This is the preprint of "The performance of estimators based on the propensity score, Journal of Econometrics, 175 (1), 1-21, 2013, doi: 10.1016/j.jeconom.2012.11.006".

The performance of estimators based on the propensity score

Martin Huber, Michael Lechner, and Conny Wunsch*

Swiss Institute for Empirical Economic Research

University of St.Gallen

Revised version: June 2012

Date this version was printed: 08 June 2016

Abstract: We investigate the finite sample properties of a large number of estimators for the average treatment effect on the treated that are suitable when adjustment for observed covariates is required, like inverse probability weighting, kernel and other variants of matching, as well as different parametric models. The simulation design used is based on real data usually employed for the evaluation of labour market programmes in Germany. We vary several dimensions of the design that are of practical importance, like sample size, the type of the outcome variable, and aspects of the selection process. We find that trimming individual observations with too much weight as well as the choice of tuning parameters are important for all estimators. A conclusion from our simulations is that a particular radius matching estimator combined with regression performs best overall, in particular when robustness to misspecifications of the propensity score and different types of outcome variables is considered an important property.

Keywords: Propensity score matching, kernel matching, inverse probability weighting, selection on observables, empirical Monte Carlo study, finite sample properties.

JEL classification: C21.

Address for correspondence: Martin Huber, Michael Lechner, Conny Wunsch, Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Varnbüelstrasse 14, CH-9000 St. Gallen, Switzerland, Michael.Lechner@unisg.ch, www.sew.unisg.ch/lechner.

---

* Michael Lechner is a Research Fellow of CEPR and PSI, London, CES-Ifo, Munich, IAB, Nuremberg, IZA, Bonn, and ZEW, Mannheim. Conny Wunsch is a Research Fellow of CES-Ifo, Munich, and IZA, Bonn. This project received financial support from the Institut für Arbeitsmarkt und Berufsforschung, IAB, Nuremberg (contract 8104). We would like to thank Patrycja Scioch (IAB), Benjamin Schünemann and Darjusch Tafreschi (both SEW, St. Gallen) for their help in the early stages of data preparation. The paper has been presented at the annual meeting of the German Statistical Society in Dortmund and the Statistische Woche in Nuremberg, as well as at seminars at EIEF, Rome, at the Economics Department of the University of Mannheim and the Center for European Economic Research (ZEW), Mannheim. We thank participants, in particular Markus Frölich and Franco Perrachi, for helpful comments and suggestions. The usual disclaimer applies. Furthermore, the current version of the paper benefit much from very helpful remarks by the editor, John Geweke, an anonymous associate editor, and four anonymous referees. All remaining errors are of course our own.

1 Introduction

Semiparametric estimators using the propensity score to adjust in one way or another for covariate differences are now well-established for either estimating causal effects in a selection-on-observables framework with discrete treatments or for simply purging the means of an outcome variable in two or more subsamples from differences due to observed variables.[1] Compared to (non-saturated) parametric regressions, they have the advantage of including the covariates in a more flexible way without incurring a curse-of-dimensionality problem and of allowing for effect heterogeneity. The former problem, which is highly relevant due to the usually large number of covariates that should be adjusted for, is tackled by collapsing the covariate information into a single parametric function. This function, the so-called propensity score, is defined as the probability of being observed in one of two subsamples conditional on the covariates. The difference to parametric regression is that this parametric function is not directly related to the outcome (as it would be in regression) and thus additional robustness with respect to specification issues can be expected.[2] These methods originate from the pioneering work of Rosenbaum and Rubin (1983) who show that balancing two samples on the propensity score is sufficient to equalize their covariate distributions.

[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] The propensity-score could also be non-parametrically estimated for maximum robustness. In practice, this is however avoided because the dimension of covariates is too large for such an estimator to have desirable properties with the samples usually available for such studies.

