Empir Econ (2015) 49:1–31
DOI 10.1007/s00181-014-0847-1

Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behaviour

Martin Huber · Michael Lechner · Andreas Steinmayr

Received: 18 June 2013 / Accepted: 15 May 2014 / Published online: 2 August 2014

Abstract Using a simulation design that is based on empirical data, a recent study by Huber et al. (J Econom 175:1–21, 2013) finds that distance-weighted radius matching with bias adjustment as proposed in Lechneret et al. (J Eur Econ Assoc 9:742–784, 2011) is competitive among a broad range of propensity score-based estimators used to correct for mean differences due to observable covariates. In this companion paper, we further investigate the finite sample behaviour of radius matching with respect to various tuning parameters. The results are intended to help the practitioner to choose suitable values of these parameters when using this method, which has been implemented in the software packages GAUSS, STATA and R.

Keywords Propensity score matching · Radius matching · Selection on observables · Empirical Monte Carlo study · Finite sample properties

JEL Classification C21

Michael Lechner is a Research Fellow of CEPR and PSI, London, CES-Ifo, Munich, IAB, Nuremberg, and IZA, Bonn.

M. Huber · M. Lechner (✉) · A. Steinmayr
Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen,
Varnbüelstrasse 14, 9000 St. Gallen, Switzerland
e-mail: michael.lechner@unisg.ch
URL: http://www.sew.unisg.ch/lechner

M. Huber
e-mail: martin.huber@unisg.ch

A. Steinmayr
e-mail: andreas.steinmayr@unisg.ch

1 Introduction

In the applied literature on the evaluation of binary treatments or policy interventions, matching estimators are often used to remove differences in the distributions of covariates across treatment states. Instead of matching on the covariates directly, these estimators are usually based on the propensity score, i.e. the conditional treatment probability given observed covariates.[1] Propensity score methods are usually implemented as semiparametric estimators, i.e. the propensity score is estimated by a parametric model, whereas the relationship between the outcome variables and the propensity score is nonparametric. This allows controlling for covariates in a more flexible way than (non-saturated) parametric regression and permits effect heterogeneity w.r.t. observables, whereas curse of dimensionality problems related to an entirely non-parametric estimation are avoided. Popular classes of propensity score methods include direct matching (Rubin 1974; Rosenbaum and Rubin 1983), kernel matching (Heckman et al. 1998a), radius matching (Rosenbaum and Rubin 1985; Dehejia and Wahba 1999), inverse probability weighting (Horvitz and Thompson 1952; Hirano et al. 2003), inverse probability tilting (Graham et al. 2012) and doubly robust estimation (Robins et al. 1992).

Huber et al. (2013), henceforth referred to as HLW13, assess the finite sample properties of a broad range of different (classes of) estimators of the average treatment effect on the treated (ATET) by constructing a—what they call—*Empirical Monte Carlo Study* (EMCS) which is based on empirical labour market data from Germany. The simulation study considers various scenarios with different sample sizes, shares of treated and non-treated, levels of selectivity into the treatment and propensity score specifications. Overall, a version of radius matching with regression-based bias adjustment as proposed in Lechner et al. (2011), henceforth LMW11, performed best in terms of root mean squared error when estimating the average treatment effect on those who received the treatment.[2] The study also reveals that estimator performance may vary with the choice of tuning parameters such as the width of the radius, i.e. the size of the local neighbourhood around the propensity score within which counterfactual observations are matched and whether matching is not solely on the propensity score, but in addition on further important covariates based on the Mahalanobis distance metric. However, due to the large variety of estimators investigated and the related computational burden, HLW13 could not assess the sensitivity of the LMW11 estimator w.r.t. to the values of these parameters in great detail. Previous simulation studies on propensity score methods (Frölich 2004; Busso et al. 2009a, b) do not even include radius matching.

Using the same simulation design as HLW13, this companion paper more thoroughly investigates the impact of tuning parameters on the root mean squared error, bias, variance, skewness and kurtosis of this estimator for the ATET. While the former three features are relevant for consistency, the latter two moments indicate whether the

[1] See for example the recent surveys by Blundell and Costa Dias (2009), Imbens (2004), and Imbens and Wooldridge (2009) for a discussion of the properties of such estimators as well as a list of recent applications.

[2] It has also been used in Wunsch and Lechner (2008), Lechner (2009), Lechner and Wunsch (2009a, b), Behncke (2010a); Behncke et al. (2010b)), and Huber et al. (2011).

estimator's distribution can be adequately approximated by the normal distribution, which is relevant for inference. The parameters considered are the size of the radius and whether matching is on the propensity score only or also on additional important predictors via Mahalanobis distance matching. The size of the radius is varied as a function of the distances of matched treated and controls in one-to-one (or pair) matching. That is, the quantile at a particular rank in the distribution of distances is multiplied by a constant term, which we call the radius multiplier, to define the radius. The latter is thus not fixed in absolute terms but may change from one application to another depending on the distribution of pair differences, an approach that has not been considered in previous simulation studies.[3] In the EMCS, we consider three choices for the quantile (0.1, 0.5 and 0.9) and four for the radius multiplier (0.25, 1, 10 and 100), i.e. 12 different definitions of the radius. In contrast, HLW13 considered three radius sizes (0.5, 1.5 and 3 times the maximum distance of matched treated and controls in pair matching). Note that compared to the maximum, a quantile may be less variable as it does not completely depend on a particular large observation. Concerning the covariates used in the Mahalanobis distance and the regression adjustment, we use none (propensity score matching), 1 or 4 additional matching variables on top of the propensity score (while HLW13 included 2 additional covariates in Mahalanobis matching). In addition, we also investigate the impact of assigning different weights to the propensity score in the Mahalanobis metric, namely 0.5 (i.e. the score receives half the weight of any other covariate), 1 and 5.

The results suggest that both the radius size and the number of covariates in the Mahalanobis metric/regression adjustment influence the estimator's behaviour importantly, while the propensity score weight does not (at least for the values investigated). Specifically, a larger choice of the radius and the number of covariates decreases the RMSE, which is mainly driven by a reduction in the standard deviation while the bias is not much affected. Because increasing these tuning parameters implicitly shifts more weight to the parametric regression adjustment, our results suggest that the latter performs well in terms of reducing the RMSE. Therefore, combining (distance-weighted) radius matching and regression in an appropriate way appears to improve the properties of the estimator.

This paper makes several contributions to the literature on matching estimators. Firstly, it thoroughly investigates the importance of tuning parameters for radius matching as proposed by LMW11. Secondly, it does so using the EMCS design of HLW13, which is likely to be closer to real world applications than arbitrarily chosen data generating processes not based on empirical data. Finally and particularly relevant for practitioners, the LMW11 estimator has been implemented as the "BinMatch" programme in the statistical software package GAUSS, and as the "radiusmatch" command in STATA, and as the R package "radiusmatching", along with options for tuning parameters, common support procedures and inference methods. These programmes constitute an alternative to other matching packages, which so far do not offer a radius matching procedure that includes all of the following features/options inherent in this command: (i) weighting of the matched controls within the radius according to their

[3] Note that HLW13 combine the radius multiplier with the maximum distance between matched, rather than a particular quantile.

distance to the treated observation, (ii) bias-adjustment based on OLS or logit regression depending on the support of the outcome variable, (iii) partially data-driven choice of the radius size as a function of the distances in pair matching and (iv) asymptotically unbiased propensity score trimming as considered in HLW13 to ensure common support in the propensity score across treatment groups. The estimator can be downloaded at http://www.alexandria.unisg.ch/publications/citation/Michael_Lechner/218871.[4]

The remainder is organized as follows. Section 2 discusses identification based on the propensity score (2.1) as well as matching estimation in general (2.2) and the LMW11 algorithm in particular (2.3). It also covers common support procedures (2.4) and inference methods (2.5) that are available in the programmes. Section 3 reviews the Empirical Monte Carlo Study design of HLW13. The simulation results are presented in Section 4. Section 5 concludes.

2 Econometrics

2.1 Identification and general estimation principle

In the treatment evaluation literature, identification strategies based on a 'selection on observables' or 'conditional independence' assumption (CIA) require that all factors jointly affecting the treatment probability, and the outcomes are observed and thus can be controlled for. That is, potential outcomes that would have been realized under either treatment state are assumed to be independent of the actual treatment assignment conditional on the observed covariates, see for instance Imbens (2004) for an in-depth discussion. To formalize the discussion, we denote the observed outcome by Y, e.g. employment or earnings in labour market applications, by D the binary treatment indicator taking either the value 1 (treated, e.g. receiving a training) or 0 (non-treated) and by X the vector of observed covariates (e.g. labour market experience, education and age). Using the potential outcome framework advocated by Rubin (1974), among many others, we let $Y(1)$ and $Y(0)$ denote the potential outcomes under treatment and non-treatment, respectively. By the observational rule, only one potential outcome can be observed, because $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$. The CIA states that

$$Y(1), Y(0) \perp D|X, \tag{1}$$

where $\perp$ denotes independence. In many empirical applications, this assumption only appears plausible when controlling for a large set of covariates. However, conditioning on a high dimensional X may be problematic, as the number of possible combinations of elements in X increases exponentially in the dimension such that (acceptably precise) estimation quickly becomes exorbitantly data hungry, a problem known as curse of dimensionality.

