac{\sigma^2_C(X)}{1-p(X)}+\left[E\left(Y^T \mid X\right)-E\left(Y^C \mid X\right)-\tau\right]^2\right]$$

Estimation of all the components above can be done, but involves additional burden to the average treatment effect estimation. A simpler alternative – commonly used in applied work, as it will be discussed below – is to use bootstrapping methods.

Up to this point, the discussion has been focused on the estimands of interest and their characteristics. Thus, the next step is to explain how matching methods can recover the average treatment effects of interest for health programme evaluation.

3.2 Matching estimation of average treatment effects

Similarly to estimations based on natural experiments, matching methods attempt to mimic an experiment using non-experimental data and, for this purpose, need to make some independence and exclusion assumptions. All the matching procedures that will be discussed below – including propensity score matching and its extensions for the multiple and continuous treatments cases – rely on some version of the fundamental unconfoundedness assumption (Assumption 1) coupled with overlap (Assumption 2), suitably adapted for the estimation of the average treatment effect of interest. For ease of exposition, most of the discussion in this section will focus on estimating the ATT; extending the ideas for the estimation of other parameters such as

the ATE is often straightforward and will be explicitly addressed according to necessity.

3.2.1. Matching on covariates

The matching method is a non-parametric approach that tries to re-establish the conditions of an experiment when only non-experimental data are available. It is non-parametric because no particular specification needs to be assumed for the outcomes, treatment decision process or the unobservable term. The broad idea is to construct a matched comparison group – containing the missing counterfactual information – based on individual observable characteristics: individuals will be compared only to their counterparts who are similar in terms of these observable factors. As it was explained above, the observable characteristics on which matching will be based should be those that affect the individual treatment status and health outcomes simultaneously. Variants of this method have proven very useful in empirical research, mainly (but not only) when the average treatment effect of interest is the ATT and when there is a large pool of comparison individuals.

Taking Blundell and Costa-Dias (2000) exposition as a starting point, let *S* be the *support* (set of all possible values) of the vector of explanatory variables *X*, and let *S** be the *common support* of *X*, the space of *X* that is observed both among treatment and comparison groups in the dataset. A consistent estimator for the ATT of a given health programme is the empirical counterpart of:

$$\frac{\int_{S^*} E\left[Y^T - Y^C \mid X, D=1\right] dF\left(X \mid D=1\right)}{\int_{S^*} dF\left(X \mid D=1\right)} \tag{13}$$

where the numerator is the expected health benefit for individuals exposed to the health intervention for whom it was possible to find a comparable (in *X* terms) unexposed individual – i.e., over the common support. Individual health gains must then be integrated over the distribution of observables among treated individuals and re-scaled by the dimension of the common support. As an illustration, let *X* take only discrete values; then, the sample analogue of expression (13) means that treated and comparison individuals will be compared in all cells formed by the combination of *x*'s and a weighted average over these cells will be taken, using as weights the proportion

of treated individuals in each of the cells. Furthermore, cells with only treated or comparison individuals will not be used for estimation purposes.

Thus, a consistent estimator for the ATT (13) is simply the mean conditional difference in health outcomes over the region of common support S^*, appropriately weighted by the distribution of treated individuals over $X \in S^*$. Clearly, matching methods will only recover the parameter of interest provided that the outcomes for comparison individuals are good approximations to the counterfactual, i.e. if matching is performed within the common support region. In this case, the assumption of unconfoundedness in the common support region can be invoked as a basis for matching estimation of the ATT:

ASSUMPTION 5 (UNCONFOUNDEDNESS IN THE COMMON SUPPORT REGION):

$$Y^C \perp D \mid X \quad for \; X \in S^*.$$

However, one of the limitations of matching methods is that they do not ensure that the support for the comparison group equals the support for the treatment group; in other words, often (13) cannot be identified for all subsets of *S* given *D=1*, and a different parameter – no longer the experimental sample average treatment effect – is being defined and estimated. Also, depending on the nature of the health programme, the weak overlap assumption (Assumption 4) for identifying the ATT can represent quite strong a requirement, for instance when the health intervention has been targeted to a very specific group. If the impact of the health programme is homogeneous within the treatment group, the only problem of not finding a suitable counterfactual for some treated individuals and hence discarding them will be the loss of information. On the other hand, if the programme effect is indeed heterogeneous and the counterfactual cannot be obtained for some subgroups of treated individuals, the loss of observations also limits the parameter that can be identified, which will be consistent only for the region of common support. In this situation, it is possible that the estimated impact does not represent the mean impact of the health programme, but insisting in the estimation of a treatment effect by matching without common support can introduce severe bias by relying on the matching of treated individuals – with possibly outlying covariate values – to substantially different comparison individuals.

The next step is to generally define the matching estimator in formal terms. The main idea of matching is to pair to each treated individual another (or a group) of comparison individual(s), associating to the health outcome Y_i^T of the treated person *i*

a matched outcome $\hat{Y}_j^C$ given by the (weighted) outcome(s) of her "neighbour(s)" *j* in the comparison group. The general form of the matching estimator for the ATT within the common support region is given by:

$$\widehat{\beta}_M^{ATT} = \sum_{i \in \{T \cap S^*\}} \left(Y_i^T - \hat{Y}_j^C \right) W_i \tag{14}$$

In (14), the summation is performed over the group of individuals belonging to the treatment group *T* and falling within the common support region S^*. The term W_i stands for the reweighting that reconstructs the health outcome distribution for the treated sample (Blundell and Costa-Dias, 2000). The matching estimator for the ATT usually takes the form:

$$\widehat{\beta}_M^{ATT} = \sum_{i \in \{T \cap S^*\}} \left(Y_i^T - \hat{Y}_j^C \right) \frac{1}{N_T^*} \tag{15}$$

where N_T^* denotes the number of treated individuals falling within the common support region. Note that $\sum \hat{Y}_j^C \frac{1}{N_T^*}$ corresponds to the estimator for the average untreated counterfactual for treated individuals, $E\left[Y^C \mid T \right]$. The general form of the estimator for the counterfactual for treated individual *i* is:

$$\hat{Y}_j^C = \hat{E}\left(Y_i^C \mid X, D = 1 \right) = \sum_{j \in \{C(X_i) \cap S^*\}} W_{ij} Y_j^C \tag{16}$$

where $C(X_i)$ defines comparable neighbours of *i* in terms of *X* characteristics and W_{ij} is the weight placed on untreated individual *j* when compared to *i*, with $W_{ij} \in [0,1]$ and $\sum_{j \in \{C(X_i) \cap S^*\}} W_{ij} = 1$. A commonly used matching estimator (e.g., Dehejia and Wahba, 2002) takes the form:

$$\widehat{\beta}_M^{ATT} = \sum_{i \in \{T \cap S^*\}} \left(Y_i^T - \sum_{j \in \{C(X_i) \cap S^*\}} \frac{1}{N_{C(X_i)}^*} Y_j^C \right) \frac{1}{N_T^*} \tag{17}$$

The particular form assumed by the ATT matching estimator for the counterfactual (16) depends on the number of neighbours that will be used in constructing the counterfactual health outcome for each treated individual and also on the weighting scheme chosen. The simplest matching procedure – called *nearest-neighbour matching (NNM)* – consists in using only the health outcome of the

observably closest untreated individual as the matched counterfactual. Consequently, the matching estimator formula collapses to:

$$\widehat{\beta}_{NNM}^{ATT} = \sum_{i \in \{T \cap S^*\}} \left(Y_i^T - Y_j^C\right) W_i \tag{18}$$

where *j* now refers to the closest comparison individual to treated individual *i*. Other matching procedures will be discussed later in the context of propensity score matching; they include *radius matching* (which uses multiple matches when there is more than one comparison observation within a tolerated distance of the treated individual in terms of observables to perform the matching, and the counterfactual is an average outcome of these comparison individuals) and *kernel matching* (where the counterfactual comes from a weighted average of the outcomes of several or all comparison individuals, weights being defined according to the "closeness" of their characteristics based on a metric function such as Gaussian or Epanechnikov kernel).

The definition of *closeness* for choosing the comparable neighbours *j* is a required step in matching estimation. As mentioned above, the set of comparable neighbours for a given treated individual may be restricted to one comparison individual or contain many of them who are considered "close enough" in terms of observables and whose health outcomes might be differently weighted according to their degree of similarity. Closeness of the individual vectors of covariates is usually measured by employing Euclidean or Mahalanobis metrics. The latter metric incorporates the former but has the advantage of taking into account the correlation between coordinates of *X*: the Mahalanobis distance between treated observation *i* and comparison observation *j* is given by:

$$d_M\left(X_i, X_j\right) = \sqrt{\left(X_i - X_j\right)' V^{-1} \left(X_i - X_j\right)}$$

where *V* corresponds to the covariance matrix of the covariates in the sample (only in the treated sample if estimating the ATT). In this way, when comparing the covariate vectors of any two individuals, the contribution of the Euclidean distance measured for a given covariate to the *total* Mahalanobis distance will depend on the precision (in variance terms) with which that particular covariate is measured: the more (less) precisely that covariate is measured in the sample, the more (less) weight its corresponding distance will be given in the computation of the total Mahalanobis distance.

Different weighting schemes for matched comparison individuals (as in nearest-neighbour, radius or kernel matching), which reflect different ways of using the available information, can potentially influence the results of the average treatment effects estimation. Abadie and Imbens (2002) show that simple matching estimators such as nearest-neighbour can suffer from bias and not to be root-N consistent if more than one continuous covariate is used (although the bias can be small or even disappear under specific conditions, for instance by having a large comparison group relatively to the size of the treatment group). The authors also find that matching estimators are not generally efficient given a fixed number of matches.[6] However, as the extensive review made by Imbens (2004) suggests, the current debate on the practical advantages of each matching estimator is still inconclusive.

For the unconfoundedness assumption to be valid in estimating the ATT, the matched comparison group cannot differ from the treatment group by any variable that is systematically related to the potential outcome Y^C. Choosing a "good" and rich set of covariates is therefore essential for the credibility of the unconfoundedness assumption in the context of matching estimation. Imbens (2004) identifies two main issues that dominate the choice of the covariates set. Firstly, in finite samples, including covariates that are only weakly correlated with the treatment variable and/or the health outcomes may decrease precision (or, in more formal words, increase the expected mean squared error). Secondly, and more importantly, covariates that are themselves affected by the health programme – such as intermediate outcomes – should not be included in the covariates set; as a general rule, health outcomes of potential interest for the impact evaluation of a particular programme must be excluded from the matching variables used.[7]

In summary, the set of matching covariates should basically include pre-treatment variables, time-invariant characteristics (such as gender and education) and variables that are deterministic with regard to time (such as age). Factors which affect only treatment status or the potential health outcomes do not need to be controlled for. Thus, for instance, if individuals with a higher unobserved trait are more likely to be

[6] In practice, since the matching estimator is the (weighted) difference between two sample means, standard methods for calculating the variance for difference in means in randomised studies have been commonly applied without additional corrections for potential biases (Imbens, 2004).