Although many of these propensity-score-based methods are not asymptotically efficient (see for example Heckman, Ichimura, and Todd, 1998, and Hahn, 1998),[3] they are the work-horses in the literature on microeconometric programme evaluations and are now rapidly spreading to other fields. They are usually implemented as semiparametric estimators: the propensity score is based on a parametric model, but the relationship between the outcome variables and the propensity score is nonparametric. However, despite the popularity of propensity-score-based methods, the issue of which version of the many different estimators suggested in the literature should be used in a particular type of application is still unresolved, despite recent advances in important Monte Carlo studies by Frölich (2004) and Busso, DiNardo and McCrary (2009a,b). In this paper we address this question and add further insights to it.

Broadly speaking, the popular estimators can be subdivided into four classes: Parametric estimators (like OLS or Probit or their so-called double-robust relatives, see Robins, Mark and Newey, 1992), inverse (selection) probability weighting estimators (similar to Horvitz and Thompson, 1952) or to the recently introduced version by Graham, Pinto and Egel (2011), direct matching estimators (Rubin, 1974, Rosenbaum and Rubin, 1983), and kernel matching estimators (Heckman, Ichimura and Todd, 1998).[4] However, many variants of the estimators exist within each class and several methods combine the principles underlying these main classes.

[3] See the paper by Angrist and Hahn (2004) for an alternative justification of conditioning on the propensity score by using non-standard (panel) asymptotic theory.

[4] There is also the approach of stratifying the data along the values of the propensity score ('blocking'), but this approach did not receive much attention in the empirical economic literature and does not have very attractive theoretical properties. It is thus omitted (see for example Imbens, 2004, for a discussion of this approach).

There are two strands of the literature that are relevant for our research question: First, the literature on the asymptotic properties of a subset of estimators provides some approximate guidance on their small sample properties. Therefore, the next section reviews this literature while discussing the various estimators. Unfortunately, such properties have not (yet?) been derived for all estimators that are used in practice, nor is it obvious how well these asymptotic properties approximate small sample behaviour. Furthermore, these results are usually not informative for the important choice of tuning parameters (e.g., number of matched neighbours, bandwidth selection in kernel matching), on which almost all of these estimators critically depend.

The second strand of the literature provides Monte Carlo evidence on the properties of the estimators of the effects.[5] As one of the first papers investigating estimators from several classes simultaneously, Frölich (2004) found that a particular version of kernel-matching based on local regressions with finite sample adjustments (local ridge regression) performs best. In contrast, Busso, DiNardo and McCrary (2009a, b) conclude that inverse probability weighting (IPW) has the best properties (when using normalized weights for estimation).[6] They explain the differences to the Frölich (2004) study by claiming i) that he considers unrealistic data generating processes and ii) that he does not use an IPW estimator with normalized weights. In other words, they point to the design dependence of the Monte Carlo results as well as to the requirement of having to use optimized variants of the estimators.

[5] There are several papers that are not interested in the properties of the estimator of the effects, but only interested in quality of covariate balancing achieved by the different matching methods. For example, King, Nielson, Coberley, Pope, and Wells (2011) motivate this interest in the fact that matching is not seen as an estimator for an effect, but as a 'pre-processor' that purges the data from differences related to observed covariates. After this preprocessing step, other estimators are used with the matched data to obtain the final result. This view is however not our view of matching nor the view usually entertained by other econometricians.

[6] Further findings from more specific Monte Carlo studies will be discussed below.

Below, we argue that their work may be subject to the same criticism. Indeed, it is this criticism that provides a major motivation for our study.

We contribute to the literature on the properties of estimators based on adjusting for covariate differences in the following way: First of all, we suggest a different approach to conduct simulations. This approach is based on 'real' data. Therefore, we call our particular implementation of this idea an 'Empirical Monte Carlo Study'.[7] The basic idea is to use the real data to simulate realistic 'placebo treatments' among the non-treated. Selection into treatment, which is potentially of key importance for the performance of the various estimators, is based on a selection process directly obtained from the data. The various estimators then use the remaining non-treated in different ways to estimate the (known) non-treatment outcome of the 'placebo-treated' exploiting the actual dependence of the outcome of interest on the covariates on which selection is based in the data. Thus, this approach is much less prone to the standard critique of simulation studies that the chosen data generating processes are irrelevant for real applications. Since our model for the propensity score mirrors specifications used in past applied work, it depends on many more covariates compared to the studies mentioned above. Although this makes the simulation results particularly plausible in our context, which is the context of labour market programme evaluation in Europe, this may also be seen as a limitation concerning its applications to other fields. Therefore, to help generalize the results outside our specific data situation, we further modify many features of the data generating process, like the type of the outcome variable and as well as various aspects of the selection process.