This motivates the use of propensity score methods frequently encountered in applied work. We denote $p(X) \equiv Pr(D = 1|X)$ the propensity score, the conditional

[4] The latest version of the GAUSS codes is available from http://www.michael-lechner.eu/software. The latest version of the STATA code is available from the SSC archive.

treatment probability given the covariates. Rosenbaum and Rubin (1983) showed that conditioning on the propensity score is asymptotically equivalent to conditioning on the covariates directly, as both X and $p(X)$ are balancing scores in the sense that they adjust the distributions of covariates in the treatment and in the control (or non-treated) group. Thus, if (1) is fulfilled, it also holds that the potential outcomes are independent of the treatment conditional on the propensity score:

$$Y(1), Y(0) \perp D | p(X). \tag{2}$$

In principle, conditioning on the propensity score, therefore, allows for the identification of causal effects such as the average treatment effect (ATE) in the entire population, $E[Y(1) - Y(0)]$, because (2) implies that

$$\begin{aligned} E[Y(0)|D = 1, p(X)] &= E[Y(0)|D = 0, p(X)] = E[Y|D = 0, p(X)], \\ E[Y(1)|D = 0, p(X)] &= E[Y(1)|D = 1, p(X)] = E[Y|D = 1, p(X)]. \end{aligned}$$

However, a large part of the applied literature focuses on the evaluation of the average treatment effect on the treated (ATET), defined as $\theta = E[Y(1) - Y(0)|D = 1]$, which is also the estimand considered in this paper.[5] In this case, (2) may be relaxed to

$$Y(0) \perp D | p(X). \tag{3}$$

Identification also requires that the following common support assumption of the propensity score holds for all values of the covariates:

$$p(X) < 1, \tag{4}$$

i.e. the treatment must not be perfectly predicted by any combination of the covariates to ensure that non-treated matches are available, at least asymptotically. Under (3) and (4) and by the law of iterated expectations,

$$\begin{aligned} \theta &= E[Y(1)|D = 1] - E[Y(0)|D = 1] \\ &= E[Y|D = 1] - E[E[Y|D = 0, p(X)]|D = 1], \end{aligned} \tag{5}$$

so that the ATET is identified.

Concerning estimation, assume that we have an i.i.d. sample of (Y, D, X) consisting of N observations denoted by i, where $i \in \{1, 2, ..., N\}$. Then, a general class of estimators of (5) can be defined as

[5] We focus on the ATET for reasons of computational costs. Note that estimating the average treatment effect on the non-treated (ATENT) is symmetric to the problem we consider (just recode D as $1 - D$) and thus not interesting in its own right. The ATE is obtained as a weighted average of the ATET and the ATENT, where the weight for the ATET is the share of treated and the weight of ATENT is one minus this share.

$$\hat{\theta} = \frac{1}{N_1} \sum_{i=1}^{N} d_i y_i - \frac{1}{N_0} \sum_{i=1}^{N} (1 - d_i) \hat{w}_i y_i, \tag{6}$$

where N_1 and N_0 are the number of treated and non-treated observations, respectively, and $\hat{w}_i$ is a weight that is a function of the estimated propensity score $\hat{p}(x_i)$. $\hat{w}_i$ reweights the non-treated observations such that they resemble the treated sample in terms of the distribution of the propensity score as well as the covariates X and differs across distinct (classes of) estimators (such as matching and inverse probability weighting). As a final remark, note that the applicability of these methods is not confined to the leading case of treatment evaluation in i.i.d. settings. They may be applied whenever the adjustment of covariate distributions across different groups is of interest, which does not necessarily imply a CIA or i.i.d. sampling. E.g. propensity score methods have been fruitfully applied to instrumental variable estimation, see for instance Frölich (2007).

2.2 Matching estimators

Prototypical one-to-one or pair matching on the propensity score matches to each treated unit exactly one control unit that is closest in terms of the propensity score. In the subsequent discussion, we focus on matching 'with replacement', implying that the same control observation may be used several times as a match, whereas in estimators 'without replacement' it is matched at most once. However, the latter principle only works well when there are many more controls than treated. The pair matching estimator based on matching with replacement is defined as

$$\hat{\theta}_{\mathrm{PM}} = \frac{1}{N_1} \sum_{i:d_i=1} \left[y_i - \sum_{j:d_j=0} \underline{1} \left(\min \left| \hat{p}(x_j) - \hat{p}(x_i) \right| \right) y_j \right], \tag{7}$$

$\underline{1}(\cdot)$ denotes the indicator function, which is one if its argument is true and zero otherwise. A direct extension of pair matching is $1{:}M$ nearest neighbour matching which uses several (i.e. M) controls instead of just one. Increasing M increases the precision but also the bias of the estimator, as even 'not so close' controls might be matched in this case. Radius matching, see for instance Rosenbaum and Rubin (1985) and Dehejia and Wahba (1999, 2002), tackles this issue using only controls that are situated within a predefined distance around the propensity score of each treated unit. Compared to $1{:}M$ matching, this may lead to a smaller bias in regions where comparable controls are sparse. Also, it increases precision compared to $1{:}M$ nearest neighbour matching in propensity score regions with many similar controls. Instead of fixing M globally, radius matching determines the number of matches, M, in the local neighbourhood of each treated observation.

Further improvements to standard propensity score matching have been proposed in the literature. Rubin (1979) suggested combining pair matching with (parametric) regression adjustments to take into account the fact that treated and controls with exactly the same propensity score are usually very rare or non-existent. Also

Abadie and Imbens (2006) consider this idea and show (however, for 1:M matching on X rather than on the propensity score) that nonparametric regression removes the asymptotic bias that may occur when X is more than one-dimensional. Furthermore, instead of matching on the propensity score alone, one may use a distance metric that (in addition to the score) accounts for differences in those covariates that are particularly good predictors of the outcome. In finite samples, this potentially improves estimation by putting a larger emphasis on balancing the most important confounders across treatment states. The intuition behind this potential improvement is that it is particularly important to balance variables that have a large influence on the outcomes, as any imbalances of those variables will lead to larger biases than imbalances of variables that are only slightly correlated with the outcomes. In this case, the Mahalanobis distance metric is commonly used to collapse the multidimensional distances between the propensity scores and predictors of the treated and the controls into a single measure (see Rosenbaum and Rubin 1985) for details. The distance between two observations is defined as

$$\sqrt{\left(\tilde{x}_i^{D=1} - \tilde{x}_j^{D=0}\right) C^{-1} \left(\tilde{x}_i^{D=1} - \tilde{x}_j^{D=0}\right)'}, \tag{8}$$

where $\tilde{x}_i^{D=1}, \tilde{x}_j^{D=0}$ are row vectors of the K factors to be matched on, i.e. the propensity score and $K-1$ further covariates, of some treated observation i and some control j, respectively. C denotes the covariance matrix of the K covariates in the control group. In Mahalanobis matching, the distances are weighted by the inverse of their covariance matrix to give higher weights to less noisy differences and those with smaller covariances.[6] As a modification of the original metric, which treats the propensity score and each of the covariates as equally important, one may assign a higher weight to the propensity score than to the other elements in $\tilde{x}_i^{D=1}$, $\tilde{x}_j^{D=0}$. This is obtained by multiplying the inverse of the variance of the propensity score in C^{-1} by a factor larger than one. As a further modification, we do not take the square root as proposed in Eq. (8), with the consequence that observations further away will receive less weight by the matching algorithm.

2.3 The radius matching algorithm of Lechner, Miquel and Wunsch (2011)

The LMW11 estimator combines the features of distance-weighted radius matching with a logit- or OLS-based regression adjustment (depending on whether the outcome is binary or not) as well as Mahalanobis matching when using further covariates besides the propensity score (which are also included in the propensity score). The first step consists of distance-weighted radius matching either on the propensity score or the

[6] In contrast, the Euclidean distance metric - defined as $\sqrt{\left(\tilde{x}_i^{D=1} - \tilde{x}_j^{D=0}\right) I \left(\tilde{x}_i^{D=1} - \tilde{x}_j^{D=0}\right)'} = \sqrt{\sum_{k=1}^{K} \left(\tilde{x}_{i,k}^{D=1} - \tilde{x}_{j,k}^{D=0}\right)^2}$, with I denoting the K-dimensional identity matrix and $\tilde{x}_{i,k}^{D=1}, \tilde{x}_{j,k}^{D=0}$ being the k^{th} elements in $\tilde{x}_i^{D=1}, \tilde{x}_j^{D=0}$- would assign equal weights to all differences, irrespective of how much they differ in terms of standard deviations and covariances.

Mahalanobis metric, respectively. Distance-weighting implies that controls within the radius are weighted proportionally to the inverse of their distance to the respective treated they are matched to when computing the local mean outcome under non-treatment. In contrast to standard radius matching algorithms, controls within the radius do not obtain the same weight independent of their location. Therefore, the LMW11 estimator can also be interpreted as a kernel matching estimator based on a truncated triangular kernel. In the second step, the weights obtained from matching are used in a weighted linear or non-linear regression in order to remove biases due to mismatches.[7]

An open, though very important, question in radius matching is the choice of the size of the radius, for which no well-established algorithm exists. LMW11 suggest – rather arbitrarily but data-driven – defining the size as a function of the maximum distance between treated and matched controls in pair matching.[8] Alternatively, one may consider the quantile at a particular rank of the distance distribution instead of the maximum distance. The latter approach might be more robust to outliers in the distances as it is less variable. Considering both options, the LMW11 estimator follows the matching protocol outlined in Table 1.