[7] Behrman et al. (2004) use the percentage of individuals correctly classified regarding treatment status (hit-or-miss criterion) as a simple test to choose among sets of regressors for estimating the propensity score.

exposed to the health intervention, but this unobserved factor has no effect on potential health outcomes, then it needs not be included in the set of covariates *X*. Treated and matched comparison individuals do not have to be similar regarding *all* observable characteristics, but instead regarding all *confounding* variables. Whether this is achieved with a particular dataset must be argued on a case-by-case basis, taking into consideration also the institutional characteristics that drive the selection into treatment. If a dataset contains high-quality data rich in covariates associated both with exposition to the programme and health outcomes, matching becomes a more sensible choice.

However, one important consequence of the above is that, the more data the researcher uses, the more difficult it will be to find similar untreated individuals and the more restricted the common support region can become – for instance, with discrete *X*, small or empty cells may be obtained. An additional general limitation of the matching method lies on the difficulty of finding matches when a wide range of *X* variables is being used. Apart from imposing linearity in the parameters (and thus coming back to ordinary least-squares regression analysis), one possibility for reducing the high dimensionality problem is to combine all covariates into a scalar measuring the distance between observations *i* and *j* (using Euclidean or Mahalanobis metrics), a procedure known generally as *inexact matching*. Another possibility (more commonly used) is to match on the propensity score, a scalar which also condenses all the information contained in the covariates vector; this will be the topic of the next sub-section.

3.2.2. Matching on the propensity score

An alternative for solving the matching version of the curse of dimensionality is to use the propensity score as the matching criterion. The problems are now only the estimation of the propensity scores of each individual of the sample as a function of the covariates, and the estimation of the mean health outcome in the comparison group as a function of the propensity scores. The former is usually done parametrically, whereas the specifications of $E\left[Y^T - Y^C \mid p(X)\right]$, $E\left[Y^C \mid p(X)\right]$ and $E\left[Y^T \mid p(X)\right]$ are left unrestricted, resulting in a *semi-parametric* method. The

conditioning vector of individual characteristics has thus its dimension reduced to one.[8]

Intuitively, matching on the propensity score works because it imposes the same distribution of covariates for the treatment and comparison groups, that is, the density of the matching covariates does not vary with treatment status; therefore, unconfoundedness given $p(X)$ implies the same given *X*. The fundamental result is that, under unconfoundedness, conditioning on the propensity score leads to the removal of the correlation between the set of covariates *X* and treatments status *D*:

$$X \perp D \mid p(X)$$

The propensity score is called a *balancing score* due to its ability of balancing the relevant covariates across the matched groups (Rosenbaum and Rubin, 1983). This approach divides the sample into sub-samples where causal comparisons can be performed and appropriately reweighs the health outcomes of the comparison group individuals. Once we condition on the propensity score, the resulting distribution of covariates should be the same in the treatment and comparison groups, and being exposed to the health programme or not should be now random for a group of individuals with similar propensity scores.[9] Thus, the omission of *X* does not lead to any bias, although it may still lead to efficiency loss due to less information used (Imbens, 2004).

In practice, however, individuals with similar propensity score values might end up being quite dissimilar regarding a few covariates deemed very important for explaining the selection into the health programme and potential health outcomes; therefore, in finite samples, it can be more efficient to match on a vector including a combination of the individual propensity score and a few important covariates (rather than solely on the propensity score), achieving a better balancing of the relevant observables in the researcher's specific context. Importantly, given that the propensity score is a balancing score, any combination of that conditional probability with

[8] One way of using the estimated propensity scores for estimating average treatment effects is in fact an extension of traditional regression methods. The main idea is to use the propensity scores as weights for the observations, which are weighted by the inverse of the probability of being assigned to the treatment actually received, so as to balance the distribution of covariates between treatment and comparison groups. The *propensity score weighting method* will not be discussed here due to the fact that it does not rely on matching procedures; the interested reader is referred to Imbens (2004) for a basic review of this estimation method. My discussion focuses on propensity score matching methods.

[9] The rigorous definition of the term "similar" will be given below and depends on the particular propensity score matching procedure to be applied.

elements of the vector of covariates X – which would contain more information than using only the propensity score – is also a balancing score.

Since the balancing of covariates between treatment and comparison groups is essential for obtaining reliable estimation results, it is good practice to assess the balancing condition after conditioning on the propensity score. This can be visually inspected by plotting the distributions of propensity scores for treatment and comparison samples, which should be now roughly equivalent. A test for the equality of means of propensity scores in the matched and unmatched sub-samples can complement this visual inspection. A regression-based test may be regressing the treatment variable on the set of covariates before and after matching has been performed; if treatment assignment is really random after matching on the propensity score, the regression will show low explanatory power (R-squared) and one will not be able to reject the null hypothesis of joint insignificance of the entire covariates' set. Other indicative tests include checking whether the calculated Mahalanobis distance between the vectors of covariates of treated and untreated groups is indeed close to zero after matching, for instance.

In principle, the propensity score could be estimated using either parametric or non-parametric procedures. Nevertheless, it is well known that non-parametric estimators tend to run into trouble when the number of covariates is large, resulting in a return to the dimensionality problem that, after all, propensity score matching seeks to avoid. For this reason, propensity scores are usually estimated by parametric techniques in applied work. Since the problem consists on estimating a conditional probability – for a binary treatment variable in the case of a single treatment – there is a good idea about the methods that can perform adequately for this task. Thus, logit or probit specifications are normally chosen for estimating the individual probabilities of being exposed to the active treatment conditional on covariates.

As previously described, the coupling of the balancing score result due to Rosenbaum and Rubin (1983) with the unconfoundedness assumption means that the ATT of a given health programme can be estimated using the fact that $E\left[Y^C \mid D=1\right] = E\left[E\left[Y^C \mid p(X)=p, D=0\right] \mid D=1\right]$. The propensity score (PS) matching estimator for the ATT takes the same general form of the corresponding matching estimator:

$$\hat{\beta}_{PS}^{ATT} = \sum_{i \in \{T \cap S^*\}} \left(Y_i^T - \hat{Y}_j^C \right) W_i$$

The counterfactual is constructed as a weighted average of the health outcomes of matched individuals not exposed to the intervention, but this time matched according to the propensity score:

$$\hat{Y}_j^C = \hat{E}\left(Y_i^C \mid p(X), D = 1 \right) = \sum_{j \in \{C(p_i(X)) \cap S^*\}} W_{ij} Y_j^C \tag{19}$$

A common formulation of the PS matching estimator for the ATT is given by:

$$\hat{\beta}_{PS}^{ATT} = \sum_{i \in \{T \cap S^*\}} \left(Y_i^T - \sum_{j \in \{C(p_i(X)) \cap S^*\}} W_{ij} Y_j^C \right) \frac{1}{N_T^*} \tag{20}$$

The weight W_{ij} accruing to comparison individual *j* when constructing the counterfactual for treated person *i* is now dependent on the distance between their propensity scores, $p_j(X)$ and $p_i(X)$ respectively, instead of being dependent on the distance between their whole sets of observables as in the matching on covariates' case. As it will be discussed below, the exact weighting scheme to be applied to comparison individuals on the basis of their propensity scores will vary according to the specific matching method being used.

An important step in performing propensity score matching is the weighting procedure adopted for constructing the counterfactual. Following Smith and Todd (2005), define p_i and p_j as the estimated propensity scores of treated individual *i* and comparison individual *j*, respectively. Let the *neighbourhood set* $C(p_i)$ – the set of comparable neighbours for *i* – be the set of individuals belonging to the comparison group *C* for whom the propensity score is close to *i*'s propensity score by some pre-defined measure. Therefore, the *matched set*, the set of comparison individuals matched to *i*, can be defined as:

$$M_i = \left\{ j \in C \mid p_j \in C(p_i) \right\} \tag{21}$$

The propensity score matching methods described below differ in the way they define "closeness" – that is, the set $C(p_i)$ – and how the weights W_{ij} are constructed.

A. Nearest-neighbour matching

The simplest form of applying this matching method is by using the health outcome of the closest comparison individual as the counterfactual, *without*

replacement (i.e. each comparison observation can serve as a match for at most one treated person). The closest comparison individual forms the singleton matched set M_i defined in (21) and is the one for whom the following is true:

$$M_i = \left\{ j \in C \mid j = \arg\min_j \left\| p_i - p_j \right\| \right\}$$

where the term $\left\| p_i - p_j \right\|$ denotes the Euclidean distance between propensity scores.[10] If there is a tie between two or more comparison observations, this tie is usually broken by a random draw.

This simplest form can lead to considerable bias if it results in many bad matches, due to treated individuals being matched to comparison counterparts which have very different propensity scores (despite being the "closest" neighbours). This is the consequence of having regions in the covariate space with low density of propensity scores for the comparison group relative to the treatment group, and of not imposing *a priori* a common support restriction. Additionally, the results obtained will depend on the order in which individuals get matched.

Some flexibility can be introduced by allowing the matching procedure to be performed with replacement and/or multiple matches. *Matching with replacement* is less demanding in terms of the overlap condition because it allows extreme observations within the comparison group to be used more than once. If re-use occurs, matching with replacement will use better matches for each treated individual thus reducing the bias, but the variance of the estimates will probably be higher than in the no-replacement case due to the smaller number of different comparison observations used to construct the counterfactual.

Therefore, if replacement is allowed but with only the closest neighbour being matched to the treated observation, very few observations in the sample might end up being heavily used even with similar comparison observations being available, leading to an unnecessary increase in variance. On the other hand, using *multiple nearest neighbours* tends to reduce the variance of the treatment effects estimates

[10] Straightforward modifications to the conditions for the neighbourhood set in each propensity score matching procedure apply to the "matching on covariates" case. For instance, if nearest-neighbour matching on a set of variables X is being performed (instead of PS matching), then the condition would be $C(X_i) = \min_j \left\| X_i - X_j \right\|$, $j \in C$, referring to the Euclidean distance between the vectors of covariates of treated individual i and comparison individual j. An alternative metric, such as the Mahalanobis distance, could also be used.

(more information is used in constructing the counterfactual) at the likely cost of increased bias due to poorer matches on average.

B. Caliper matching

This is an extension of nearest-neighbour matching that avoids bad matches by constructing the matched set M_i using only comparison individuals whose propensity scores are within a tolerated distance from p_i. In formal terms, the matched set is denoted by:

$$M_i = \left\{ j \in C \mid \left\| p_i - p_j \right\| < \varepsilon \right\} \tag{22}$$

where $\varepsilon > 0$ is a pre-defined tolerance distance, the *caliper*. If, for instance, the interest lies on estimating the ATT, the nearest neighbour will only be matched to the corresponding treated observation if the comparison individual's j propensity score falls within the aforementioned tolerated distance.

The definition of the caliper must be left to the researcher's subjective concept of reasonability. If the caliper criterion leads to the exclusion of treated individuals from the analysis (those without a suitable comparison match), one may have to redefine the treatment effect parameter being estimated and focus on an ATT for individuals within a particular range of covariates values. Caliper matching can be made more flexible by using multiple matches when there is more than one suitable comparison observation to perform the matching; in this case, the counterfactual would be an average health outcome of comparison individuals within the caliper. This procedure has been denominated *radius matching*.

Therefore, caliper matching enforces the common support condition by excluding from the analysis those individuals exposed to the health intervention for whom it was not possible to find at least one good match, that is, at least one comparison individual whose propensity score falls within the tolerance distance.