[7] Stigler (1977) is probably the first paper explicitly suggesting a way to do a type of Monte Carlo study with real data (we thank a referee of this journal for this reference). See Section 3.1 for more recent references using the same basic idea of informing the simulations by real data.

Secondly, we consider standard estimators as well as their modified (optimised?) versions based on different tuning parameters such as bandwidth or radius choice. This leads to a great number of estimators to evaluate, but it also provides us with more information on particular important choices regarding the parameters on which the various estimators depend. Such estimators may also consist of combinations of estimators, like combining matching with weighted regression, which have not been considered in any simulation so far.

Finally, we reemphasise the relevance of trimming to improve the finite sample properties of all estimators. The rule we propose is (i) a data driven trimming rule, (ii) easy to implement, (iii) identical for all estimators, and (iv) avoids asymptotic bias. We show that for all estimators considered, including the parametric ones, trimming based on this rule effectively improves their performance.

Overall, we find that (i) trimming observations that have 'too large' a weight is important for all estimators; (ii) the choices of the various tuning parameters play an important role; (iii) simple matching estimators are inefficient and have considerable small sample bias; (iv) no estimator is superior in all designs and for all outcomes; (v) inverse probability tilting performs best for a binary outcome but is unsatisfactory for a semi-continuous outcome; (vi) particular bias-adjusted radius (caliper) matching estimators perform best on average, but may have fat tails if the number of controls is not large enough; and finally, (vii) flexible, but simple parametric approaches do almost as well in the smaller samples, because their gain in precision frequently compensates (in part) for their larger bias which, however, dominates when samples become larger. Strictly speaking these properties relate to our particular data generating process (DGP) only. However, at least such a DGP is typical for an important application of matching methods, namely labour market evaluations.

The plan of the paper is as follows: In the next section we discuss the basic setup of each of the relevant estimators and their properties, as well as the issue of trimming, while

relegating the technical details of the estimators to Appendix A. Section 3 describes our Monte Carlo design, again relegating many details as well as descriptive statistics to Appendix B as well as to Appendix C, which contains a description of the support features of our data. The main results are presented in Section 4, while the full set of results is given in Appendix D. Section 5 concludes and Appendix E contains further sensitivity checks. The website of this paper (www.sew.unisg.ch/lechner/matching) will contain additional material that has been removed from the paper for the sake of brevity, in particular Appendices B, C, D, and E as well as the Gauss, Stata, and R codes for the preferred estimators.

2 Estimators

2.1 Notation and targets for the estimation

The outcome variable, *Y*, denotes earnings or employment. The group of treated units (treatment indicator *D=1*) are the participants in training in our empirical example. We are interested in comparing the mean value of *Y* in the group of treated (*D=1*) with the mean value of *Y* in the group of non-treated (*D=0*), the non-participants, free of any mean differences in outcomes that are due to differences in the observed covariates *X* across the groups.[8]

$$\begin{aligned}\theta &= E(Y \mid D=1) - E[E(Y \mid X, D=0) \mid D=1] \\ &= E(Y \mid D=1) - \int_{\chi} E[Y \mid D=0, X=x] f_{X|D=1}(x)dx \\ &= E(Y \mid D=1) - \int_{0}^{1} E[Y \mid D=0, p(X)=\rho] f_{p(x)|D=1}(\rho)d\rho,\end{aligned}$$

[8] As a convention, capital letters denote random variables, while small letters denote particular realisations of the random variables. If the small letters are indexed by another small letter, typically *i* or *j*, it means that this is the value realised for the sample unit *i* or *j*.

where $f_{X|D=1}$ denotes the conditional density of X and χ its support. The propensity score is defined by $P(D=1|X=x) =: p(x)$. The second equality is shown in the seminal paper by Rosenbaum and Rubin (1983).