The estimator depends on several tuning parameters. Besides choosing the maximum distance (*maxdist*) or a particular quantile in the distance distribution (*quantdist*), which we henceforth refer to as distance quantile, in step D-1, one also needs to define the radius multiplier R in step B-2. The product of R and *maxdist* or *quantdist*, respectively, determines the absolute size of the radius, which may vary from application to application because it is partially data-driven by the distances in pair matching. Finally, (the number of) additional covariates entering the Mahalanobis distance as well as the weight, the propensity score receives relative to the covariates have to be selected in B-1. The sensitivity of the estimator's properties to the choice of these tuning parameters will be investigated in Section 4.

2.4 Distributional overlap

The issue of thin or even lacking common support (or overlap) in the propensity score across treatment states has been discussed extensively in the literature (see the surveys by Heckman et al. 1999; Imbens 2004, and Imbens and Wooldridge 2009), because it may hamper estimation due to a non-comparability of treated and controls. If particular values of $p(x)$ that are observed for the treated are either very rare ('thin common

[7] Note that this estimator satisfies the so-called 'double robustness property': it is consistent if either the matching step is based on a correctly specified propensity score model or if the bias-adjustment step is based on a correctly specified regression model (see for instance Joffe et al. 2004, and Rubin 1979). However, in our implementation the propensity score and the variables included in the Mahalanobis metric are used as regressors in the local adjustment. Therefore, the relevance of the double robust property in our context is not clear.

[8] We acknowledge that cross-validation might be an alternative data-driven approach worth considering. See Frölich (2005), whose simulations suggest that cross-validation performs rather well for bandwidth selection in kernel matching (and in particular better than a selection method based on an asymptotic approximation of the estimator's mean squared error), even though it does asymptotically not provide the optimal bandwidth. Similar arguments could carry over to radius matching as considered in this paper.

Table 1 Matching protocol for the estimation of a counterfactual outcome and the effects

Step A-1	Choose one observation in the subsample defined by $d = 1$ and delete it from that pool
Step B-1	Find an observation in the subsample defined by $d = 0$ that is as close as possible to the one chosen in step A-1) in terms of either (i) $p(x)$ (matching on the propensity score only), or (ii) $p(x)$ and additional predictors (matching on the propensity score and a subset of X). In the latter case, 'closeness' is based on the Mahalanobis distance, in which $p(x)$ and the additional predictors may or may not be weighted
Step C-1	Repeat (A-1) and (B-1) until no observation with $d = 1$ is left
Step D-1	Compute the maximum distance (*maxdist*) obtained for any comparison between a member of the reference distribution and matched comparison observations. Alternatively, one may also compute the quantile at a particular rank in the distribution of distances (*quantdist*)
Step A-2	Repeat (A-1)
Step B-2	Repeat (B-1). If possible, find other observations in the subsample of $d = 0$ that are at least as close as R_* *maxdist* or R_* *quantdist*, respectively, to the one chosen in step A-2), where R denotes the radius multiplier. Do not remove these observations, so that they can be used again. Compute weights for all chosen comparisons observations that are proportional to their distance. If no control observation is at least as close as the chosen radius, find the closest observation outside the radius. Normalise the weights such that they add to one
Step C-2	Repeat (A-2) and (B-2) until no participant in $d = 1$ is left
Step D-2	For any potential comparison observation, add the weights obtained in (A-2) and (B-2)
Step E	Using the weights of the comparison observations obtained in (D-2), run a weighted linear regression of the outcome variable on an intercept, the propensity score, its square, and any further variables used to define the distance
Step F-1	Predict the potential outcome $y^0(x_i)$ of every observation using the coefficients of this regression: $\hat{y}^0(x_i)$
Step F-2	Estimate the bias of the matching estimator for $E(Y^0\|D = 1)$ as: $\sum_{i=1}^{N} \frac{(1-d_i)w_i \hat{y}^0(x_i)}{N_0} - \frac{d_i \hat{y}^0(x_i)}{N_1}$
Step G	Using the weights obtained by weighted matching in (D-2), compute a weighted mean of the outcome variables in $d = 0$. Subtract the bias from this estimate to get $\widehat{E(Y^0\|D = 1)}$

For estimation of the ATENT the counterfactual distribution can be obtained by replacing d by 1-d and repeating steps A–G

support') or absent (lack of common support) among the controls, as it may happen in particular close to the boundary of $p(x) = 1$, control observations with such values, or very close to them, receive a large weight $\hat{w}_i$. In the case of thin common support, these observations may dominate the estimator of the ATET which may entail a possible explosion of the variance. In the case of lacking common support, this even introduces asymptotic bias by giving a large weight to controls that are not comparable to the treated in terms of the propensity score.

There have been different proposals in the literature on how to tackle the common/thin support problem, which, however, all introduce asymptotic bias, see Heckman et al. (1998a), Dehejia and Wahba (1999), Ho et al. (2007) and Crump et al. (2009). In contrast, HLW13 suggests using a trimming procedure that was first discussed in Imbens (2004, p. 23) and is asymptotically unbiased in DGPs where common

support holds asymptotically (such as the simulation design presented in Section 3). The idea is to set the weight of any control observation to zero whose relative share of all weights exceeds a particular threshold value in percent (denoted by t):

$$\hat{w}_{i|d_i=0} = \hat{w}_i \, \underline{1}\left[\hat{w}_i \Big/ \sum_{j=1}^{N} (1-d_j)\hat{w}_j \leq t\%\right]. \tag{9}$$

As the trimming procedure is applied before the estimation, this raises the question of how to obtain the weights in (9). In principle, one could apply any propensity score-based method (including matching) as a preliminary procedure to compute $\hat{w}_i$. As in HLW13, we use normalized inverse probability weighting, which is computationally inexpensive and implies the following weights:

$$\hat{w}_i = \frac{\frac{(1-d_i)\hat{p}(x_i)}{1-\hat{p}(x_i)}}{\sum_{j=1}^{N} \frac{(1-d_j)\hat{p}(x_j)}{1-\hat{p}(x_j)}}. \tag{10}$$

To avoid a severely unbalanced sample induced by trimming the controls only, also all treated observations with a value of $\hat{p}(x)$ larger than the largest value of $\hat{p}(x)$ among the remaining controls are removed (if such observations exist). Strictly speaking, this changes the estimand due to discarding extreme support areas, but ensures common support prior to matching. Note that the matching algorithm then produces its own (normalized) weights which are the base for the actual estimator and for inference, such that the weights defined in (10) are no longer used after trimming. Besides the trimming procedure, the available programmes also include the conventional common support procedure suggested by Dehejia and Wahba (1999), which removes all treated with propensity scores that are larger than the largest propensity score among controls.[9] The study by Lechner and Strittmatter (2014) provides an in-depth investigation of the properties of various procedures aiming at reducing common support problems.

2.5 Inference methods

Under i.i.d. sampling, the variance of the ATET estimator is asymptotically simply the sum of the variances of the estimators of the treated population's mean potential outcomes under treatment and non-treatment (ignoring any correlation that may occur due to the estimation of the propensity score). Denoting the variance estimator by $\hat{V}(\cdot)$, a consistent estimator of the variance of the mean potential outcome under treatment is $\hat{V}\left(E(y_i|d_i=1)\right) = \frac{1}{N_1-1}\sum_{i:d_i=1}^{N_1}\left(y_i - \left(\frac{1}{N_1}\sum_{i:d_i=1}^{N_1} y_i\right)\right)^2 / N_1$. To approximate the variance under non-treatment, an estimator of $\sigma_i^2 = E[(y_i-\mu_i)^2|w_i, d=0]$, the conditional variance among controls given the matching weight, is required, with $\mu_i = E(y_i|w_i, d_i=0)$ denoting the conditional mean. To this end, we first estimate the

[9] If both procedures are used at the same time, the common support restriction of Dehejia and Wahba (1999) is enforced prior to trimming the weights of the remaining observations.

latter by $\hat{\mu}_i = \hat{E}(y_i|\hat{w}_i, d_i = 0)$, where $\hat{E}(\cdot|\cdot)$ denotes a local regression estimator. In a second step, the conditional variance is estimated by plugging in the first-step estimate $\hat{\mu}_i : \hat{\sigma}_i^2 = \hat{E}[(y_i - \hat{\mu}_i)^2|\hat{w}_i, d = 0]$. In our programme, both $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are obtained from Nadaraya-Watson kernel regression using the Epanechnikov kernel, where the bandwidth is chosen by a Silverman (1986)-type rule of thumb for Epanechnikov kernels.[10] Finally, the variance of $\hat{\theta}$ is approximated by

$$\hat{V}(\hat{\theta}) = \hat{V}\left(E(y_i|d_i = 1)\right) + \sum_{i=1}^{N} (1 - d_i)\hat{w}_{norm,i}^2 \hat{\sigma}_i^2. \tag{11}$$

The second part on the right hand side is the variance estimate of the estimated mean potential outcome under non-treatment. Note that $\hat{w}_{norm,i}$ is the normalized weight of the ATET estimator based on $\hat{w}_i$. The normalization is such that the non-treated weights add up to unity: $\hat{w}_{norm,i} = \hat{w}_i \big/ \sum_{j=1}^{N} (1 - d_j)\hat{w}_j$. Even though (10) might be a reasonable approximation, it has to be stressed that it is not a consistent variance estimator. Firstly, it omits the fact that the propensity score entering the matching weights is itself an estimate which in general affects the distribution of $\hat{\theta}$. Secondly, also the bias correction may affect the variance, which is not considered in (11). Thirdly, if the bias correction is based on a logit regression (under binary outcomes), the matching weights taken for inference are those obtained prior to the bias correction and may therefore differ somewhat from the final matching weights. In contrast, under linear bias correction the (correct) matching weights after bias correction are used.