C. Stratification or interval matching

The main idea of this matching method is to divide the common support region into intervals (or "blocks") and then calculating one mean treatment effect for each interval. The overall ATT is computed as a weighted average of mean interval effects, with weights being defined as the number of treated individuals in each interval (analogously, weighting by the total number of individuals in each interval leads to

the estimate of the ATE for that particular block). Implementation of this method is therefore to be preceded by the definition of the matched set M_i – and thus of the common support region – by one of the procedures described herein.

In formal terms, let the common support of $P(D=1|X)=p(X)$ be partitioned in I intervals containing both treated and comparison individuals – hence discarding observations in intervals in which there are either only treated or only comparison individuals. Also, let $\widehat{ATT_k}$ represent the difference between average treated and comparison outcomes within the kth interval (as if randomisation of treatment had occurred within that particular block) and $N^*_{T\in k}$ denote the number of treated individuals falling within the common support region and in interval k. If we use the following estimator for the ATT in the kth interval:

$$\widehat{ATT_k}=\left(\sum_{i\in k}Y_i^T\frac{1}{N^*_{T\in k}}\right)-\left(\sum_{j\in k}Y_j^C\frac{1}{N^*_{C\in k}}\right),$$

then the overall estimated ATT will be given by:

$$\hat{\beta}^{ATT}_{IM}=\sum_{k=1}^{I}\left(\widehat{ATT_k}\times N^*_{T\in k}\right)\frac{1}{N^*_T} \qquad (23)$$

One possibility (performed by Dehejia and Wahba, 1999) is to define the intervals so as to have statistically insignificant differences between the estimated propensity scores of treated and comparison individuals within each interval; the balancing of important covariates within each block might be separately assessed as well. Alternatively, the common support region can be divided into five categories – quintiles – so as to have estimates of the ATT of a health programme according to quintiles of income, for instance.

If the model for the propensity score has been adequately specified, one should expect that the distribution of covariates among treatment and comparison groups are well balanced within each interval. This works in fact as an informal test of the statistical model. When covariates in a given interval end up not being well balanced among groups, the researcher can interpret this either as an evidence of the need for additional splitting of intervals, or for improving the statistical model of the propensity score by, for example, adding more covariates. According to a review made by Imbens (2004), no formal algorithm has been proposed for dealing with the issue of the optimal number of blocks in finite samples, although based in asymptotic properties the author sees no apparent harm in choosing a large number of intervals.

Interval matching as described so far allows the estimation of average treatment effects in a non-parametric way (of course, usually after the parametric estimation of the propensity scores, as in the other matching methods); this is done by approximating the unknown function by a step function with fixed jump points, which leads to substantial difficulties for establishing asymptotic properties for this estimator (Imbens, 2004). As an alternative, least squares regression might be used for estimating the average treatment effect of interest within each interval. This is equivalent to taking the "unadjusted" regression of the health outcome on the treatment variable within each block *k* and adding covariates to it, resulting in an estimated average treatment effect $\hat{\beta}$ given by the regression $Y = \alpha_k + \beta_k' D + \gamma_k' X + \varepsilon.$

D. Kernel and local linear matching

In this matching procedure, the counterfactual for each treated individual is constructed by using a kernel-weighted average over multiple comparison individuals. The general form of the kernel matching estimator is given by:

$$\hat{\beta}_{KM}^{ATT} = \frac{1}{N_T^*} \sum_{i \in \{T \cap S^*\}} \left(Y_i^T - \frac{\sum_{j \in C(p_i)} Y_j^C K\left(\frac{p_j - p_i}{h_n} \right)}{\sum_{j \in C(p_i)} K\left(\frac{p_j - p_i}{h_n} \right)} \right) \tag{24}$$

where $K(\cdot)$ is a chosen kernel function with mean zero and which integrates to one; also, h_n is a bandwidth parameter which tends to zero as *n* tends to infinity (Smith and Todd, 2005). The weighting term W_{ij} in (20) is given by $\left(K(\cdot) \Big/ \sum_{j \in C(p_i)} K(\cdot) \right)$ and, as usual, depends on the distance between the propensity score of each comparison observation and the treated individual for which the counterfactual is being constructed. For kernel functions taking non-zero values only on the interval [-1,1], the matched set is denoted by:

$$M_i = \left\{ j \in C \,|\, \left| \frac{p_i - p_j}{h} \right| \leq 1 \right\} \tag{25}$$

The kernel counterfactual estimate is a weighted average of the health outcomes of comparison observations within the bandwidth at the current point of evaluation (treated individual *i*), with greater weights placed on comparison observations with propensity scores closer to p_i. In matching estimation, the most common kernel

functions are second-order kernels such as the uniform, Epanechnikov, biweight, tricubic and Gaussian. The choice of a particular kernel function will not only affect the specific weight accruing to a given comparison observation, but also the composition of the common support region *S**: for instance, while the Gaussian kernel uses *all* comparison observations in constructing the counterfactual, the other functions restrict the neighbourhood set to those comparison observations for which $\left|p_i - p_j\right| \leq h$.[11]

More important than the particular kernel function chosen is the choice of the bandwidth. Kernel estimates are obtained by slicing the data into ever smaller intervals as the sample size tends to infinity and estimating local behaviour within each interval. As such, the kernel counterfactual estimator belongs to the category of *local averaging methods* for which the definition of "localness" influences results in finite samples, turning the bandwidth choice for smoothing the density estimate an important step (Cameron and Trivedi, 2005). It can be noticed by looking at (25) that, keeping the distance between propensity scores constant, choosing a smaller bandwidth is equivalent to being stricter in defining the common support region, by excluding some previously used comparison observations – those who no longer satisfy the requirement $\left|p_i - p_j\right| \leq h$ – and placing a heavier weight on the health outcomes of those comparison individuals who are closer to the treated observation.[12] Conversely, increasing the bandwidth means being more tolerant in terms of the closeness requirements imposed to comparison observations.[13]

Smith and Todd (2005) note that kernel matching can be seen as a weighted regression of Y_j^C on an intercept with kernel weights which vary according to the evaluation point, with the estimated intercept providing an estimate of the counterfactual mean. Additional adjustments for asymmetries of comparison observations around treated observations can be made by including in the regression a linear term in p_i. This is equivalent to a more general specification of the kernel

[11] Common support restrictions may be placed on the treatment group as well, e.g. by excluding treated individuals whose propensity scores are larger than the largest propensity score in the comparison group. A similar though more refined procedure is to "trim" the common support region; this procedure is explained in Smith and Todd (2005).

[12] In the limit, continuously reducing *h* would lead to nearest-neighbour matching.

[13] There are different procedures for defining an "optimal" bandwidth (which will vary with the particular kernel function chosen). One common method is to use a plug-in estimate for the bandwidth, a simple formula which depends on the sample size and the sample standard deviation, such as the Silverman's plug-in estimate (Cameron and Trivedi, 2005).

matching estimator (24) called *local linear matching*, proposed by Heckman et al. (1997) in order to correct for the fact that the probability mass of the estimated propensity scores of treated and comparison observations in their sample tended to be concentrated at the boundaries (close to one and zero, respectively). In local linear matching, the weight W_{ij} is given by a combination of the kernel functions for comparison individuals when matched to a given treated observation:

$$W_{ij} = \frac{K_{ij}\sum_{k\in C} K_{ik}\left(p_k - p_i\right)^2 - \left[K_{ij}\left(p_j - p_i\right)\right]\left[\sum_{j\in C} K_{ik}\left(p_k - p_i\right)\right]}{\sum_{j\in C} K_{ij}\sum_{k\in C} K_{ij}\left(p_k - p_i\right)^2 - \left[\sum_{k\in C} K_{ik}\left(p_k - p_i\right)\right]^2} \qquad (26)$$

where $K_{ij} = K\left(p_j - p_i / h_n\right)$ and $K_{ik} = K\left(p_k - p_i / h_n\right)$. Heckman et al. (1997) argue that the advantages of local linear regression over using standard kernel weights include a faster rate of convergence near boundary points and greater robustness to different data densities.

***

As noted by Imbens (2004), there are no formal results for the variance of the propensity score matching estimators when the propensity score is unknown and needs to be estimated.[14] Therefore, a common procedure is to estimate standard errors by bootstrapping; however, the theoretical properties of bootstrap have not yet been established for matching estimators (albeit there is some evidence that bootstrapping does not lead to valid confidence intervals for some nearest-neighbour estimators; see Abadie and Imbens, 2006). Although further research is clearly needed so as to establish the reliability of bootstrapped standard errors for multiple matches-based estimators, the vast majority of empirical research to date has relied on bootstrapping as a feasible alternative for constructing confidence intervals in matching estimation settings.

Finally, given its importance for a credible propensity score matching estimation of treatment effects (and for matching procedures in general), the overlap assumption should ideally be assessed using the real data. With the propensity score as the only covariate, its distributions among treatment and comparison groups can be directly plotted and compared, but the validity of the result depends on an adequate

[14] When parametric modelling is performed, it is often the case that the estimate of the asymptotic variance of the treatment effect estimators ignores the fact that there is an error component associated to the estimation of the propensity score (and also to the ordering of the matching process itself), hence being considered a "conservative" estimator.

specification of the propensity score model. When a high dimensional matrix of covariates is being used, a good test of the overlap assumption is to verify the distribution of the covariates deemed important for the evaluation in both groups. If in fact there is lack of common support in one or more regions of the covariates distribution, it can be advisable to drop matches where individual covariates are far apart from each other – which, once again, may lead to a necessary redefinition of the parameter being estimated.

3.2.3. Matching and difference-in-differences

The identification assumption which underpins matching estimation requires that programme exposure and potential health outcomes are independent once the relevant covariates are controlled for. Thus, if there remain any systematic differences between the health outcomes of treated and comparison individuals, matching estimation will not recover the parameter of interest. This will occur, for example, if individuals select into the health programme based on unobserved characteristics which are themselves correlated with their potential outcomes (e.g., more health-concerned people, with "better" lifestyles, tend to migrate to areas where the health programme of interest is available exactly because of these services availability), or when there were differences on health endowments between the areas in which the programme was implemented and those where it was not.

However, even in the cases just described, semi-parametric estimation of average treatment effects can still be performed by relying on weaker identification assumptions than the ones described in previous sections. If we have reasons to think that there remain systematic differences between treated and comparison individuals even after matching, but it can be assumed that those differences are time-invariant, a difference-in-difference (DD) strategy can be adopted to eliminate the remaining biases.

One important characteristic of matching methods is that they are flexible enough to be combined with other estimation strategies such as DD, therefore allowing identification of the parameters of interest to be based on weaker, more credible assumptions for some settings. Moreover, a combined matching-DD approach has two advantages over the standard DD method: firstly, there is no

imposition of a linear relationship between treatment and health outcomes, with the linking function between health outcomes and treatment status allowed to change over time; secondly, comparison observations are reweighed according to their similarity – in terms of observable covariates – to their treated counterparts (Smith and Todd, 2005). One weakness of the DD approach is that usually there are no guidelines regarding which covariates, if any, should be included in the regression.