If there are no other (perhaps unobserved) covariates that influence the choice of the different values of *D* as well as the outcomes that would be realised for a particular value of *D* (the so-called potential outcomes), this comparison of means yields a causal effect, namely the average treatment effect on the treated (ATET). This is the mean effect of *D* on individuals observed with *D=1*.[9] The assumption required to interpret θ as a causal parameter is called either unconfoundedness, the conditional independence assumption (CIA) or selection on observed variables (e.g., Imbens, 2004). The plausibility of the CIA depends on the particular empirical problem considered and on the richness of the data at hand. That is, in labour market applications estimating the effects of training programmes on employment, X should include variables reflecting education, individual labour market history, age, family status, and local labour market conditions, among others, in order to plausibly justify the CIA (e.g. Gerfin and Lechner, 2002). Therefore, in applications exploiting the CIA, *X* is typically of high dimension, as in most cases many covariates are necessary to make this assumption plausible. However, for this paper, which focuses on the finite sample properties of estimation, it does not matter whether θ has a causal interpretation or not. It is important to note that other semiparametric estimators also rely on propensity-score-based covariate adjust-

[9] For reasons of computational costs (which are a severe restriction in our analysis due to the complexity of the design and the numbers of estimators) we focus entirely on reweighting the controls towards the distribution of *X* among the treated. Common alternatives are reweighting the treated towards the covariate distribution of the controls, or weighting the outcomes of both groups towards the covariate distribution of the population at large. The resulting parameters are called the average treatment effect on the non-treated (ATENT) and the average treatment effect (ATE). Estimating the 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. We conjecture that having a good estimate of the components of the ATE will lead to a good estimate of the ATE.

ments, like, for example, the instrumental variable estimator proposed by Frölich (2007a), the decomposition-type of approach suggested by DiNardo, Fortin and Lemieux (1996) and semi-parametric versions of the difference-in-difference estimator (e.g., Abadie, 2005, Blundell, Meghir, Costas Dias, and van Reenen, 2004, Lechner, 2010).

2.2 General structure of the estimators considered

As discussed by Smith and Todd (2005), Busso, DiNardo, and McCrary (2009a) and Angrist and Pischke (2009) among many others, all estimators adjusting for covariates can be understood as different methods that weight the observed outcomes using weights, $\hat{w}_i$.

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

where N denotes the sample size of an i.i.d. sample and N_1 is the size of the treated subsample. Reweighting is required to make the non-treated comparable to the treated in terms of the propensity score. See for example the afore-mentioned references for formulas of the weighting functions implied by various estimators. In almost all cases we will set $\hat{w}_i = 1$ for the treated, i.e. we estimate the mean outcome under treatment for the treated by the sample mean of the outcomes in the treated subsample. Therefore, the different estimators discussed below represent different ways to estimate $E[E(Y \mid X, D=0) \mid D=1]$. Following Busso, DiNardo, and McCrary (2009a), we normalize the weights of all semi-parametric estimators such that $\frac{1}{N_0}\sum_{i=1}^{N}(1-d_i)\hat{w}_i = 1$.

Next, we will briefly introduce the estimators considered in this study, namely inverse probability weighting, direct matching, kernel matching, linear and non-linear regressions as well as combinations of direct matching and inverse probability weighting with regression.

All of these estimators, or at least similar versions of them, have been applied in empirical studies,[10] which is the motivation for analysing them in this paper.

2.3 Inverse probability weighting

As already mentioned, the idea of inverse-probability-of-selection weighting (henceforth abbreviated as IPW) goes back to Horvitz and Thompson (1952). IPW can attain the semi-parametric efficiency bound derived by Hahn (1998) when using a non-parametric estimate of the propensity score. It is generally not efficient when based on the true or a parametrically estimated propensity score (see Hirano, Imbens and Ridder, 2003, for results and an excellent summary of the literature on the efficiency of IPW).[11]