As an alternative to analytical approximations, inference for matching is frequently based on the bootstrap (see Efron 1979, or Horowitz (2001), and MacKinnon (2006), for more recent surveys in economics). This is in spite of the results of Abadie and Imbens (2008), which suggest that the bootstrap may not be valid for standard (i.e. pair or $1 : M$) matching because of the non-smoothness of the estimator. However, the LMW11 estimator is by construction smoother thanks to a variable number of (weighted) matched controls and the regression-based bias adjustment. Therefore, the bootstrap appears to be an attractive inference method, which we recommend in applications rather than relying on the approximation in (11). In contrast to the latter, the bootstrap is consistent because it accounts for the estimation of the propensity score and all further issues raised before.

While one could in principle bootstrap the ATET estimate directly to obtain standard errors and p-values, the bootstrap is known to have better properties when using a pivotal statistic such as the t-statistic. We, therefore, suggest computing the t-statistic based on the variance estimator in (11) as first step of the bootstrap procedure: $T_N = \frac{\hat{\theta}}{\sqrt{\hat{V}(\hat{\theta})}}$. In the second step, one randomly draws B bootstrap samples of size N with replacement to compute the ATET $\hat{\theta}^b$ as well as the t-statistic $T_N^b = \frac{\hat{\theta}^b - \hat{\theta}}{\sqrt{\hat{V}(\hat{\theta}^b)}}$ in each draw, where b is the index of the bootstrap sample, $b \in \{1, 2, ..., B\}$. Finally, accounting for the fact that the t-statistic is symmetrically distributed around zero, the

[10] $\hat{\sigma}_i^2$ may also be obtained from different methods as for instance the Abadie and Imbens (2006) variance estimator based on matching within the same treatment group.

p-value is computed as the share of absolute bootstrap t-statistics that are larger than the absolute value of the t-statistic in the original sample:

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} \underline{1}\left(\left|T_N^b\right| > |T_N|\right), \tag{12}$$

where $|\cdot|$ denotes the absolute value of the argument.

Analogously, the proposed method can be used for the estimation of the variance of the ATENT. Concerning the variance of the ATE, one may replace (11) by

$$\sum_{i=1}^{N} d_i \hat{w}_{1,norm,i}^2 \hat{\sigma}_i^2 + \sum_{i=1}^{N} (1-d_i)\hat{w}_{0,norm,i}^2 \hat{\sigma}_i^2, \tag{13}$$

with $\hat{w}_{1,norm,i}$, $\hat{w}_{0,norm,i}$ being the normalized matching weights of the ATE estimator, where the normalization is such that the weights add up to unity within the treatment and control groups, respectively. Thus, Eq. (13) approximates the sum of the variances of the mean potential outcomes under treatment and control of the entire population. All remaining steps are equivalent to those of the inference for the ATET.

3 Empirical Monte Carlo Study

3.1 Idea and data base

In contrast to conventional simulation studies where all features of the data generating process (DGP) are specified by the researcher, the idea of an *Empirical Monte Carlo Study* (EMCS) is to exploit empirical data (e.g. observed outcomes and covariates) to better imitate real world applications when investigating the finite sample behaviour of estimators.[11] As in HLW13, the simulations in this paper are based on a large German administrative data set, which consists of a 2 % random sample of employees subject to social insurance[12] from 1990 to 2006 and combines information from four different sources: (i) employer-provided employee records to the social insurance agency (1990–2006), (ii) unemployment insurance records (1990–2006), (iii) the programme participation register of the Public Employment Service (PES, 2000–2006) and (4) the jobseeker register of the PES (2000–2006).[13] As in LMW11 and Lechner and Wunsch (2009b), those individuals who start training courses that provide job-related vocational classroom training[14] within the first 12 months of unemployment are defined as treated (3,266 observations). The non-treated are those not participating in any active labour market programme in the same period (114,349).

[11] Papers with related approaches include Abadie and Imbens (2002), Bertrand et al. (2004), Diamond and Sekhon (2008), Lee and Whang (2009), Khwaja et al. (2010) and Huber (2012).

[12] This covers 85 % of the German workforce. It excludes the self-employed as well as civil servants.

[13] Further details regarding the data can be found in Appendix 2.

[14] The programmes we consider correspond to *general training* in Wunsch and Lechner (2008) and to *short and long training* in LMW11.

3.2 Simulation design

The EMCS in HLW13 consists of three steps: (i) estimation of the propensity score (the conditional probability to receive the training) in the 'population', which is then considered to be the true propensity score in the simulations; (ii) drawing a sample of control observations in which a (placebo-)treatment is simulated and the treatment effect is estimated (with the true effect being zero by definition); and (iii) repeating the second step many times to assess the performance of the estimators.

Selection into treatment, which is relevant for step (i), is displayed in Table 2. Firstly, the upper part presents descriptive statistics for the two outcome variables

Table 2 Descriptive statistics of the 'population'

Variable	Treated		Control		Standardized difference in %	Probit estimation of selection equation	
	Mean	Std.	Mean	Std.		Marg. eff. in %	Std. error
3 Years since beginning of UE spell some unsubsidized employ.	0.63	0.48	0.56	0.50	9	–	–
Av. monthly earnings (EUR)	1193	1115	1041	1152	9	–	–
Age / 10	3.67	0.84	3.56	1.11	8	7.3	0.5
… squared / 1000	1.42	0.63	1.39	0.85	3	−9.1	0.6
20–25 years old	0.22	0.41	0.36	0.48	22	0.9	0.2
Women	0.57	0.50	0.46	0.50	15	−5.8	1.5
Not German	0.11	0.31	0.19	0.39	16	−0.5	0.1
Secondary degree	0.32	0.47	0.22	0.42	15	1.1	0.1
University entrance qualification	0.29	0.45	0.20	0.40	15	1.0	0.1
No vocational degree	0.18	0.39	0.34	0.47	26	−0.3	0.1
At least one child in household	0.42	0.49	0.28	0.45	22	−0.2	0.1
Last occupation: Non-skilled worker	0.14	0.35	0.21	0.41	13	0.3	0.1
Last occupation: Salaried worker	0.40	0.49	0.22	0.41	29	1.8	0.2
Last occupation: part time	0.22	0.42	0.16	0.36	12	2.1	0.3
UI benefits: 0	0.33	0.47	0.44	0.50	16	−0.6	0.1
>650 EUR per month	0.26	0.44	0.22	0.41	7	0.7	0.1
Last 10 years before UE: share empl.	0.49	0.34	0.46	0.35	8	−1.4	0.2
Share unemployed	0.06	0.11	0.06	0.11	1	−2.5	0.5
Share in programme	0.01	0.04	0.01	0.03	9	5.1	1.2
Last year before UE: share minor em*	0.07	0.23	0.03	0.14	15	−1.0	0.7
Share part time	0.16	0.33	0.11	0.29	10	−1.0	0.2

Table 2 continued

Variable	Treated		Control		Standardized difference in %	Probit estimation of selection equation	
	Mean	Std.	Mean	Std.		Marg. eff. in %	Std. error
Share out-of-the labour force (OLF)	0.28	0.40	0.37	0.44	14	−1.3	0.2
Entering UE in 2000	0.26	0.44	0.19	0.39	13	1.6	0.2
2001	0.29	0.46	0.26	0.44	5	0.9	0.1
2003	0.20	0.40	0.27	0.44	12	0.0	0.1
Share of pop. living in/ close to big city	0.76	0.35	0.73	0.37	6	0.4	0.1
Health restrictions	0.09	0.29	0.15	0.36	13	−0.6	0.1
Never out of labour force	0.14	0.34	0.11	0.31	6	0.6	0.2
Part time in last 10 years	0.35	0.48	0.29	0.45	9	−0.5	0.1
Never employed	0.11	0.31	0.20	0.40	17	−1.0	0.1
Duration of last employment > 1 year	0.41	0.49	0.43	0.50	4	−0.6	0.1
Av. earn. last 10 years when empl./1,000	0.59	0.41	0.52	0.40	13	−0.4	0.2
Women x age / 10	2.13	1.95	1.65	1.94	17	2.6	0.6
x squared / 1000	0.83	0.85	0.65	0.90	15	−2.6	0.8
x no vocational degree	0.09	0.28	0.16	0.36	15	−0.9	0.1
x at least one child in household	0.32	0.47	0.17	0.37	25	0.9	0.2
x share minor employment last year	0.06	0.22	0.02	0.13	16	3.2	0.7
x share OLF last year	0.19	0.36	0.18	0.35	3	1.0	0.2
x average earnings last 10 years. if empl.	0.26	0.34	0.19	0.30	16	−1.0	0.2
x entering UE in 2003	0.10	0.30	0.13	0.33	6	−0.6	0.1
$x_i\hat{\beta}$	−1.7	0.42	−2.1	0.42	68	–	–
$\Phi(x_i\hat{\beta})$	0.06	0.03	0.05	0.03	59	–	–
Number of obs., Pseudo-R^2in %	3266		114349			3.6	

* Minor em. is minor employment with earnings of no more than 400 EUR per month, which are not or only partially subject to social insurance contributions. 'binary': indicates a binary variable (standard deviation can be directly deduced from mean). $\hat{\beta}$ is the estimated probit coefficients and $\Phi(a)$ is the c.d.f. of the standard normal distribution evaluated at *a*. *Pseudo-*R^2 is the so-called Efron's R^2 $\left(1-\sum_{i=1}^{N}\left[d_i-\hat{p}(x_i)\right]/\sum_{i=1}^{N}\left[d_i-\sum_{i=1}^{N}(d_i)/N\right]\right)$. The *Standardized difference* is defined as the difference of means normalized by the square root of the sum of estimated variances of the particular variables in both subsamples (see e.g. Imbens and Wooldridge 2009, p. 24). Marg. effect: Average marginal effects based on discrete changes for binary variables and derivatives otherwise

considered: average monthly earnings over the three years after entering unemployment (semi-continuous with 50 % zeros), and an indicator whether there has been some (unsubsidized) employment in that period (binary). Secondly, Table 2 includes the descriptive statistics for the 38 confounders (among these eight interaction terms) that are considered in the 'true' selection equation for the estimation of the propensity score.[15] It also contains the normalized differences between treated and controls as well as the marginal effects of the covariates at the means of all other covariates according to the estimation of the true propensity score. Both results suggest considerable selection into treatment due to imbalances in several variables.