If t_1 denotes the time period after the programme was implemented, the evaluation problem for estimating the ATT can be expressed as estimating the counterfactual in:

$$E\left(Y_i^T - Y_i^C \mid D=1, t_1\right) = E\left(Y_i^T \mid D=1, t_1\right) - E\left(Y_i^C \mid D=1, t_1\right) \tag{27}$$

With two time periods, where t_0 represents the time period before exposure to the health programme, and under the assumption that the evolution of health outcomes in treatment and comparison groups would have been the same in the absence of the programme, (27) is equivalent to:

$$\begin{aligned} & E\left(Y_i^T \mid D=1, t_1\right) - \left[E\left(Y_i^C \mid D=1, t_0\right) + E\left(Y_i^C \mid D=0, t_1\right) - E\left(Y_i^C \mid D=0, t_0\right)\right] \\ & = E\left(Y_i^T \mid D=1, t_1\right) - E\left(Y_i^C \mid D=1, t_0\right) - \left[E\left(Y_i^C \mid D=0, t_1\right) - E\left(Y_i^C \mid D=0, t_0\right)\right] \\ & = E\left(Y_i^T \mid D=1, t_1\right) - E\left(Y_i^C \mid D=0, t_1\right) - \left[E\left(Y_i^C \mid D=1, t_0\right) - E\left(Y_i^C \mid D=0, t_0\right)\right] \end{aligned} \tag{28}$$

Reliable estimation of the last three terms in any line of (28) requires the comparison of health outcomes across similar groups. Consider the case of propensity-score matching estimation. Heckman et al. (1997) suggest the following difference-in-differences propensity score matching (PSDD) estimator for the ATT:

$$\hat{\beta}_{PSDD}^{ATT} = \frac{1}{N_T^*} \sum_{i \in \left\{T \cap S^*\right\}} \left(\left(Y_{it_1}^T - Y_{it_0}^C\right) - \sum_{j \in \left\{C \cap S^*\right\}} W_{ij} \left(Y_{jt_1}^C - Y_{jt_0}^C\right) \right) \tag{29}$$

The idea behind this estimation procedure is to eliminate any systematic time-invariant differences – e.g., health endowments or geographic mismatches – between groups exposed and unexposed to the health programme, conditional on the propensity score. This will be true only if a modified version of the weak unconfoundedness assumption holds:

ASSUMPTION 6 (WEAK UNCONFOUNDEDNESS FOR PSDD):

$$\left(Y_{t_1}^C - Y_{t_0}^C\right) \perp D \mid p(X).$$

Unconfoundedness as defined in Assumption 6 states that, conditional on the propensity score, it must be the case that comparison individuals have evolved in terms of their (average) health outcomes in the same way treated individuals would have evolved had they not received treatment; this is the matching-modified version of the *"parallel trend"* assumption invoked for standard DD estimation. This is weaker than the assumptions previously defined in this paper because now it is not a problem if there are unobserved factors which affect exposure to the programme and health outcomes simultaneously, as long as the effect of these unobserved variables exhibits the same variation, on average, for treated and comparison individuals. This guarantees that the following equality holds and can be used for estimation purposes:

$$E\left(Y_{t_1}^C - Y_{t_0}^C \mid P(X), D=1\right) = E\left(Y_{t_1}^C - Y_{t_0}^C \mid P(X), D=0\right)$$

The overlap assumption must hold in both periods: for a given propensity score value, there must be both comparison and treated observations in both periods. Also, as it can be easily noticed, panel data are required to estimate (29); however, it can be the case that *i)* panel data are not available at least for the comparison individuals, or *ii)* only repeated cross-sections data are available for both groups. Smith and Todd (2005) modify the estimator (29) to deal with the former case; for the latter case, Blundell and Costa-Dias (2000) propose a repeated cross-section PSDD estimator of the ATT that can be expressed as:

$$\hat{\beta}_{PSDD}^{ATT} = \frac{1}{N_{T_1}^*} \sum_{i \in \{T_1 \cap S^*\}} \left(\left(Y_{it_1}^T - \sum_{j \in \{T_0 \cap S^*\}} W_{ij} Y_{jt_0}^C \right) - \left(\sum_{j \in \{C_1 \cap S^*\}} W_{ij} Y_{jt_1}^C - \sum_{j \in \{C_0 \cap S^*\}} W_{ij} Y_{jt_0}^C \right) \right) \quad (30)$$

In (30), T_0, T_1, C_0 and C_1 stand for the treatment and comparison groups in the periods before and after exposure to the health programme, respectively. Therefore, the estimator proposed by Blundell and Costa-Dias (2000) requires matching to be performed three times for each treated individual: to find comparable treated individual(s) before treatment and comparison individuals before and after the programme.

As to the relevant propensity score to be used when performing the matching procedure in a repeated cross-sections context, an alternative is proposed by Blundell et al. (2002). Since there are two non-random assignments – to treatment and time of observation – and the distribution of covariates must be the same in the four cells defined by combining these assignments, the authors use a vector of two propensity scores (one for each assignment category) as the matching variable. Once the three

counterfactuals have been constructed by a chosen matching procedure, the programme ATT is estimated by DD as in (30) under the additional assumptions of additive separability of the group and time effects.

4 Matching estimation of health programme average treatment effects with multiple treatments

The matching methods discussed so far were originally developed for assessing the impact of social programmes when the treatment variable is defined as a binary category: the individual is *either* in the active treatment group *or* in the comparison group (the latter usually defined as the "untreated" case). There has been though some recent interest in investigating the potential that matching methods have for jointly evaluating the impact of alternative programmes, mainly in labour market policies settings (see, for instance, Lechner, 2002; Frölich et al, 2004). This calls of course for an extension of the matching procedures described above so as to contemplate the case of multiple alternative interventions.

Obviously, it can be very relevant to evaluate the impact of alternative programmes in the context of health interventions as well. Health programmes aimed at improving the same health outcomes may contain different components (for example, offering a different mix of health services) or the same health intervention may be available in several geographic areas, but with different population coverage levels across these localities. Another evaluation setting could require the impact assessment of the same health programme when individuals have been exposed to the intervention for different periods of time. All the settings just described can be seen as involving a comparison between the health impacts of alternative "programmes", where for instance coverage levels and length of exposure play the role of compared alternatives.

The required extensions of matching procedures for the case of multiple treatments have been proposed almost simultaneously in two papers, Imbens (2000) and Lechner (2000). They show that the main results obtained for the case of matching with a single treatment apply for the multivariate case as well. I will present these developments relating them to the aforementioned health sector relevant case of

a health programme with different levels of population coverage between localities, the so-called “dose-response” context.

4.1 Average treatment effects: definition and identification

4.1.1. Definition of average treatment effects

Applying the definitions introduced by Lechner (2000) to a more specific context, let a given health programme be implemented in a group of localities according to sequentially increasing, mutually exclusive coverage levels *l* denoted by $l = \{0,1,2,...,L\}$. A given individual *i* who lives in a locality with a coverage level *l* will have then only one element of the outcomes set $\{Y^0, Y^1, Y^2, ..., Y^L\}$ observed at a given point in time, the remaining being her counterfactual outcomes. The treatment variable *D* can now assume one of *(L+1)* discrete values: $D \in \{0,1,2,...,L\}$.

The average treatment effects of interest defined for the single treatment case are expanded so as to encompass the availability of multiple treatments, although the focus remains on pair wise comparisons between the health effects of two different coverage levels, e.g., $l_1 = 1$ and $l_2 = 2$, $l_2 > l_1$. The causal effects of interest are now related to the difference $Y_i^2 - Y_i^1$, that is, the effect of being exposed to treatment level 2 and not being exposed to treatment level 1.

Average Treatment Effect (ATE): this is the expected health effect of living in a locality with a coverage level $l = 2$ instead of living in an area with $l = 1$ for an individual randomly drawn from the entire population.

$$ATE^{2,1} = E\left[Y^2 - Y^1\right] = E\left[Y^2\right] - E\left[Y^1\right] \tag{31}$$

Note that the average treatment effect of being subjected to coverage level 2 instead of 1 is symmetric to the treatment effect of being subjected to the latter instead of the former, i.e. $ATE^{2,1} = -ATE^{1,2}$. Note also that $E\left[Y^{l_2} - Y^{l_1}\right]$ involves the computation of the expected value of every counterfactual outcome for that particular pair wise comparison, as can be seen by expanding (31):

$$
\begin{aligned}
ATE^{2,1} = E\left[Y^2 - Y^1\right] = E\left[Y^2 - Y^1 \mid D=0\right]P(D=0) + E\left[Y^2 - Y^1 \mid D=1\right]P(D=1) + \\
+ E\left[Y^2 - Y^1 \mid D=2\right]P(D=2) + ... + \\
+ E\left[Y^2 - Y^1 \mid D=L\right]P(D=L) \\
= \sum_{l=0}^{L} E\left[Y^2 - Y^1 \mid D=l\right]P(D=l)
\end{aligned}
\tag{32}
$$

Thus, in the matching estimator case, the ATE can be obtained by estimating the expected outcomes conditional on covariates in both groups and weighting them by the distribution of these covariates in the full sample:

$$
\begin{aligned}
ATE^{2,1} &= E\left[Y^2\right] - E\left[Y^1\right] \\
&= E\left[E\left(Y^2 \mid X\right)\right] - E\left[E\left(Y^1 \mid X\right)\right] \\
&= \int \left[E\left(Y^2 \mid X = x, D=2\right) - E\left(Y^1 \mid X = x, D=1\right)\right] \cdot f_X(x)\,dx
\end{aligned}
\tag{33}
$$

where $f_X(x)$ denotes the density of *X* in the whole sample. A modified version of the unconfoundedness assumption is necessary for (33) to be valid; this assumption is rigorously defined in the next sub-section.

Lechner (2000) suggests also a redefinition of the ATE so as to refer only to the population exposed to the two coverage levels of a given pair wise comparison. This *"restricted ATE"* is similar to the definition of the ATE when the treatment variable is binary:

$$
ATE_R^{2,1} = E\left[Y^2 - Y^1 \mid D = 1,2\right] = E\left[Y^2 \mid D = 1,2\right] - E\left[Y^1 \mid D = 1,2\right] \tag{34}
$$

The above treatment effect is also symmetric in the sense that $ATE_R^{2,1} = -ATE_R^{1,2}$.

Average Treatment Effect on the Treated (ATT): this parameter corresponds to the average effect among those who reside in a locality with a coverage level $l = 2$ when compared to those who live in a locality with coverage level $l = 1$.

$$
ATT^{2,1} = E\left[Y^2 - Y^1 \mid D = 2\right] = E\left[Y^2 \mid D = 2\right] - E\left[Y^1 \mid D = 2\right] \tag{35}
$$

Thus, in the context of a health programme with multiple possible levels of coverage, the ATT is equivalent to the marginal gain (in terms of health outcomes) accruing to a randomly selected individual from a locality with coverage level 2, relative to what would have been her outcome if she had lived in a locality with coverage level 1. Again, under unconfoundedness, the ATT can be identified in the matching estimation context as:

$$
\begin{aligned}
ATT^{2,1} &= E\left[Y^2 - Y^1 \mid D=2\right] = E\left[Y^2 \mid D=2\right] - E\left[Y^1 \mid D=2\right] \\
&= E\left[Y^2 \mid D=2\right] - E\left[E\left(Y^1 \mid X, D=1\right) \mid D=2\right] \\
&= E\left[Y^2 \mid D=2\right] - \int E\left(Y^1 \mid X=x, D=1\right) \cdot f_{X|D=2}\left(x\right) dx
\end{aligned} \tag{36}
$$

where $f_{X|D=2}\left(x\right)$ denotes the density of *X* only among individuals exposed to treatment level 2.