Several IPW estimators for the ATET have recently been analysed by Busso, DiNardo and McCrary (2009a, b). In this Monte Carlo study we consider the following implementation:

$$\hat{\theta}_{IPW} = \frac{1}{N_1}\sum_{i=1}^{N} d_i y_i - \sum_{i=1}^{N} \frac{(1-d_i)\dfrac{\hat{p}(x_i)}{1-\hat{p}(x_i)}}{\displaystyle\sum_{j=1}^{N}\frac{(1-d_j)\cdot\hat{p}(x_j)}{1-\hat{p}(x_j)}} y_i \,.$$

The normalization $\displaystyle\sum_{j=1}^{N}\frac{(1-d_j)\cdot\hat{p}(x_j)}{1-\hat{p}(x_j)}$ ensures that the weights add up to one in the sample as well as in expectation. This estimator directly reweights the non-treated outcomes to control for differences in the propensity scores between treated and non-treated observa-

[10] For inverse probability weighting see DiNardo, Fortin, and Lemieux (1996), for one-to-one matching Rosenbaum and Rubin (1983), for kernel matching see Heckman, Ichimura, and Todd (1998), for caliper matching see Dehejia and Wahba (1999), and for double-robust estimation see Robins, Mark, and Newey (1992). Of course, many more studies than those mentioned as (early) examples use these estimators in various applications.

[11] Hirano, Imbens, and Ridder (2003) prove that the efficiency bound is reached when the propensity score is estimated non-parametrically by a particular series estimator. The results by Newey (1984) on two-step GMM estimators imply that IPW estimators based on a parametric propensity score are consistent and asymptotically normally distributed (under standard regularity conditions).

tions. It is the estimator recommended by Busso, DiNardo and McCrary (2009a). When a parametric propensity score is used inference for IPW is straightforward, because one could either rely, for example, on the GMM methodology (Hansen, 1982), or on the bootstrap.

This estimator is attractive because it is computationally easy and fast, it is probably close to being asymptotically efficient, and (in principle) there is no need to choose any tuning parameters. However, there is also evidence that this or related IPW estimators may be sensitive to large values of $\hat{p}(x)$ that might lead to fat tails in its distribution (see, for example, Frölich, 2004, as well as the discussion in Busso, DiNardo and McCrary, 2009b). Furthermore, as this estimator exploits the propensity score directly, there is a potential concern that it might be more sensitive to small misspecifications of the propensity score than other estimators that do not exploit the actual value of the propensity score, but compare treated and controls with same value of the score, whatever that value is (e.g., Huber, 2011).

A method closely related to IPW is inverse probability tilting (IPT), proposed by Graham, Pinto and Egel (2011). In contrast to conventional IPW, where propensity scores typically are estimated by maximum likelihood methods (for a correct parametric model maximum likelihood is efficient for the parameters of the propensity score but not necessarily for the ATET estimated by using these scores to form the IPW weights), the propensity score is estimated by a particular method of moments estimator. The latter determines the coefficients in the propensity score model such that the moments of the covariates in the treated sample reweighted by the propensity score exactly coincide with the unweighted moments in the full sample. In other words, the estimator satisfies the following moment conditions:

$$\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}\dfrac{d_i}{\hat{p}(x_i)}\\ \dfrac{d_i x_i'}{\hat{p}(x_i)}\end{pmatrix}=\begin{pmatrix}1\\ \dfrac{1}{N}\sum_{i=1}^{N}x_i'\end{pmatrix}.$$

The resulting exact balancing property of the propensity score is the key difference to standard IPW. Therefore, further refinements are not required, such that the propensity scores are plugged into the non-normalized IPW estimator in order to estimate the ATET. Graham, Pinto and Egel (2011) argue that IPT is attractive in terms of efficiency and robustness when compared to other forms of IPW.

2.4 Direct matching

Pair, one-to-one, or single nearest neighbour matching is considered to be the prototype of a matching estimator (with replacement)[12]. The pair matching estimator (PM) is defined as:

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

$\underline{1}(\cdot)$ denotes the indicator function, which is one if its argument is true and zero otherwise. This estimator is not efficient, as only one non-treated observation is matched to each treated observation, independent of the sample size. All other control observations obtain a weight of zero even if they are very similar to the observations with positive weight.