After having estimated the propensity in the full population, the treated are discarded and no longer play a role in the simulations. The next step is to draw a random sample of size N from the population of controls (independent draws with replacement). HLW13 use sample sizes of 300, 1,200 and 4,800 and thoroughly motivate this choice. In each sample, (pseudo-) treated observations are simulated based on the propensity score in the population. For each individual in the sample, $\hat{p}_i(x_i) = \Phi(x_i\hat{\beta})$ is computed, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, x_i is the observed covariate vector of observation i (including the constant), and $\hat{\beta}$ are the coefficient estimates. We consider three choices of selection into treatment based on the following equation:

$$d_i = \underline{1}(\lambda\, x_i\hat{\beta} + \alpha + u_i > 0), \quad u_i \sim N(0,1), \quad \lambda \in \{0, 1, 2.5\}, \tag{14}$$

where u_i denotes a standard normally distributed i.i.d. random number and λ determines the magnitude of selection (0 = random, 1 = observed, and 2.5 = strong selection). Finally, α gauges the shares of treated and controls. It is chosen such that the expected number of treated equals 10, 50, or 90 %, respectively.[16] Note that due to the assignment of a pseudo-treatment, the true treatment effect on any individual in any scenario is zero.

At least in expectation, this simulation routine ensures common support. Nevertheless, when strong selection is combined with the large share of treated, overlap of the distributions of the propensity score in the treated and control sample becomes very thin in the right tail of the treated population, as documented in HLW13. In addition, combining the small sample size with extreme shares of participants would frequently include cases in which the number of covariates exceeds the number of treated or non-treated observations. Hence, in the small sample the unconditional treatment probability is 0.5. Table 3 summarizes the 21 scenarios that are used in the EMCS and gives statistics about the amount of selection implied by each.[17]

[15] Note that the descriptive statistics in Table 2 seemingly differ from those in Table 1 of HLW13, even though they refer to the same data. The reason is that in HLW13, the non-treated covariate means are incorrectly displayed in the column which claims to provide the standard deviations of the covariates of the treated, while the latter are given in the column which claims to show the non-treated covariate means. Therefore, Table 2 is correct, while the statistics in Table 1 of HLW13 are partially misplaced.

[16] Note that the simulations are not conditional on D. Thus, the share of treated in each sample is random.

[17] The standardized differences as well as the pseudo-R^2s are based on a re-estimated propensity score in the population with simulated treated (114,349 obs.). However, when reassigning controls to act as simulated treated this changes the control population. Therefore, this effect, and the fact that the share of

Table 3 Summary statistic of DGP's

Magnitude of selection	Share of treated in %	Standardized difference of p score	Pseudo-R^2 of probit in %	Sample size
Random	10	0	0	1200, 4800
	50	0	0	300, 1200, 4800
	90	0	0	1200, 4800
Observed	10	0.5	6	1200, 4800
	50	0.4	10	300, 1200, 4800
	90	0.5	6	1200, 4800
Strong	10	1.1	27	1200, 4800
	50	0.8	36	300, 1200, 4800
	90	0.8	27	1200, 4800

See note of Table 2

In the analysis, we investigate performance not only when using the correct propensity score model, but also under misspecification omitting the eight interaction terms and the two higher order terms of age. As in HLW13, the number of Monte Carlo replications is proportional to the sample size, consisting of 16,000 replications for the small, 4,000 for the medium and 1,000 for the large sample. The latter is computationally most expensive, but has the least variability in results across simulation samples.

4 Results

This section discusses how the properties of the DGP and the four tuning parameters affect the small sample behaviour of the LMW11 estimator. The latter parameters are the radius size, which is determined by (i) the distance quantile and (ii) the radius multiplier, (iii) the additional covariates in Mahalanobis matching and in the regression adjustment and (iv) the weight of the propensity score relative to the additional covariates. Concerning the choice of the distance quantile, the values at the 0.1, 0.5, and 0.9 quantiles of the distribution of minimum distances in pair matching are considered. To obtain the radius size, the quantile is multiplied by the radius multiplier which is set to 0.25, 1, 10 and 100 in the simulations. We therefore cover a more extensive range of radius sizes than HLW13, who only investigated three choices: 0.5, 1.5 and 3 times the maximum distance in pair matching. Note that if a radius is empty, which may happen only if the product of the distance quantile and the multiplier is smaller than the maximum distance, the algorithm picks the nearest control.

With regard to additional covariates in the Mahalanobis metric and the regression adjustment, we consider 0 (propensity score matching only), 1 (woman) and 4 covariates (woman, no vocational degree, UI benefits of zero, average earnings in the last 10

Footnote 17 continued
treated differs from the original share leads to different values of those statistics even in the case that mimics selection in the original population.

years when employed / 1000). To alter the weight of the propensity score in the metric, the inverse of its variance is multiplied by 0.5 (propensity score receives less weight than the covariates), 1 (propensity score and each covariate are equally weighted) and 5 (propensity score receives more weight than the covariates).

All results are based on trimming as described in Section 2.4 and Eq. (9), with the trimming threshold set to $t = 4\,\%$. This choice has been made because it dominated the non-trimmed version of the estimator as well as larger t (e.g. 6 %) in HLW13 in terms of the mean squared error (RMSE). Furthermore, we remove all treated units with larger propensity scores than the largest control observation prior to matching. Moreover, we use bias adjustment based on logit regression (for the binary employment outcome) and OLS (for earnings), as this resulted in a lower RMSE of the estimator in HLW13 than an unadjusted version. Table 4 presents the impact of the DGP features and the tuning parameters of the estimator on the RMSE, whereas the results for the bias and the standard deviation are presented in Appendix 1. Similarly to HLW13, the analysis is based on an OLS regression in which the RMSE is the outcome variable and the DGP features and tuning parameters serve as regressors. All in all, our simulations provide us with 648 data points in the small sample and 1,944 in the medium and large samples (which consider more shares of treated). As expected, the baseline RSME, which is captured by the constant, decreases in the sample size for both the binary outcome (employment) and the semi-continuous outcome (earnings) and does so roughly at root-N rate. Taking a look at the DGP features, we see that a stronger selection into treatment significantly increases the RMSE across all sample sizes and outcomes (the reference point is the selectivity observed in the data, i.e. $\lambda = 1$). This is due to both a larger bias and a higher standard deviation (see Tables 7 and 8 in the Appendix). With regard to the share of treated, the estimator performs best in terms of the RMSE for a share of 50 %. Even though the bias is slightly (but not significantly) larger than for 10 % treated, where in both relative and absolute terms more potential matches are available, the standard deviation is considerably lower due to a higher number of treated observations. The 90 % share does worse than the 50 % share in terms of bias and standard deviation, as too few comparisons among the controls are available. In conclusion, none of the effects of the DGP features comes with a surprise.

Under the misspecification of the propensity score, the bias is increased because an incorrect functional form is assumed. At the same time, the propensity score is more precisely estimated due to omitting the interaction and higher order terms of covariates, which also reduces the variance of the radius matching estimator. In the smaller sample, the variance reduction outweighs the bias increase such that misspecification reduces the RMSE. In the medium and large samples, the contrary holds true.

We now analyse the impact of the tuning parameters, starting with the additional covariates. For both outcomes the RMSE decreases in the number of covariates in the Mahalanobis metric and the regression/logit adjustment suggesting that controlling for the most important confounders may be beneficial, as long as the curse of dimensionality does not kick in. The reduction is largest in the small sample. As Table 8 reveals, the effect is primarily driven by a reduction in the standard deviation (in particular when using four covariates). The impact on the bias is more ambiguous. For employment, it is significantly negative when using one covariate, but insignificant when using four. For earnings, it is economically negligible and insignificant in any sample

Table 4 Impact of the features of the DGP and the estimator on the RMSE (OLS regression)

	Employment			Earnings		
	300	1200	4800	300	1200	4800
Constant	8.95***	4.05***	1.87***	207.32***	101.02***	46.92***
Features of the data generating process						
Selection						
Random	−0.52***	−0.50***	−0.67***	−20.54***	−20.33***	−22.60***
Observed	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Strong	1.95***	1.79***	2.03***	43.26***	47.39***	52.99***
Share treated						
10 %		1.90***	0.73***		52.34***	20.87***
50 %		Ref.	Ref.		Ref.	Ref.
90 %		2.89***	1.75***		52.64***	39.04***
Misspecified p score	−0.75***	0.23***	0.95***	−8.76***	11.59***	27.12***
Features of the estimator						
Additional matching variables						
0 (only p score)	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
1	−0.23***	−0.05**	−0.09**	−7.27***	−2.79***	−0.20
4	−1.31***	−0.61***	−0.41***	−33.53***	−18.26***	−8.90***
Scoreweight						
0.5	0.01	0.00	−0.00	0.15	0.08	−0.01
1	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
5	0.00	0.01	0.01	−0.07	0.10	0.26