The symmetry property for the ATE measures mentioned above does not necessarily hold in the case of the ATT. To see this, notice that $ATT^{1,2} = E\left[Y^1 - Y^2 \mid D=1\right] = E\left[Y^1 \mid D=1\right] - E\left[Y^2 \mid D=1\right]$, which is different from $\left(-ATT^{2,1}\right)$ because the conditioning sets differ. This difference will hold if individuals living in the compared areas systematically differ in a way related to their health outcomes.

Summing up, as in the case of a single treatment, estimating the ATT is less demanding in terms of required information than estimating the ATE (even if the latter is pair wise restricted). Estimating $ATT^{2,1}$ requires only identification of $E\left[Y^1 \mid D=2\right]$; estimation of $ATE_R^{2,1}$ requires identification of two counterfactuals, $E\left[Y^1 \mid D=2\right]$ and $E\left[Y^2 \mid D=1\right]$. The more demanding situation is that of estimation of the ATE, for which the series of counterfactuals $E\left[Y^1 \mid D \neq 1\right]$ and $E\left[Y^2 \mid D \neq 2\right]$ need to be identified.[15]

4.1.2. Identification of average treatment effects

The treatment effects defined above can be consistently estimated using matching methods even in a multiple treatment setting. Analogously to the single treatment case, identification is based on two fundamental assumptions about treatment – or, in our case, coverage level – assignment: *unconfoundedness* and *overlap*. Let $p^l\left(X\right)$ be the individual probability of being assigned to coverage level *l*

[15] Since many alternative comparisons of programmes are possible when multiple treatments are available, Lechner (2001) suggested further treatment effects parameters which are not going to be discussed here. In the present setting, these composite measures would refer to the average treatment effect for an individual of being exposed to coverage level *l* compared to the treatment effect of being randomly assigned to any of the other available coverage levels with the probabilities valid in the population.

given the vector of individual covariates *X*; then, the two fundamental assumptions are:

ASSUMPTION 7 (UNCONFOUNDEDNESS FOR MULTIPLE TREATMENTS):

$$\left(Y^0, Y^1, Y^2, ..., Y^L\right) \perp D \mid X. \tag{37}$$

ASSUMPTION 8 (OVERLAP FOR MULTIPLE TREATMENTS):

$$0 < p^l\left(X\right) < 1, \ \forall l \in \left\{0, 1, 2, ..., L\right\}. \tag{38}$$

Unconfoundedness given covariates for the multiple treatments case (37) is just an extension of the same assumption (in its stronger version) for the single treatment case. Lechner (2000) proves for the multiple treatments case that, if unconfoundedness holds given covariates, it also holds when conditioning solely on a particular function of the covariates, the *generalised balancing score*:

$$\left(Y^0, Y^1, Y^2, ..., Y^L\right) \perp D \mid X = x \ \Rightarrow \ \left(Y^0, Y^1, Y^2, ..., Y^L\right) \perp D \mid b\left(X\right) = b\left(x\right) \tag{39}$$

which is true if $E\left[P\left(D = l \mid X = x\right) \mid b\left(X\right) = b\left(x\right)\right] = P\left(D = l \mid X = x\right) = p^l\left(x\right)$. By using the same reasoning as in the single treatment case, under unconfoundedness we must have:

$$E\left[P\left(D = l \mid X\right) \mid Y^0, Y^1, ..., Y^L, b\left(X\right)\right] = E\left[P\left(D = l \mid X\right) \mid b\left(X\right)\right]$$
$$= P\left(D = l \mid b\left(X\right)\right) = P\left(D = l \mid X\right)$$

and a valid balancing score is the vector of all but one (the linearly independent) individual propensity scores $\vec{p}\left(x\right) = \left[p^1\left(x\right), p^2\left(x\right), ..., p^L\left(x\right)\right]$. All the relevant counterfactuals – and therefore all the relevant treatment effects – are identified by relying on (37) and (39). As in the single treatment case, the mechanical balancing score property is combined with unconfoundedness so as to make valid the propensity score matching approach, although now the conditioning set $\vec{p}\left(X\right)$ is not of single dimension anymore.

An important result derived by Lechner (2000) states that the evaluation problem with multiple treatments is substantially simplified when the interest lies in pair wise comparisons, e.g. when separately estimating the average health effect of increasing the coverage level of a given programme from l_0 to l_1 , l_1 to l_2 , and so forth. Weaker versions of the basic assumptions (37) and (38) are now required, and a reduction of the conditioning set to one is possible. To see this, let the treatment effect of interest be that associated to increasing the programme's coverage level from $l = 1$

to $l=2$. In this case, for identifying the ATE, ATE_R and ATT, the sufficient assumptions are, respectively:

(i) Weak unconfoundedness and overlap assumptions for estimating the ATE:

$$Y^2, Y^1 \perp D \mid X = x \Rightarrow Y^2, Y^1 \perp D \mid \left[p^1(X) = p^1(x), ..., p^L(X) = p^L(x) \right] \quad (40)$$

$$0 < p^l(x) < 1, \ \forall l \in \{0, 1, 2, ..., L\} \quad (41)$$

(ii) Weak unconfoundedness and overlap assumptions for estimating the ATE_R:

$$Y^2, Y^1 \perp D \mid X = x, D \in \{1,2\} \Rightarrow Y^2, Y^1 \perp D \mid p^{1|1,2}(X) = p^{1|1,2}(x), D \in \{1,2\} \quad (42)$$

where $p^{1|1,2}(x) = P(D=1 \mid D \in \{1,2\}, X = x) = \dfrac{p^1(x)}{p^1(x) + p^2(x)}$ is the *generalised propensity score*, and:

$$0 < p^l(x) < 1, \ l \in \{1,2\} \quad (43)$$

Notice that the latter case is similar to that of a binary treatment variable for which $p^1(x) + p^2(x) = 1$, but recall that, in the general case of multiple treatments, $p^1(x) + p^2(x) < 1$.

(iii) Weak unconfoundedness and overlap assumptions for the $ATT^{2,1}$:

$$Y^1 \perp D \mid X = x, D \in \{1,2\} \Rightarrow Y^1 \perp D \mid p^{1|1,2}(X) = p^{1|1,2}(x), D \in \{1,2\} \quad (44)$$

$$0 < p^l(X) < 1, \ l \in \{1,2\} \quad (45)$$

With the set of assumptions *(i)*, all relevant counterfactuals for the $ATE^{2,1}$ and $ATE^{1,2}$ are identified, because it is implied that:

$$E\left[Y^2 \mid X = x, D = l\right] = E\left[Y^2 \mid X = x, D = 2\right], \forall l \neq 2$$

$$E\left[Y^1 \mid X = x, D = l\right] = E\left[Y^1 \mid X = x, D = 1\right], \forall l \neq 1$$

Notice that the conditioning sets above are still of L dimension. Unconfoundedness is relaxed in *(ii)* by referring only to assignment to the pair of compared treatment levels 1 and 2 (and their respective subpopulations); if both potential outcomes of interest are independent of assignment to any coverage level l as stated in *(i)*, then it implies that the same is true when comparing only the groups of individuals assigned to any pair of treatment levels. Thus, *(ii)* is a logical implication of *(i)* and identification is based on the following equalities:

$$E\left[Y^2 \mid p^{1|1,2}(X) = p^{1|1,2}(x), D = 1\right] = E\left[Y^2 \mid p^{1|1,2}(X) = p^{1|1,2}(x), D = 2\right]$$

$$E\left[Y^1 \mid p^{1|1,2}\left(X\right) = p^{1|1,2}\left(x\right), D=2\right] = E\left[Y^1 \mid p^{1|1,2}\left(X\right) = p^{1|1,2}\left(x\right), D=1\right]$$

In this case, the counterfactual expected outcome can be consistently identified by adjusting for the adequate distribution of the (single dimensional) generalised propensity score $p^{1|1,2}\left(X\right)$. For example, for the $ATT^{2,1}$:

$$\begin{aligned} &E\left[Y^1 \mid D=2\right] = E\left[E\left[Y^1 \mid p^{1|1,2}\left(X\right), D=1\right] \mid D=2\right] \\ &= \int E\left[Y^1 \mid p^{1|1,2}\left(X\right), D=1\right] \cdot f_{p^{1|1,2}|D=2}\left(p^{1|1,2}\right) dp^{1|1,2} \end{aligned} \tag{46}$$

where $f_{p^{1|1,2}|D=2}$ corresponds to the distribution of the generalised propensity score in the sub-sample in the active treatment, i.e. coverage level 2.

Conditioning can be based on the generalised propensity score, but obviously a finer balancing score can also be used, including combinations of the generalised propensity score and covariates deemed to be particularly relevant in a given context. It is also valid to condition directly on the vector $\left[p^1\left(x\right), p^2\left(x\right)\right]$, which is finer than the conditional probability $p^{1|1,2}\left(x\right)$ in the sense that the latter is equivalent to its expectation conditional on $p^1\left(x\right)$ and $p^2\left(x\right)$, that is:

$$E\left[p^{1|1,2}\left(X\right) \mid p^1\left(X\right), p^2\left(X\right)\right] = E\left[\frac{p^1\left(X\right)}{p^1\left(X\right)+p^2\left(X\right)} \mid p^1\left(X\right), p^2\left(X\right)\right] = p^{1|1,2}\left(X\right)$$

In each case, the parameters ATE_R and ATT can now be estimated by relying on the set of assumptions *(ii)*, as $E\left[Y^1 \mid D=2\right]$ and $E\left[Y^2 \mid D=1\right]$ are identifiable counterfactuals.

If, however, interest lies only in estimating the $ATT^{2,1}$, the set of assumptions *(iii)* suffices for identifying the required counterfactual $E\left[Y^1 \mid D=2\right]$ by relying on the equality $E\left[Y^1 \mid p^{1|1,2}\left(X\right) = p^{1|1,2}\left(x\right), D=2\right] = E\left[Y^1 \mid p^{1|1,2}\left(X\right) = p^{1|1,2}\left(x\right), D=1\right]$. Recall that $ATT^{2,1} = E\left[Y^2 \mid D=2\right] - E\left[Y^1 \mid D=2\right]$ and that, by unconfoundedness, the second expectation term is equal to $E\left[E\left[Y^1 \mid p^{1|1,2}\left(X\right), D=1\right] \mid D=2\right]$.

A careful examination of the statements above leads to two important implications. Firstly, a *sample reduction property* for pair wise comparisons of treatment levels is derived (Lechner, 2000). If coverage levels 1 and 2 are being compared and the interest lies in estimating only the parameters ATE_R and ATT,

unconfoundedness can be assumed to hold only for the sub-sample of individuals subjected to treatment levels 1 and 2, implying that this sub-sample is the only one required for the empirical analysis; in other words, individuals exposed to treatment levels $l \neq \{1,2\}$ – and thus the existence of multiple treatments – can be ignored for this particular analysis.