Despite its inefficiency, PM also has its merits. Firstly, using only the closest neighbour should reduce bias (at the expense of additional variance). Secondly, PM is likely to be more robust to propensity score misspecification than IPW as it remains consistent if the misspecified propensity score model is a monotone transformation of the true model (see the

[12] 'With replacement' means that a control observation can be used many times as a match, whereas in estimators 'without replacement' it is used at most once. Since the latter principle works only when there are many more controls than treated, it is rarely used in econometrics and will be omitted from this study in which we consider treatment shares of up to 90%. For matching without replacement, many more matching algorithms have appeared in the literature that differ on how to use the scarce pool of good controls optimally (as they can only be used once). See, for example, Augurzky and Kluve (2007) and Hansen (2004) for some discussion of these issues.

simulation results in Drake, 1993, Zhao, 2004, 2008, Millimet and Tchernis, 2009, and Huber, 2011, suggesting some robustness to the specification of the propensity score).

A direct extension of PM is the *1:M* propensity score matching estimator which, instead of using just one control, uses several controls. Thus, increasing *M* increases the precision but also the bias of the estimator. This class of estimators has been analysed by Abadie and Imbens (2009) for the ATE and has been found to be consistent and asymptotically normal for a given value of *M*. Yet, it appears that there are no results on how to optimally choose *M* in a data dependent way. Thus, we focus on 1:1 matching, which is the most frequently used variant in this class of estimators.

The third class of direct matching estimators considered is the one-to-many caliper matching algorithm as, for example, discussed by Rosenbaum and Rubin (1985) and used by Dehejia and Wahba (1999, 2002). Caliper or radius matching uses all comparison observations within a predefined distance around the propensity score of each treated unit. This allows for higher precision than fixed nearest neighbour matching in regions of the χ-space in which many similar comparison observations are available. Also, it may lead to a smaller bias in regions where similar controls are sparse. In other words, instead of fixing *M* globally, *M* is determined in the local neighbourhood of each treated observation.

There are further matching estimators evaluated in the literature. For example, Rubin (1979) suggested combining PM with (parametric) regression adjustment to take into account the fact that treated and controls with exactly the same propensity score are usually very rare or non-existent.[13] This idea has been taken up again by Abadie and Imbens (2006) who show that for a *1:M* matching estimator (directly on *X*) nonparametric regression can be used to

[13] This idea has been applied by Lechner (1999, 2000) in a programme evaluation study.

remove the bias from the asymptotic distribution that may occur when X is more than one-dimensional.

An additional suggestion for improving naïve propensity score matching estimators is to use a distance metric that not only includes the propensity score, but in addition those covariates that are particularly good predictors of the outcome (in addition to the treatment). Since this distance metric has many components, usually a Mahalanobis distance is used to compute the distance between the treated and the controls (again, see the discussion in Rosenbaum and Rubin, 1985).

The estimator proposed by Lechner, Miquel and Wunsch (2011) and used in several applications by these authors,[14] combines the features of caliper matching with additional predictors and linear or nonlinear regression adjustment. After the first step of distance-weighted caliper matching with predictors (which could be reinterpreted as kernel matching, see below, with a truncated triangle kernel[15]), this estimator uses the weights obtained from matching in a weighted linear or non-linear regression in order to remove any bias due to mis-matches. The matching protocol of this estimator is shown in Appendix A.

Inference for the matching estimators is usually performed by bootstrap, although the results in Abadie and Imbens (2008) suggest at least for (pure) nearest neighbour matching, the bootstrap may not be valid. For the case of using a parametric propensity score as in this paper, Abadie and Imbens (2009) suggest alternative procedures.

[14] See Wunsch and Lechner (2008), Lechner (2009), Lechner and Wunsch (2009a, b), Behncke, Frölich and Lechner (2010a, b), and Huber, Lechner and Wunsch (2011).

[15] We thank Jeff Smith for pointing out this relation between kernel and radius matching.