Table 4 continued

	Employment			Earnings		
Radius (quantile × multiplier)						
0.1 × 0.25	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1 × 1	0.00	−0.00	0.00	−0.01	−0.01	−0.01
0.1 × 10	−0.03	−0.02	−0.01	−0.64	−0.38	−0.11
0.1 × 100	−0.26***	−0.15***	−0.05	−5.55***	−3.02***	−0.96
0.5 × 0.25	−0.02	−0.02	−0.01	−0.36	−0.32	−0.12
0.5 × 1	−0.07	−0.05	−0.03	−1.52	−1.15	−0.50
0.5 × 10	−0.49***	−0.33***	−0.13	−10.61***	−6.93***	−2.86
0.5 × 100	−0.97***	−0.62***	−0.23***	−21.66***	−13.84***	−5.81***
0.9 × 0.25	−0.20***	−0.18***	−0.08	−4.37***	−3.85***	−1.77
0.9 × 1	−0.49***	−0.37***	−0.16**	−10.67***	−7.99***	−3.67*
0.9 × 10	−1.06***	−0.70***	−0.27***	−23.90***	−15.79***	−7.01***
0.9 × 100	−1.22***	−0.79***	−0.29***	−27.86***	−18.95***	−8.77***
Statistics						
Observations	648	1,944	1,944	648	1,944	1,944
Adjusted *R*-squared	0.96	0.92	0.80	0.97	0.96	0.82

Dependent variable: RMSE. *ref* reference group. Significance levels are indicated as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

size. In contrast to the number of covariates, the values of the propensity score weights considered in the simulations do not play any role. The effects on the RMSE, bias and standard deviation are consistently close to zero and insignificant in all scenarios.

Finally, we consider the 12 different combinations of the distance quantile and the multiplier that determine the radius size. The clear cut result of our simulations is that the larger the radius, the smaller the RMSE. For any sample size and outcome, increasing the quantile while holding the multiplier fixed or doing it vice versa reduces the RMSE. This is entirely driven by a decrease in the standard deviation, as a larger radius uses more controls for the estimation of the local mean outcome under non-treatment and therefore increases precision. On the other hand, including controls that are more distant and thus, less comparable to the treated observations may increase bias, giving rise to a potential bias-variance-trade-off. However, Table 8 shows that the bias is not significantly affected by the radius size in any but the two cases with the largest radius. Clearly, this finding is dependent on the ability of the parametric bias removal to be effective. That is, in the DGPs considered, using a larger radius does not come with the cost of an increased bias, but allows realizing gains in efficiency such that the RMSE is reduced. Note that this need not hold for estimation without any bias correction (which is not considered in this paper), where the unadjusted use of more distant and less comparable controls can possibly entail a larger bias. In this light, a bias adjustment appears particularly advisable in the case of a large radius size (leading to heavy oversmoothing).

Our results on the effects of additional covariates and the radius size suggest that the regression/logit adjustment performs well in terms of reduction of the RMSE. We arrive at this conclusion because additional covariates and a larger radius implicitly shift more weight to the parametric component of the estimator. In particular, using the 0.9 quantile (of minimum distances in pair matching) times 100 approaches global parametric estimation due to the large radius size. Therefore, our findings are in line with those of HLW13 showing that the parametric OLS and logit estimators (although more flexibly specified than here) fair surprisingly well when estimating the ATET.

Tables 5 and 6 present the effects of the number of covariates in the Mahalanobis metric/regression adjustment and of the radius multiplier on the behaviour of the estimator in terms of RMSE, bias, standard deviation, skewness and kurtosis within strata defined by the sample size, selection into treatment, correct/incorrect propensity score specification and the share of treated (the latter for the medium and large sample sizes only). This allows investigating the heterogeneity of the effects across DGP features, while averaging over all remaining tuning parameters, e.g. the choices of the distance quantile and the propensity score weight. Note that the propensity score weight itself is no longer investigated due to its obvious irrelevance, at least for the values considered. In Table 5, the number of additional covariates in the Mahalanobis metric and adjustment procedure is varied. Clearly, choosing four covariates performs best in terms of the RMSE in any stratum and for both outcomes. This result is driven by a considerable reduction in the standard deviation, while the bias is often non-monotonic in the number of predictors, but overall barely affected.

A similar picture arises when looking at the impact of the multiplier in Table 6, where the distance quantile is now fixed at 0.9 (rather than averaging over all choices),

Table 5 Estimator properties as function of additional covariates

	Employment					Earnings				
	RMSE	Bias	Std. dev.	Skew	Kurtosis	RMSE	Bias	Std. dev.	Skew	Kurtosis
Covars in Mahal	$N = 300$									
0	8.6	1.4	8.5	0.1	5.5	201.5	33.3	195.9	−0.3	5.5
1	8.4	1.3	8.3	−0.1	4.2	194.2	33.6	188.6	−0.2	3.7
4	7.3	1.3	7.2	0.1	3.0	167.9	32.8	162.4	−0.0	3.0
$N = 1{,}200$										
0	5.9	1.5	5.5	0.1	3.0	144.8	38.9	134.5	−0.1	3.1
1	5.9	1.4	5.5	0.1	3.0	142.0	39.5	130.7	−0.1	3.1
4	5.3	1.5	5.0	0.1	3.0	126.5	38.0	115.0	−0.0	3.1
$N = 4{,}800$										
0	3.5	1.5	2.8	0.1	3.0	88.0	35.1	72.3	−0.0	3.0
1	3.4	1.4	2.8	0.1	3.0	87.8	37.7	70.1	−0.1	3.1
4	3.1	1.4	2.5	0.1	3.0	79.1	36.6	60.7	−0.0	3.0
Normal selection										
0	4.8	1.2	4.5	0.1	3.4	118.8	29.9	112.0	−0.1	3.4
1	4.7	1.2	4.5	0.1	3.2	116.4	32.3	107.9	−0.1	3.2
4	4.3	1.2	4.0	0.1	3.0	103.9	30.6	95.4	0.0	3.0
No selection										
0	4.2	0.1	4.2	0.0	3.5	97.1	2.7	97.1	−0.1	3.6
1	4.2	0.1	4.2	0.0	3.3	94.8	3.3	94.7	−0.1	3.2
4	3.7	0.3	3.7	0.0	3.0	83.1	3.0	83.0	−0.0	3.1
Heavy selection										
0	6.9	3.1	5.7	0.1	3.2	169.8	76.8	140.8	−0.1	3.2
1	6.7	2.9	5.7	0.1	3.0	167.5	78.0	136.4	−0.1	3.1
4	6.0	2.8	5.0	0.1	3.0	149.4	76.3	117.0	−0.1	3.0
Correctly specified pscore										
0	5.1	0.7	5.0	0.1	3.7	123.5	14.2	121.9	−0.1	3.8
1	5.0	0.7	4.9	0.0	3.3	118.4	12.6	117.0	−0.1	3.3
4	4.4	0.8	4.3	0.1	3.0	102.0	13.3	100.0	−0.0	3.0
Misspecified pscore										
0	5.5	2.3	4.6	0.1	3.0	133.6	58.7	111.3	−0.1	3.0
1	5.4	2.1	4.6	0.1	3.0	134.1	63.1	109.1	−0.1	3.1
4	4.9	2.0	4.2	0.1	3.0	122.2	60.0	97.0	−0.0	3.0
10 % treated										
0	4.8	1.2	4.4	0.1	3.1	125.7	29.6	118.4	0.0	3.0
1	4.7	1.1	4.4	0.1	3.0	122.9	31.7	113.0	0.0	3.1
4	4.3	1.1	4.0	0.1	3.0	112.9	32.3	101.8	0.0	3.0
50 % treated										
0	3.5	1.2	2.9	0.1	3.0	87.3	30.3	74.2	−0.1	3.1

Table 5 continued

	Employment					Earnings				
	RMSE	Bias	Std. dev.	Skew	Kurtosis	RMSE	Bias	Std. dev.	Skew	Kurtosis
1	3.4	1.1	2.9	0.1	3.0	86.5	31.8	71.8	−0.1	3.1
4	3.1	1.2	2.6	0.0	3.0	77.9	31.0	62.6	−0.0	3.0
90 % treated										
0	5.9	2.1	5.2	0.1	2.9	136.3	51.1	117.7	−0.1	3.0
1	5.8	2.0	5.2	0.1	2.9	135.3	52.2	116.5	−0.2	3.1
4	5.2	2.0	4.6	0.1	3.0	117.7	48.5	99.1	−0.1	3.1