Secondly, a *conditioning set reduction* is achieved when making pair wise comparisons and estimating the ATE_R and ATT parameters. Propensity score matching can be based on the single dimension conditioning set $p^{1|1,2}(X)$, a composite individual index. Importantly, Imbens (2000) and Lechner (2000) show that, for the pair wise comparison case, a similar reduction can also be derived for estimating the ATE, which involves the computation of $E\left[E\left[Y^l \mid p^l(X), D=l\right] \mid D \neq l\right]$. Also in this case, $p^{1|1,2}(X)$ and the vector of propensity scores $\vec{p}(x)=\left[p^1(x), p^2(x)\right]$ are valid balancing scores.

Finally, variance bounds for the estimators of the expected potential outcome $E\left[Y^l\right]$, the ATE $E\left[Y^2-Y^1\right]$, the ATT $E\left[Y^2-Y^1 \mid D=2\right]$ and the mean counterfactual outcome $E\left[Y^1 \mid D=2\right]$ have been derived by Frolich (2004).

4.2 Matching estimation of average treatment effects

4.2.1. Matching on the propensity score

Using the same notation as in Section 3, estimators of the treatment effects discussed above can be expressed as follows:

$$\widehat{ATT}^{2,1} = \frac{1}{N_2}\sum\nolimits_{i\in\{D=2\}} Y_i - \frac{1}{N_2}\sum\nolimits_{j\in\{D=1\}} W_{ij} Y_j \tag{47}$$

$$\widehat{ATE_R}^{2,1} = \widehat{ATT}^{2,1}\cdot P\left(D=2 \mid D\in\{1,2\}\right) - \widehat{ATT}^{1,2}\cdot P\left(D=1 \mid D\in\{1,2\}\right) \tag{48}$$

$$\widehat{ATE}^{2,1} = \sum_{l=0}^{L}\left[\left(\frac{1}{N_l}\sum\nolimits_{i\in\{D=2\}} W_{il} Y_i - \frac{1}{N_l}\sum\nolimits_{j\in\{D=1\}} W_{il} Y_j\right) P\left(D=l\right)\right] \tag{49}$$

where *i* indexes an individual belonging to the active treatment group and *j* denotes a matched individual belonging to the comparison treatment group.

The actual estimation of average treatment effects with multiple treatments by propensity score matching (PSM) follows the same general steps performed as in the single treatment case. Consistent estimates of $P(D=l)$ and $E\left[Y^l \mid D=l\right]$ are obtained, respectively, by cell frequencies and the average outcomes for individuals exposed to treatment level *l*. The PSM estimator of the counterfactual can then be constructed according to the steps suggested by Lechner (2000):[16]

1) Specify and estimate a multiple choice model in order to obtain the individual vectors of (generalised) propensity scores (one for each individual): $\left[\hat{p}_i^0(X), \hat{p}_i^1(X), \hat{p}_i^2(X), ..., \hat{p}_i^L(X)\right]$. For the particular case of different coverage levels of the same health programme, an ordered choice model would be appropriate, but other settings may rather call for a multinomial choice model;

2) Estimate the counterfactual expectations of the outcome variables conditional on the respective propensity scores. In the case of investigating the ATT of being subjected to treatment level 2 instead of 1, having already computed in the first step, for each individual, $\hat{p}^{1|1,2}(X) = \dfrac{\hat{p}^1(X)}{\hat{p}^1(X) + \hat{p}^2(X)}$ or $\left[\hat{p}^1(X), \hat{p}^2(X)\right]$, this is achieved by:

a) choosing one individual from the sub-sample belonging to $D=2$ and temporarily excluding her from the sample;

b) finding an individual in the sub-sample $D=1$ who is the closest one to the individual chosen in *a)*, either in terms of $\hat{p}^{1|1,2}(X)$ or in terms of the vector $\left[\hat{p}^1(X), \hat{p}^2(X)\right]$. Obviously, this is to be preceded by the definition of a "closeness" measure for the analysis, such as the Euclidean or Mahalanobis distances. The comparison individual chosen in this step will be replaced in the corresponding sub-sample to allow her possible re-use as a match to other $D=2$ individuals;

[16] Lechner relies on nearest-neighbour with replacement as the matching procedure, although the suggested protocol can be easily adapted to make use of the alternative PSM procedures discussed for the single treatment case. However, since the role of each sub-sample as treatment and comparison group is reversed for estimating all the treatment effects parameters, it is necessary to rely on matching with replacement when the number of individuals is different in the treatment level sub-samples.

c) repeating steps *b)* and *c)* until all individuals belonging to the $D=2$ group are matched to a comparison counterpart;

d) using the sample mean of the outcomes in the resulting comparison group formed above (which may contain repeated comparison individuals as matches) to compute the counterfactual conditional expectation. The conditional expectation $E\left[Y^2 \mid D=2\right]$ is estimated as the mean outcome in the sub-sample $D=2$;

3) Repeat step *2)* for all the relevant pair wise combinations of l_1 and l_2 (if estimating average treatment effects for additional pair wise comparisons of other treatment levels);

4) Calculate the estimated average treatment effect of interest using the results obtained in steps *2)* and *3)*;

5) Obtain the estimated variance of treatment effects. For the parameter $ATT^{2,1}$, for instance, Lechner (2002) suggests calculating the variance of $\hat{E}\left[Y^1 \mid D=2\right]$ by $\frac{\sum_{j\in\{D=1\}}\left(\hat{W}_{ij}\right)^2}{\left(N_2\right)^2}\times\widehat{Var}\left(Y \mid D=1\right)$, and the variance of $\hat{E}\left[Y^2 \mid D=2\right]$ by $\frac{\widehat{Var}\left(Y \mid D=2\right)}{N_2}$. The term $\widehat{Var}\left(Y \mid D=l\right)$ denotes the empirical variance in the respective sub-sample, N_2 represents the number of individuals in the $D=2$ group and $\hat{W}_{ij}$ denotes the number of times observation $j\in\{D=1\}$ appears in the comparison group formed for the counterfactual estimation. The estimated variance of the treatment effect will be given by the summation of the two variance terms defined above. Another alternative is to use bootstrapping.

As mentioned in step *1)*, two possibilities emerge for estimating the generalised propensity score or selection probabilities (Lechner, 2002). The first alternative is to specify and estimate each conditional binary choice equation separately – using probit or logit models – to obtain the pair wise conditional probabilities $\hat{p}_i^{l|l_1,2}\left(x\right)$. Estimation of each binary choice equation requires only data for individuals belonging to the sub-samples involved in the corresponding pair wise comparison; therefore, if all possible

pair wise comparisons are of interest, this procedure will have to be repeated $L(L-1)/2$ times (where L denotes the total number of treatment levels). One advantage of this procedure is that it does not require the "independence of irrelevant alternatives" assumption; also, since the conditional probabilities are not interdependent, misspecification of one binary choice equation does not imply that all conditional probabilities are misspecified (and different sets of regressors might in principle be used across the binary models). A drawback of the method is the fact that, with many treatment levels, many binary choice models will have to be estimated and interpreted, one for each pair wise comparison.

A second possibility for obtaining the generalised propensity score is to specify a choice problem incorporating all the possible treatment levels and estimating it in one step using the full sample, through a multinomial/ordered choice model. In the multinomial case, a probit specification might be preferred because, unlike the logit, it does not rely on the "independence of irrelevant alternatives" assumption. Moreover, compared to the estimation of several binary choice models, it seems richer in the sense that it allows investigation of the relevant factors which determine individual selection into alternative treatment levels. An important drawback of this procedure is, however, its somewhat restrictive character: restrictions on the covariance matrix of error terms need to be imposed, and there is the danger of misspecification of all conditional probabilities if one choice equation is misspecified, since the derived conditional probabilities are interdependent. Furthermore, the existence of more than four treatment alternatives makes necessary the utilisation of simulated maximum likelihood methods in order to approximate the results of a multinomial probit model.[17]

Checks should be performed regarding the common support condition. In the pair wise comparison setting between treatment levels 1 and 2, this means that overlap must be observed between the distributions of $\hat{p}^{1|1,2}(x)$ or $\left[\hat{p}^{1}(x), \hat{p}^{2}(x)\right]$. In cases where all the pair wise average treatment effects are of interest, it can be good practice to restrict the estimation over the joint common support – the overlap region

[17] The empirical application of Lechner (2002) found that the correlation of the conditional probabilities obtained via multinomial and binary choices models was very high (between 0.980 and 0.998). Thus, in his context, no significant differences in the evaluation results should be expected using one or another approach, although no generalisation of such a result is warranted for other settings.

of the distribution of propensity scores across all treatment levels – such that all treatment effects refer to the same sub-population and no comparability problems arise.

The balancing of covariates across treatment and comparison groups is essential for the reliability of treatment effects estimates. Although no clear procedure dominates for checking which of the alternatives for obtaining the conditional probabilities (namely, matching on the vector of propensity scores or on the generalised propensity score obtained either by a multinomial or binary choice models) does a better job in balancing the covariates, some possibilities are the Mahalanobis distance and the (median absolute or mean squared) standardised bias. The latter is used by Lechner (2002) for measuring the matching quality through the distance between the marginal distributions of the relevant covariates.

Additionally, simple diagnostic checks can be performed regarding the (average) number of times a comparison observation is used as match in a given pair wise comparison, when using each of the three alternatives mentioned above for performing the matching. One way of doing that is to compute, for each of the three procedures, the mean of the weights for matched comparison observations, with weights being defined as the number of treated individuals a given comparison individual is matched to. Better matching procedures will use more comparison observations as matches without loss from the point of view of covariates balancing, therefore leading to smaller estimated standard errors of treatment effects.

4.2.2. Propensity score matching and difference-in-differences

As in the single treatment case, the identification assumption behind propensity score coupled with difference-in-differences (PSDD) is weaker than the unconfoundedness assumption used for simple propensity score matching, because the former allows the bias – between active and comparison treatment groups – to be different from zero, requiring only bias stability over time. Estimating average treatment effects by a PSDD approach is a feasible option due to the mechanical validity of the balancing property of the (generalised) propensity score, in the sense that the balancing of covariates is achieved whether or not unconfoundedness holds. In other words, the resulting equality of counterfactuals (36) and (46) can be used in a

difference-in-differences estimation context when unconfoundedness does not seem a plausible identification assumption.