2.5 Kernel matching

Propensity score kernel matching is based on the idea of consistently estimating the regression function $E[Y \mid D=0, \hat{p}(X)=\rho] =: m(\rho)$ with the control observations and then averaging the estimated function by the empirical distribution of $\hat{p}(X)$ for the observed treated:

$$\hat{\theta}_{kernel} = \frac{1}{N_1} \sum_{i=1}^{N} d_i \left[y_i - \hat{m}(\hat{p}(x_i)) \right],$$

where $\hat{m}(\cdot)$ denotes the nonparametrically estimated conditional expectation function. Heckman, Ichimura and Todd (1998) is an early example of an analysis of the type of kernel regression estimators that could achieve Hahn's (1998) semiparametric efficiency bound if the covariates were used directly instead of the propensity score (see also Imbens, Newey and Ridder, 2006). Due to the curse-of-dimensionality problem, however, the latter has very undesirable small sample properties in a typical application.

Considering a continuous outcome, Frölich (2004) investigated several kernel matching estimators and found the estimator that is based on ridge regressions to have the best finite sample properties. Ridge regression may be considered an extension to local linear kernel regression. The latter is superior to the local constant kernel estimator in terms of boundary bias (which is the same as in the interior, see Fan, 1992), but is prone to variance problems entailing rugged regression curves when data are sparse or clustered (see Seifert and Gasser, 1996). Therefore, a ridge term is added to the estimator's denominator to avoid division by values close to zero (thus in effect dealing with ‘extreme’ observations in a particular way). The details of the estimator used in the simulation (including the choice of the bandwidth) can be found in Appendix A.2. As we also consider a binary outcome variable (see Section 3.2.3), we apply (in addition to ridge regression) kernel matching based on local logit regression as used in Frölich (2007b). Note that the latter does not include a ridge term, which is not neces-

sary because of the finite support of the expectation of the outcome variable (even under very large coefficients) due to the logit link function.

Inference for kernel regression is usually performed by bootstrap methods.

2.6 Parametric models

The parametric estimators used here are similar to kernel matching estimators with two exceptions. The first difference is that we use a parametric specification for the conditional expectation function, $m(\cdot)$, such as a probit or linear model. The second difference is that instead of using the propensity score as regressor, we use the covariates that enter the propensity score directly in a linear index specification, as it is done in typical applications.[16] This approach may be regarded as unusually flexible (given how regressions are used in many applications) in that estimation only takes place in the non-treated subsample.[17] However, specifying a joint model for the treated and controls that just includes a treatment dummy is unnecessarily restrictive. I.e., it can lead to large biases (because it essentially estimates a treatment effect for a different population[18]) and, thus, is not competitive with the more flexible semiparametric models consider in this paper.

We also combine IPW with parametric linear and non-linear regression, an approach that has been termed "double-robust regression" (DR) in the (epidemiology) literature. DR estimation follows in two steps. First, we run weighted regressions of Y on X in the pool of

[16] Using the propensity score as regressor is less attractive in a parametric setting compared to kernel matching, because in parametric regressions functional forms play a crucial role, and the propensity score is obviously not an attractive choice because it does not relate the variation of the covariates directly to the variation of the outcome variables. Furthermore, the curse of dimensionality problem is less relevant in parametric regressions.

[17] We also examined a specification that uses a regression model for the treated, too. However, as the results are almost identical, we do not consider this case explicitly.

[18] For the linear model, details can be found in Angrist (1998) and Angrist and Pischke (2009).

non-treated individuals, where the weights are proportional to $\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}$. This reweights the controls according to the distribution of $\hat{p}(X)$ among the treated. Second, let $\hat{g}(x, \hat{p}(x))$ denote the weighted predicted outcome under non-treatment, which is the estimate of $g(x, p(x)) := q(x)\frac{p(x)}{1-p(x)}$, $(q(x) := E(Y \mid X = x, D = 0))$. After an appropriate specification of the model for the conditional expectation of the outcome, $q(x)$, the DR estimator of the ATET is given by:

$$\hat{\theta}_{DR} = \frac{1}{N_1}\sum_{i=1}^{N} d_i \left[y_i - \hat{g}(x_i, \hat{p}(x_i)) \right].$$