** Contains only results for $N = 1{,}200$ and $N = 4{,}800$

Table 6 Estimator properties as function of the radius multiplier

	Employment					Earnings				
	RMSE	Bias	Std. dev.	Skew	Kurtosis	RMSE	Bias	Std. dev.	Skew	Kurtosis
Radius multiplier	$N = 300$									
0.25	8.3	1.3	8.2	0.0	4.1	192.4	33.1	187.1	−0.1	4.0
1	8.0	1.3	7.9	0.0	4.3	186.1	33.0	180.6	−0.1	4.1
10	7.5	1.4	7.3	0.0	4.6	172.9	33.7	166.6	−0.1	4.4
100	7.3	1.4	7.1	0.0	4.8	168.9	33.7	162.4	−0.1	4.5
$N = 1{,}200$										
0.25	5.8	1.4	5.4	0.1	3.0	139.9	38.4	129.3	−0.1	3.1
1	5.6	1.4	5.2	0.1	3.0	135.8	38.5	124.8	−0.1	3.1
10	5.3	1.6	4.9	0.1	3.0	128.0	39.9	115.7	−0.1	3.1
100	5.2	1.6	4.7	0.1	3.0	124.8	40.5	112.1	−0.1	3.1
$N = 4{,}800$										
0.25	3.4	1.4	2.8	0.1	3.0	85.9	36.2	68.8	−0.0	3.0
1	3.3	1.4	2.7	0.1	3.0	84.0	36.3	66.6	−0.0	3.0
10	3.2	1.5	2.5	0.1	3.0	80.6	37.2	62.2	−0.0	3.0
100	3.2	1.6	2.4	0.1	3.0	78.9	38.1	59.6	−0.0	3.0
Normal selection										
0.25	4.7	1.2	4.4	0.1	3.2	114.9	30.5	107.3	−0.1	3.2
1	4.5	1.2	4.2	0.1	3.2	111.5	30.7	103.5	−0.0	3.2
10	4.3	1.3	3.9	0.1	3.3	105.2	31.9	96.2	−0.0	3.3
100	4.2	1.3	3.8	0.1	3.3	102.7	32.5	93.3	−0.0	3.3
No selection										
0.25	4.1	0.1	4.1	0.0	3.3	93.7	2.8	93.6	−0.1	3.3
1	3.9	0.1	3.9	0.0	3.3	89.8	2.7	89.7	−0.1	3.3
10	3.7	0.3	3.6	0.0	3.3	83.3	3.5	83.2	−0.0	3.4
100	3.6	0.4	3.6	0.0	3.4	81.3	4.3	81.1	−0.0	3.4

Table 6 continued

	Employment					Earnings				
	RMSE	Bias	Std. dev.	Skew	Kurtosis	RMSE	Bias	Std. dev.	Skew	Kurtosis
Heavy selection										
0.25	6.6	2.9	5.6	0.1	3.0	164.2	76.7	134.1	−0.1	3.1
1	6.5	2.9	5.4	0.1	3.0	161.1	77.0	130.3	−0.1	3.1
10	6.1	3.0	5.0	0.1	3.1	153.9	78.2	120.7	−0.1	3.1
100	6.0	3.0	4.8	0.1	3.1	150.3	78.7	116.0	−0.1	3.1
Correctly specified pscore										
0.25	4.9	0.7	4.8	0.1	3.3	117.0	13.3	115.3	−0.1	3.3
1	4.8	0.7	4.7	0.1	3.4	113.1	13.3	111.4	−0.1	3.4
10	4.4	0.7	4.3	0.1	3.5	105.1	13.6	103.4	−0.1	3.4
100	4.3	0.8	4.2	0.0	3.5	101.6	14.0	99.8	−0.0	3.5
Misspecified pscore										
0.25	5.3	2.1	4.5	0.1	3.0	131.5	60.0	108.0	−0.1	3.0
1	5.2	2.1	4.4	0.1	3.0	128.4	60.3	104.2	−0.1	3.0
10	5.0	2.3	4.1	0.1	3.0	123.1	62.2	96.7	−0.1	3.0
100	4.9	2.4	3.9	0.1	3.0	121.3	63.1	93.8	−0.0	3.0
Radius multiplier	10 % treated**									
0.25	4.7	1.1	4.3	0.1	3.0	121.9	30.7	112.6	0.0	3.0
1	4.5	1.1	4.1	0.1	3.0	117.2	31.0	107.6	0.0	3.0
10	4.2	1.2	3.9	0.1	3.0	110.6	32.5	100.2	0.0	3.1
100	4.2	1.4	3.8	0.1	3.0	108.6	34.2	97.4	0.0	3.1
50 % treated**										
0.25	3.3	1.2	2.9	0.1	3.0	84.8	30.6	70.7	−0.1	3.0
1	3.3	1.2	2.8	0.1	3.0	83.0	30.7	68.6	−0.1	3.0
10	3.1	1.3	2.6	0.1	3.0	79.6	32.1	64.2	−0.0	3.1
100	3.1	1.4	2.5	0.0	3.0	77.5	33.0	61.4	−0.0	3.1
90 % treated**										
0.25	5.7	2.0	5.1	0.1	2.9	132.0	50.5	113.9	−0.1	3.1
1	5.6	2.0	5.0	0.1	2.9	129.5	50.5	110.9	−0.1	3.1
10	5.3	2.1	4.6	0.1	3.0	122.7	51.1	102.4	−0.1	3.0
100	5.2	2.1	4.4	0.1	3.0	119.4	50.7	98.7	−0.1	3.0

Contains only specifications with rquantil = 0.9 as a larger radius always dominates a smaller one
** Contains only results for $N = 1{,}200$ and $N = 4{,}800$

as higher quantiles always dominate lower ones (given equal multipliers). The RMSE decreases in the radius size in any scenario. Even though the bias generally increases slightly, this is more than offset by a reduction in the standard deviation. Interestingly, the decrease of the RMSE is much larger when switching from 1 to 10 than when switching from 10 to 100, suggesting that the marginal effect of further increases of the radius is a decreasing function. Finally, we take a look at the skewness and kurtosis of the estimator, telling us whether it is approximately normally distributed. In general,

this appears to be the case. The skewness is always close to zero and the kurtosis is close to three in most scenarios and only somewhat higher in the small sample. In the latter case, a larger number of covariates in the Mahalanobis metric/regression adjustment shifts the kurtosis back to three, while a larger radius size appears to slightly shift the kurtosis further away from that of a normal distribution.

In conclusion, the EMCS suggests that Mahalanobis matching on the propensity score and several important covariates is preferable to matching on the propensity score only. Secondly, a radius that is at least several times larger than the maximum distance in pair matching appears to be superior to smaller choices, at least in the DGPs and empirical data considered in our simulation design.

5 Conclusion

In this paper, we investigate the finite sample properties of a distance-weighted radius matching estimator with regression-based bias adjustment proposed in LMW11 using a simulation design based on empirical labour market data as suggested in HLW13. We find that the choice of tuning parameters, such as the radius size, and whether matching is on the propensity score only or additionally also on the most important confounders via the Mahalanobis metric affects the performance of the estimator, in particular its root mean squared error. Across all simulations, our results consistently suggest picking a large radius dominates smaller choices. Likewise (and related), including the most important covariates (on top of the propensity score) in the matching algorithm and the regression adjustment performs always well in terms of the root mean squared error. Because increasing the radius and the number of covariates implicitly shifts more weight to the parametric regression adjustment, our results suggest that the latter performs well in terms of reducing the RMSE. Therefore, combining radius matching and regression in an appropriate way appears to improve estimation. The study also reveals that the estimator is close to being normally distributed in almost all scenarios. The estimator is available as GAUSS, STATA and R code. It includes options for the choice of the various tuning parameters, common support procedures and inference methods.

Acknowledgments Martin Huber gratefully acknowledges financial support from the Swiss National Science Foundation grant PBSGP1_138770. We would like to thank Conny Wunsch (SEW) for her help in the early stages of the paper.

Appendix 1: More details on the features of the DGP and the estimator

Table 7 Impact of the features of the DGP and the estimator on the bias (OLS regression)

		Employment			Earnings		
		300	1,200	4,800	300	1,200	4,800
Constant		0.89***	0.38***	−0.03	12.37***	4.08*	−6.65**
		Features of the data generating process					
Selection	Random	−0.94***	−1.09***	−1.01***	−25.61***	−29.89***	−26.71***
	Observed	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	Strong	1.57***	1.77***	1.77***	41.40***	46.38***	47.42***
Share treated	10 %		−0.06	−0.04		2.51**	−2.19
	50 %		Ref.	Ref.		Ref.	Ref.
	90 %		0.87***	0.82***		18.75***	20.32***
Misspecified *p* score		0.51***	1.21***	1.96***	31.04***	43.27***	56.51***
		Features of the estimator					
Additional matching	0 (only *p* pscore)	Ref.	Ref.	Ref.	ref.	Ref.	Ref.
variables	1	−0.08***	−0.08*	−0.11**	0.25	0.54	2.63*
	4	−0.02	−0.06	−0.10*	−0.49	−0.93	1.49
Scoreweight	0.5	−0.00	0.00	−0.00	−0.04	0.03	−0.02
	1	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	5	0.02	0.01	0.01	0.38	0.17	0.27
Radius (quantile × multiplicator)	0.1 × 0.25	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	0.1 × 1	−0.00	0.00	−0.00	−0.01	0.00	−0.01
	0.1 × 10	−0.00	0.00	−0.00	−0.05	0.02	−0.00
	0.1 × 100	−0.01	0.00	−0.00	−0.11	0.14	−0.00
	0.5 × 0.25	−0.00	0.00	−0.00	−0.06	0.01	−0.00
	0.5 × 1	−0.01	0.00	−0.00	−0.10	0.07	−0.05

Table 7 continued

		Employment			Earnings		
	0.5×10	−0.02	0.01	0.01	−0.15	0.35	0.20
	0.5×100	0.06	0.12	0.11	0.46	1.64	1.18
	0.9×0.25	−0.01	0.01	0.01	−0.08	0.12	0.18
	0.9×1	−0.02	0.01	0.02	−0.14	0.27	0.31
	0.9×10	0.05	0.12	0.12	0.52	1.68	1.17
	0.9×100	0.07	0.20**	0.21*	0.53	2.31	2.08
		Statistics					
Observations		648	1,944	1,944	648	1,944	1,944
Adjusted R squared		0.92	0.74	0.71	0.87	0.79	0.71

Dependent variable: Bias. Significance levels are indicated as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 8 Impact of the features of the DGP and the estimator on the std.dev. (OLS reg.)