In the multiple treatments context, the difference-in-differences approach recovers the treatment effect of being exposed to treatment level *l* versus the non-participation version of treatment status, say $l=1$(being equivalent to "no coverage"); it generally does not allow comparisons between alternative treatment levels because pre-programme health outcomes provide information only about untreated outcomes (Frolich, 2004). Thus, with panel data, the necessary counterfactual for identifying the $ATT^{l,0}$ in a difference-in-differences framework can be obtained by using PSM and conditioning only on the generalised propensity score:

$$\begin{aligned} & E\left[E\left[Y_{t_1}^0 - Y_{t_0} \mid X, D=0\right] \mid D=l\right] = E\left[E\left[Y_{t_1}^0 - Y_{t_0} \mid p^{0|l,0}(X), D=0\right] \mid D=l\right] \\ & = \int E\left[Y_{t_1}^0 \mid p^{0|l,0}(X), D=0\right] \cdot f_{p^{0|l,0}|D=l}\left(p^{0|l,0}\right) dp^{0|l,0} \\ & \quad - \int E\left[Y_{t_0} \mid p^{0|l,0}(X), D=0\right] \cdot f_{p^{0|l,0}|D=l}\left(p^{0|l,0}\right) dp^{0|l,0} \end{aligned} \tag{50}$$

where $Y_{t_1}^0$ denotes the health outcome of a matched untreated (i.e. not covered by the programme) individual in the post-programme period, Y_{t_0} represents the health outcome of the same individual in the pre-programme period and $p^{0|l,0}$ denotes the conditional probability of not being exposed to the programme within the sub-sample of uncovered individuals and those exposed to the treatment level of interest *l*.

5 Matching estimation of health programme average treatment effects with a continuous treatment

When the interest lies in evaluating the impact of only one health programme for which data are available on several different levels of population coverage across geographic regions, a more sensible evaluation strategy can be the utilisation of propensity score methods adapted to the continuous treatment case. In this particular context, the treatment variable might not be amenable to be naturally discretised: the definition of a number of discrete categories can be a very arbitrary process, the magnitude of the estimated treatment effects (and therefore of other parameters, such as the estimated "optimum" coverage level) can be sensitive to the criteria used for definition of categories, and information regarding treatment effects within each

category is lost. Nevertheless, the observed variation in exposures to coverage levels can still be used to identify the programme impact.

It has been only recently that a methodological contribution for the continuous treatment context has been advanced in the literature by Hirano and Imbens (HI, 2004), who discuss and illustrate the application of propensity scores in a regression context for evaluating the impact of continuous treatments. Following their work, a brief discussion of matching procedures in the continuous treatment case is presented by Flores (2004) and another empirical application of the HI method is performed by Aguero et al. (2006).

The discussion herein adapts the suggestions made by HI (2004) to the different case of a health programme with several different levels of population coverage across geographic areas. Although the authors suggest a complete parametric procedure which does not involve matching methods, I discuss some alternatives for evaluating such programmes through a semi-parametric approach. These alternatives apply to the more general case of treatments offered in different dosages.

5.1 Average treatment effects: definition and identification

5.1.1. Definition of average treatment effects

As in the preceding discussion, let $Y_i(l)$ be the set of potential outcomes for individual $i = 1,...,N$, or in other words the individual *dose-response function* according to all possible coverage levels (doses) $l \in [l_{\min}, l_{\max}]$. For each individual, we observe the coverage level to which they were actually exposed, $L_i \in [l_{\min}, l_{\max}]$, the associated potential outcome corresponding to that particular coverage level $Y_i = Y_i(L_i)$ and their vector of covariates X_i.Due to the missing data problem inherent to programme evaluation, the research question of interest is to identify the curve of average potential outcomes; that is, the parameter of interest is the *average dose-response function*:

$$\mu(l) = E\{Y_i(l)\} \tag{51}$$

which represents the function of the average potential health outcomes computed over all possible programme coverage levels (e.g., 0% to 100%).

Since the treatment variable is modelled as continuous, other policy-relevant parameters can be estimated apart from the entire average dose-response curve. For example, as suggested by Flores (2004), the coverage level at which the expected health outcome is maximised can be derived, along with the corresponding maximum value of this outcome. The *optimal treatment level* – i.e. the treatment level that maximises the expected health outcome[18] – can be expressed as:

$$l^* = \arg \max_{l \in [l_{\min}, l_{\max}]} E\left\{Y_i(l)\right\} \tag{52}$$

and the corresponding *expected potential health outcome at the optimal treatment level* would then be:

$$\mu(l^*) = E\left[Y_i(l^*)\right] \tag{53}$$

Additionally, as in HI (2004), we might be interested in constructing the curve for the *marginal impact of each coverage level* of interest on the health outcome, by pair wise calculating programme average health effects for every observed "jump" in the value of the coverage level. This parameter can be defined as:

$$\Delta(l_1) = \mu(l_1) - \mu(l_0) = E\left[Y_i(l_1)\right] - E\left[Y_i(l_0)\right],\ l_1, l_0 \in [l_{\min}, l_{\max}]. \tag{54}$$

In the latter case, l_0 might alternatively be a reference coverage level (e.g., no treatment or the smallest positive coverage observed in a given context) to which all other coverage levels are compared. If l_0 is the lowest possible coverage level (e.g. zero) and l^* is the optimal treatment level as defined in (52), the computation of (54) leads to an estimate of the *maximum individual health gain* that can be expected from the intervention:

$$\Delta^* = \mu(l^*) - \mu(l_0) = E\left[Y_i(l^*)\right] - E\left[Y_i(l_0)\right],\ l^*, l_0 \in [l_{\min}, l_{\max}]. \tag{55}$$

5.1.2. Identification of average treatment effects

Similarly to the single and multiple treatments settings, identification of the required counterfactuals can be achieved by relying on (suitably adapted) "selection on observables" assumptions (HI, 2004).

ASSUMPTION 9 (WEAK UNCONFOUNDEDNESS FOR A CONTINUOUS TREATMENT):

[18] The use of the word "optimal" in this context is very specific to the definition given in the text – that is, in terms of maximum health effects – and does not encompass considerations regarding any costs incurred to achieve such effects.

$$Y_i(l) \perp L_i \mid X_i, \text{ all } l \in [l_{\min}, l_{\max}]. \tag{56}$$

This is a weak version of the unconfoundedness assumption because it does not require the joint independence of all potential outcomes $\{Y(l)\}_{l\in[l_{\min},l_{\max}]}$; rather, it requires pair wise conditional independence for each of the potential health outcomes at a given treatment level with the actual treatment assignment.[19] Therefore, this assumption states that there are no unobserved factors that affect both individual potential health outcomes and the coverage levels to which individuals have been exposed, given the pre-treatment covariates.

For propensity score matching procedures, it is necessary to redefine the propensity score so as to take into account the continuous nature of the treatment being evaluated. Let the conditional density of the programme coverage given covariates be expressed as:

$$r(l,x) = f_{L|X}(l \mid x) \tag{57}$$

The *generalised propensity score (GPS)* – that is, the conditional density of the coverage level given pre-treatment covariates – is then given by:

$$R_i = r(L_i, X_i) \tag{58}$$

HI (2004) show that, by the standard results presented before, the GPS for the continuous case is also a balancing score in the sense that, within strata with the same value of $r(l,X)$, the probability that $L_i = l$ for a given individual does not depend on the value of their covariates, it is a random event. This is again a mechanical result and does not require unconfoundedness. But HI (2004) also show that *weak unconfoundedness given pre-treatment covariates implies the same result given the generalised propensity score*:

$$\begin{aligned} &Y(l) \perp L \mid X \Rightarrow Y(l) \perp L \mid r(l,X), \text{ all } l \in [l_{\min}, l_{\max}] \\ &\Leftrightarrow f_L(l \mid r(l,X), Y(l)) = f_L(l \mid r(l,X)), \text{ all } l \in [l_{\min}, l_{\max}] \end{aligned} \tag{59}$$

The last term in (59) is the conditional density of coverage level *l* (this is analogous to the probability of an individual being exposed to the particular coverage level *l* in the discrete treatment case), given the GPS evaluated at that same coverage level. Propensity scores will then be calculated for all observed coverage levels, but only one will be used at one time.

[19] However, as stressed by Imbens (1999), in practice it would be difficult to find a situation where the weak unconfoundedness assumption should be valid but not its strong version.

As a corollary of the unconfoundedness assumption (56) coupled with the GPS result given by (59), we can write the fundamental result for propensity score matching:

$$\begin{aligned} E\left[Y(l)\right] &= E\left[E\left[Y(l) \mid r(l,X) = r\right] \mid r(l,X)\right] \\ &= E\left[E\left[Y(l) \mid L_i = l, r(l,X) = r\right] \mid r(l,X)\right] \\ &= E\left[E\left[Y \mid L_i = l, R_i = r\right] \mid r(l,X)\right], \text{ all } l \in [l_{\min}, l_{\max}]. \end{aligned} \tag{60}$$

Notice again that the outer average is taken over the GPS evaluated at l (and not over $r(L_i, X_i)$). As usual, this averaging procedure is performed in order to control for systematic differences in the observed covariates across groups of individuals exposed to different treatment levels. The components of (60) can be estimated with the observed data and a logical estimation procedure would be to firstly regress the observed health outcome on the observed individual treatment level exposure and its corresponding GPS, and then taking the expectation of that regression over the GPS evaluated at each relevant treatment level.

5.2 Propensity score matching estimation of average treatment effects

HI (2004) suggest an entirely parametric procedure for estimating the average dose-response function, as opposed to the semi-parametric approach of propensity score matching in the case of single or multiple treatments.[20] Nevertheless, similarly to the discussion in previous sections, a more flexible approach to the estimation of the average dose-response function is possible, such as relying on the matching methods already described. In this case, parametric assumptions may be made only for estimating the GPS[21], placing no restrictions on the relationship between programme exposure, covariates (summarised by the GPS) and the relevant health outcome. Among the matching methods previously presented, possible adaptations of the propensity score matching (PSM) procedure will be the focus herein. Two

[20] The suggested three-stage estimation procedure will not be presented and discussed here because it does not make use of matching methods at any stage. The interested reader is referred to HI's (2004) original paper.

[21] A non-parametric procedure could also be used for estimating the GPS, as in Flores (2004). Although more flexibility is introduced in the average dose-response function estimation, it is well known that non-parametric procedures tend to run into trouble when a large number of covariates are being considered. Non-parametric estimation of the GPS will not be discussed in this paper.

possible ways of performing PSM in the continuous treatment case are briefly discussed by Flores (2004). Here, I adapt and extend one of the alternatives.

Let the parameter of interest be the entire average dose-response function $\mu(l)$. In order to obtain this curve, we have to estimate the average expected health outcome at each treatment level of interest. Thus, for a given coverage level *l*, we need to estimate the expectation $E\left[Y(l)\right]$. Estimation of the conditional distribution of the treatment (coverage) level given covariates, $L_i = f(X_i)$, can follow HI (2004) suggested approach. The authors propose using the normal distribution, such as:[22]

$$L_i \mid X_i \sim N\left(\beta_0 + \beta_1' X_i, \sigma^2\right) \tag{61}$$

A maximum likelihood method is suggested for the above estimation. In this case, the estimated GPS is given by the predicted values of the regression, based on the probability density function of L_i, $f\left(L_i \mid X_i, \beta_i, \sigma^2\right)$:

$$\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}\left(L_i - \hat{\beta}_0 - \hat{\beta}_1' X_i\right)^2\right) \tag{62}$$

As pointed out by Flores (2004), for estimating the average dose-response function at coverage level *l*, it might be unfeasible to find observations with exactly that value of the coverage level in the continuous treatment case. This difficulty implies that matching has now to be performed not only on the GPS, but also on the treatment level. By following this procedure, in addition to the potential bias introduced by not matching exactly on the GPS, there is potential bias resulting from the need of using observations belonging to the *neighbourhood of* (instead of *exactly exposed to*) treatment level *l* in order to get information about the potential health outcomes at that particular treatment level.