		Employment			Earnings		
		300	1,200	4,800	300	1,200	4,800
Constant		8.93***	4.05***	1.98***	208.61***	104.05***	54.49***
		Features of the data generating process					
Selection	Random	−0.43***	−0.30***	−0.27***	−17.42***	−13.73***	−11.92***
	Observed	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	Strong	1.63***	1.13***	1.00***	32.83***	27.72***	22.80***
Share treated	10 %		2.04***	0.92***		56.45***	26.67***
	50 %		Ref.	Ref.		Ref.	Ref.
	90 %		2.79***	1.59***		49.78***	33.32***
Misspecified *p* score		−0.90***	−0.20***	−0.15***	−17.20***	−5.91***	−5.07***
		Features of the estimator					
Additional matching	0 (only *p* pscore)	Ref.	Ref.	Ref.	Ref.	Ref.	ref.
Variables	1	−0.22***	−0.01	0.01	−7.28***	−3.85***	−2.15***
	4	−1.31***	−0.59***	−0.32***	−33.50***	−19.60***	−11.59***
Scoreweight	0.5	0.00	−0.00	−0.00	0.10	0.02	−0.05
	1	Rref.	Ref.	Ref.	Ref.	Ref.	Ref.
	5	0.00	0.00	0.01	−0.07	0.10	0.22
Radius (quantile × multiplicator)	0.1 x 0.25	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	0.1 × 1	−0.00	−0.00	−0.00	−0.03	−0.01	−0.02
	0.1 × 10	−0.03	−0.02	−0.01	−0.64	−0.40	−0.15
	0.1 × 100	−0.26***	−0.16***	−0.06	−5.74***	−3.26***	−1.17
	0.5 × 0.25	−0.02	−0.02	−0.01	−0.36	−0.34	−0.17

Table 8 continued

		Employment			Earnings		
		300	1,200	4,800	300	1,200	4,800
Constant		8.93***	4.05***	1.98***	208.61***	104.05***	54.49***
	0.5 × 1	−0.07	−0.06*	−0.03	−1.55	−1.22	−0.61
	0.5 × 10	−0.50***	−0.35***	−0.16***	−10.96***	−7.54***	−3.52***
	0.5 × 100	−1.01***	−0.69***	−0.33***	−22.67***	−15.54***	−7.47***
	0.9 × 0.25	−0.20***	−0.19***	−0.10**	−4.48***	−4.09***	−2.22***
	0.9 × 1	−0.50***	−0.39***	−0.20***	−10.96***	−8.59***	−4.48***
	0.9 × 10	−1.10***	−0.78***	−0.38***	−24.99***	−17.65***	−8.89***
	0.9 × 100	−1.27***	−0.91***	−0.47***	−29.13***	−21.30***	−11.41***
		Statistics					
Observations		648	1,944	1,944	648	1,944	1,944
Adjusted R squared		0.94	0.95	0.82	0.95	0.96	0.90

Dependent variable: standard error. Significance levels are indicated as *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Appendix 2: Dataset description

The data comprise all aspects of an individual's employment, earnings and unemployment insurance history since 1990 (e.g. type of employment such as full/part-time and high/low-skilled, occupation, earnings, type and amount of unemployment insurance benefits and remaining claims), participation in major labour market programmes from 2000 onwards (including the exact start date, end date, planned end date and type of programme), individual characteristics (e.g. date of birth, gender, educational attainment, marital status, number of children, age of youngest child, nationality, occupation, the presence of health impairments and disability status) and job search activities (the type of job looked for such as full/part-time, high/low-skilled and the occupation, mobility within Germany and health impairments affecting employability). Furthermore, a variety of regional variables has been matched to the data, including information about migration and commuting, average earnings, unemployment rate, long-term unemployment, welfare dependency rates, urbanisation codes, and measures of industry structure and public transport facilities.

The sample used for the simulations covers all entries into unemployment in the period 2000–2003, however, excluding East Germany and Berlin since they are still affected by the aftermath of reunification. Furthermore, unemployment entries in January-March 2000 are discarded because with programme information starting only in January 2000, it should be prevented that entries from employment programmes (which we would consider as unemployed) are accidentally classified as entries from unsubsidized employment due to missing information regarding the accompanying programme spell. Entries after 2003 are not considered such that the outcome variables, employment and earnings, are observed for at least three years after entering unemployment. Moreover, the analysis is restricted to the prime-age population aged 20–59 in order to limit the impact of schooling and (early) retirement decisions and to individuals who were not unemployed or in any labour market programme in the last 12 months before becoming unemployed to make the sample more homogeneous. Finally, the very few cases whose last employment was any non-standard form of employment such as internships were excluded.

References

Abadie A, Imbens GW (2006) Large sample properties of matching estimators for average treatment effects. Econometrica 74:235–267

Abadie A, Imbens GW (2008) On the failure of the bootstrap for matching estimators. Econometrica 76:1537–1557

Behncke S, Frölich M, Lechner M (2010a) Unemployed and their case workers: should they be friends or foes? J R Stat Soc Ser A 173:67–92

Behncke S, Frölich M, Lechner M (2010b) A caseworker like me: does the similarity between unemployed and caseworker increase job placements? Econ J 120:1430–1459

Blundell R, Costa Dias M (2009) Alternative approaches to evaluation in empirical microeconomics. J Hum Resour 44:565–640

Busso M, DiNardo J, McCrary J (2009a) Finite sample properties of semiparametric estimators of average treatment effects. J Bus Econ Stat 27:397–415

Busso M, DiNardo J, McCrary J (2009b) New evidence on the finite sample properties of propensity score matching and reweighting estimators, IZA Discussion Paper, 3998

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2009) Dealing with limited overlap in estimation of average treatment effects. Biometrika 96:187–199
Dehejia RH, Wahba S (1999) Causal effects in non-experimental studies: reevaluating the evaluation of training programmes. J Am Stat Assoc 94:1053–1062
Dehejia RH, Wahba S (2002) Propensity score: matching methods for nonexperimental causal studies. Rev Econ Stat 84:151–161
Diamond A, Sekhon JS (2008) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Mimeo, Berkeley
Efron B (1979) Bootstrap methods: another look at the Jackknife. Ann Stat 7:1–26
Frölich M (2004) Finite-sample properties of propensity-score matching and weighting estimators. Rev Econ Stat 86:77–90
Frölich M (2005) Matching estimators and optimal bandwidth choice. Stat Comput 15:197–215
Frölich M (2007) Nonparametric IV estimation of local average treatment effects with covariates. J Econom 139:35–75
Graham BS, Pinto C, Egel D (2012) Inverse probability tilting for moment condition models with missing data. Rev Econ Stud 79:1053–1079
Heckman JJ, Ichimura H, Todd P (1998) Matching as an econometric evaluation estimator. Rev Econ Stud 65:261–294
Heckman JJ, Ichimura H, Smith J, Todd P (1998) Characterizing selection bias using experimental data. Econometrica 66:1017–1098
Heckman JJ, LaLonde R, Smith J (1999) The economics and econometrics of active labor market programs. In: Ashenfelter O, Card D (eds) Handbook of labour economics. Elsevier Science, Amsterdam, pp 1865–2097
Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 2003:1161–1189
Ho D, Imai K, King G, Stuart E (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, political analysis, pp 199–236. 15 Aug 2007
Horowitz JL (2001) The bootstrap. In: Heckman JJ, Leamer E (eds) Handbook of econometrics. Elsevier Science, Amsterdam, pp 3159–3228
Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite population. J Am Stat Assoc 47:663–685
Huber M (2012) Identification of average treatment effects in social experiments under alternative forms of attrition. J Educ Behav Stat 37:443–474
Huber M, Lechner M, Wunsch C (2013) The performance of estimators based on the propensity score. J Econom 175:1–21
Huber M, Lechner M, Wunsch C (2011) Does leaving welfare improve health? Evid Ger Health Econ 20:484–504
Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86:4–29
Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. J Econ Lit 47:5–86
Joffe MM, Ten Have TR, Feldman HI, Kimmel SE (2004) Model selection, confounder control, and marginal structural models. Am Stat 58:272–279
Khwaja A, Salm GPM, Trogdon JG (2010) A comparison of treatment effects estimators using a structural model of AMI treatment choices and severity of illness information from hospital charts. J Appl Econom. doi:10.1002/Jae.1181
Lechner M (2009) Long-run labour market and health effects of individual sports activities. J Health Econ 28:839–854
Lechner M, Wunsch C (2009a) Active labour market policy in East Germany: waiting for the economy to take off. Econ Trans 17:661–702
Lechner M, Wunsch C (2009b) Are training programs more effective when unemployment is high? J Lab Econ 27:653–692
Lechner M, Miquel R, Wunsch C (2011) Long-run effects of public sector sponsored training in West Germany. J Eur Econ Assoc 9:742–784
Lechner M, Strittmatter A (2014) Practical procedures to deal with common support problems in matching estimation. Mimeo,

Lee S, Whang Y-J (2009) Nonparametric tests of conditional treatment effects, Cowles Foundation Discussion Paper 1740
MacKinnon JG (2006) Bootstrap methods in econometrics. Econ Rec 82:2–18
Robins JM, Mark SD, Newey WK (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics 48:479–495
Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55
Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39:33–38
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
Rubin DB (1979) Using multivariate matched sampling and regression adjustment to control bias in observational studies. J Am Stat Assoc 74:318–328
Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London
Wunsch C, Lechner M (2008) What did all the money do? On the general ineffectiveness of recent West German Labour Market Programmes. Kyklos 61:134–174