The matching problem arising from the continuous nature of the treatment variable and the resulting difficulty in finding observations exposed exactly to level *l* can be circumvented by assuming that a given individual has been exposed to *l* if her actually assigned coverage level is sufficiently close to *l*. Formally, define two groups that exhaust the sample (in the sense that any sampled observation must belong to one and only one of the groups):

[22] Of course, more general models are also possible, such as heteroskedastic normal distributions for instance.

$$group \begin{cases} i \mid L_i \in [l-h, l+h] \\ j \mid L_j \notin [l-h, l+h] \end{cases}$$

where *h* is a chosen bandwidth that determines the set of treatment levels considered sufficiently close to the level of interest *l*. For the first group of individuals (*i*), their observed individual health outcomes Y_i can be seen as suitable approximations to their potential outcomes at coverage level *l*, $Y_i(l)$. The basic evaluation problem appears in obtaining the counterfactual potential health outcomes, at coverage level *l*, for those individuals who have been exposed to treatment levels far apart from *l*. Yet, assuming unconfoundedness, PSM can be used to construct the required counterfactuals. For any individual *j*, all that has to be done is to find at least one comparable observation *i* in the first group whose health outcome could be reliably used to impute the missing data.

In formal terms, let $M_j(r_j)$ denote the matched set for individual *j*, i.e. the set of individuals within the first group matched to *j*. In the simplest case of nearest-neighbour matching, the singleton matched set will be formed by the observation *i* for which the following two conditions are valid:

a) $L_i \in [l-h, l+h]$, and

b) $M_j(r_j) = \left\{ i \mid i = \arg\min_i \left\| r_i(l, X_i) - r_j(l, X_j) \right\| \right\}$.

Condition *b)* requires the definition of some metric for the distance $\|\cdot\|$ between GPS, such as the Euclidean or Mahalanobis distances. As in the preceding single and multiple treatments settings, flexibility can be introduced by allowing a given observation *i* to be matched to multiple neighbours. Also, a common support requirement can be enforced by imposing a maximum "tolerated distance" between GPS when looking for matches. For instance, caliper matching with multiple neighbours can be performed, in which case the definition of the matched set in *b)* would be replaced by $M_j(r_j) = \left\{ i \mid \left\| r_i(l, X_i) - r_j(l, X_j) \right\| < \varepsilon \right\}$, where ε is the pre-defined tolerated distance between GPS (the caliper). The procedure for constructing the matching set should then be repeated, with replacement, for every observation *j*.

Having constructed the set $M_j(r_j)$ which contains at least one matched *i* observation to a given individual *j*, the counterfactual health outcome of the latter can

be imputed. The general form of the GPS matching estimator for the relevant individual counterfactual can be written as:

$$\hat{Y}_j(l) = \sum_{i \in M_j} W_{ji} Y_i(L_i) \tag{63}$$

where W_{ji} stands for the weight attributed to observation *i* when matched to *j* (with the sum of weights being equal to one). In the nearest-neighbour matching case, the weight attributed to the matched individual *i* is equal to one and the imputed counterfactual is the health outcome of this closest neighbour, i.e.:

$$\hat{Y}_j(l) = Y_i(L_i),\ i \in M_j$$

For the multiple-matches case, the simplest counterfactual estimator would be:

$$\hat{Y}_j(l) = \frac{1}{N_{M_j}} \sum_{i \in M_j} Y_i(L_i)$$

where N_{M_j} denotes the number of individuals belonging to the matched set for individual *j*. In principle, different and more complex weighting schemes could alternatively be applied, such as kernel weights; moreover, besides weighting by the GPS, one could devise weights depending also on the distance $\|L_i - l\|$.

As a final step, the *GPS matching estimator* of the average dose-response function at a given treatment level *l* can be written as:

$$\hat{E}\left[Y(l)\right] = \frac{1}{N}\left(\sum_i Y_i(L_i) + \sum_j \hat{Y}_j(l)\right),\ L_i \in [l-h, l+h] \tag{64}$$

Having this estimator, the treatment effect parameters defined from (52) to (55) can also be estimated, and bootstrap methods can be used for calculating standard errors and the corresponding confidence intervals for the estimated treatment effect parameters.

Regardless of the matching procedure implemented, diagnostic checks should be performed in order to assess the balancing of covariates across individuals with similar GPS for a given treatment level, and the quality of the matches used. This can be done by applying suitably adapted versions of previously discussed methods for the multiple treatments case and/or the procedures suggested by HI (2004), who address some of the difficulties posed to diagnostic checking when the treatment variable is continuous. For instance, in order to assess the degree to which covariates are balanced by conditioning on the estimated GPS, HI (2004) suggest dividing the range of variation of the treatment variable – and therefore the sample of observations

– into intervals of the form $l_a \leq L_i \leq l_b$ (three in their empirical application) and investigate, for each covariate and via *t*-tests, whether the mean in one of these intervals is different from the mean (for the same interval) in the remaining groups combined. The authors compare the change in the achieved covariates balance when conditioning on the GPS, by comparing the number of *t*-statistics which lead to rejection of the null of means equality in the unadjusted versus GPS-adjusted intervals.

However, in HI (2004) procedure, the treatment variable is indeed discretised and no general criteria are suggested for a "suitable" intervals definition, which is important because the result of the test will depend on this particular decision. Moreover, the GPS needs also to be discretised for implementing the check $X \perp 1\{l_a \leq L_i \leq l_b\} \mid r(l, X_i)$, where $1(\cdot)$ is an indicator function of whether an individual's observed treatment level belongs to the corresponding interval. HI (2004) perform this test by evaluating the GPS at the median treatment level of the sub-sample of individuals determined by the interval $[l_a \leq L_i \leq l_b]$, that is, the test is whether $X_i \perp 1\{l_a \leq L_i \leq l_b\} \mid r_i\left(l_{med[l_a,l_b]}, X_i\right)$. The authors implement this test by blocking on that particular GPS $r(\cdot)$ and testing equality of means within quintiles of the values for that GPS in the interval. In other words, covariates means within groups defined by $L_i \in [l_a, l_b]$ and $L_i \notin [l_a, l_b]$ are being compared for individuals who have similar values – belong to the same quintile – of $r\left(l_{med[l_a,l_b]}, X_i\right)$ (i.e., who have similar conditional densities of being exposed to the median coverage level $l_{med[l_a,l_b]}$).[23]

6 Concluding remarks

In this paper, I reviewed the state of the art of the literature on matching methods, with a special focus on its propensity score variant. The broad usefulness of this approach was discussed from the specific point of view of health programmes evaluation. Extensions of the classical matching estimators for the multiple and continuous treatments cases were presented and their relevance for impact evaluations

[23] Five different comparisons should then be performed for each covariate; nevertheless, HI (2004) combine these five differences in means, weighting by the number of observations in each GPS group, and get a summary *t*-statistic for the difference in means across the five quintiles.

in the health sector was illustrated using the example – fairly common in real settings – of a health programme implemented with different levels of population coverage in several geographic areas.

The discussion performed in this paper made clear that the validity of matching estimates of treatment effects depends crucially on the "selection on observables" identification assumption. If, in a given context, such assumption seems reasonable, then matching methods can go a long way in providing reliable answers for evaluation questions. There remain, however, important aspects of matching estimators still to be fully investigated, such as analytic closed forms for the variance of the treatment effects matching estimates, the relative performance of different matching procedures for obtaining estimates of average treatment effects in the multiple treatments case and a fully developed and tested matching protocol for the continuous treatment setting. Further research on such topics is certainly needed.

References

Abadie, A., and G. Imbens (2002). "Simple and bias-corrected matching estimators for average treatment effects". NBER Technical Working Paper 283.

Abadie, A., and G. Imbens (2006). "On the failure of the bootstrap for matching estimators". NBER Technical Working Paper 325.

Aguero, J., M. Carter and I. Woolard (2006). "The impact of unconditional cash transfers on nutrition: the South African Child Support Grant". Mimeo, University of California, Riverside.

Behrman, J., Y. Cheng and P. Todd (2004). "Evaluating preschool programs when length of exposure to the program varies: a nonparametric approach". Review of Economics and Statistics 86(1): 108-132.

Blundell, R., and M. Costa-Dias (2000). "Evaluation methods for non-experimental data". Fiscal Studies 21(4): 427-468.

Blundell, R., M. Costa-Dias, C. Meghir and J. Van Reenen (2002). "Evaluating the employment impact of a mandatory job search assistance program". IFS Working Paper WP01/20.

Cameron, A.C., and P. Trivedi (2005). "Microeconometrics: Methods and Applications". Cambridge: Cambridge University Press.

Dehejia, R., and S. Wahba (1999). “Causal effects in nonexperimental studies: reevaluating the evaluation of training programs”. Journal of the American Statistical Association 94: 1053-1062.

Dehejia, R., and S. Wahba (2002). “Propensity score-matching methods for nonexperimental causal studies”. Review of Economics and Statistics 84(1): 151-161.

Flores, C. (2004). “Estimation of dose-response functions and optimal doses with a continuous treatment”. Mimeo, Department of Economics, University of California, Berkeley.

Frolich, M. (2004). “Programme evaluation with multiple treatments”. Journal of Economic Surveys 18(2): 181-224.

Frolich, M., A. Heshmati and M. Lechner (2004). “A microeconometric evaluation of rehabilitation of long-term sickness in Sweden”. Journal of Applied Econometrics 19: 375-396.

Hahn, J. (1998). “On the role of the propensity score in efficient semiparametric estimation of average treatment effects”. Econometrica 66(2): 315-331.

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(4): 605-654.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998). “Characterizing selection bias using experimental data”. Econometrica 66(5): 1017-1098.

Hirano, K. and G. Imbens (2004). “The propensity score with continuous treatments”. In: A. Gelman and X.L. Meng (Eds.), Missing Data and Bayesian Methods in Practice. New York: Wiley.

Imbens, G. (2000). “The role of the propensity score in estimating dose-response functions”. Biometrika 87(3): 706-710.

Imbens, G. (2004). “Nonparametric estimation of average treatment effects under exogeneity: A review”. Review of Economics and Statistics 86(1): 4-29.

Lechner, M. (2000). “Identification and estimation of causal effects of multiple treatments under the conditional independence assumption”. Mimeo, Swiss Institute for International Economics and Applied Economic Research, University of St. Gallen.

Lechner, M. (2002). “Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies”. Review of Economics and Statistics 84(2): 205-220.

Miguel, E., and M. Kremer (2004). “Worms: Identifying impacts on education and health in the presence of treatment externalities”. Econometrica 72(1): 159-217.

Rosenbaum, P., and D. Rubin (1983). “The central role of the propensity score in observational studies for causal effects”. Biometrika 70(1): 41-55.

Rosenzweig, M. and K. Wolpin (1988). “Migration selectivity and the effects of public programs”. Journal of Public Economics 37(3): 265-289.

Smith, J., and P. Todd (2005). “Does matching overcome LaLonde’s critique of nonexperimental estimators?”. Journal of Econometrics 125: 305-353.