This estimator possesses the double robustness property as it remains consistent if either the model for the propensity score or the regression model, or both, are correctly specified. However, the estimator is not necessarily efficient if misspecification appears in one of the models.[19]

The parametric estimators have the advantage that they are very easy to compute, their asymptotic properties are well known, inference procedures are usually known and reliable, they are efficient if correctly specified, they do not depend on tuning parameters, and they may be used to impute counterfactuals outside the common support, i.e. they can be used even if the propensity cannot be estimated (which may happen if one variable is a perfect predictor of the treatment status). Clearly, the disadvantage is their sensitivity to the correct specification of the models involved.

[19] Robins, Rotnitzky and Zhao (1994) and Robins and Rotnitzky (1995) show that DR is semi-parametrically efficient if both model components are correctly specified (see also the discussions in Robins, Mark, and Newey, 1992, Scharfstein, Rotnitzky, and Robins, 1999, Hirano and Imbens, 2001, Lunceford and Davidian, 2004, Bang and Robins, 2005, and Wooldridge, 2007, as well as the introduction into these methods by Glynn and Quinn, 2010). Concerning the robustness of regressions, the paper by Kline (2011) shows that linear regressions also have DR properties related to a propensity score that is linear in the covariates.

3 The simulation design

3.1 Basic idea

A typical Monte Carlo study specifies the data generation process of all relevant random variables and then conducts estimation and inference from samples that are generated by independent draws from those random variables based on pseudo random number generators. The advantage of such a design is that all dimensions of the true data generating process (DGP) are known and can be used for a thorough comparison with the estimates obtained from the simulations. However, the disadvantage is that the DGPs are usually not closely linked to real applications in terms of the number and types of variables used for covariates and outcomes. Furthermore, the outcome and selection processes are also quite arbitrary (irrespective of the fact that the respective papers usually claim that their design reflects the key features of the applications they have in mind).[20] Given that the results in the literature mentioned above suggest that the small sample behaviour of some of the estimators is design dependent, we propose an alternative method that we call an *Empirical Monte Carlo Study* (EMCS).

The idea of an EMCS is to base the DGP not entirely on relations specified by the researcher, but to exploit real data instead as much as possible, e.g. to use observed outcomes and covariates instead of simulated ones as well as an observed selection process. Of course, this approach has its limits as the researcher still requires the ability to control some key parameters, such as, for example, the share of the treated or the sample size, to allow for some generalizations. Furthermore, the data must be very large to be able to treat the sample as

[20] All Monte Carlo studies mentioned here suffer from this problem. They are also more restrictive on many other, usually computationally expensive dimensions, like the types of estimators, the sample sizes, and the number of covariates considered.

coming from an infinite population and it has to be relevant for the estimators under investigation. That is, it should be a typical data set for countries whose governments provide rich individual data for evaluation purposes, like for example the German-speaking and the Nordic countries. Of course, it is exactly in this case of informative individual data when matching becomes an attractive evaluation method.

Since the estimators we consider are heavily used for the evaluation of active labour market programmes for the unemployed based on (typically European) administrative data, we choose a large German administrative data set as our population. Our EMCS basically consists of three steps: First, we estimate the propensity score in the 'population' and use it as the true propensity score for the simulations. Second, we draw a sample of control observations, simulate a (placebo-) treatment for this draw, and estimate the effects with the different estimators for this sample. By definition, the true effect of this treatment is zero. Third, we repeat the second step many times to evaluate the performance of the estimators.

In other contexts related ideas appeared in the literature. For example Bertrand, Duflo, and Mullainathan (2004) use so-called placebo-laws (i.e. artificial law changes that never happened in the real world) to investigate inference procedures for difference-in-difference estimators. Abadie and Imbens (2002) and Diamond and Sekhon (2008) use a data generating process that tries to closely mimic the LaLonde (1986) National Supported Work (NSW) data to investigate the performance of a new class of matching estimators. Lee and Whang (2009) draw samples from the NSW data to study the performance of tests for zero treatment effects. Finally, Khwaja, Salm, and Trogdon (2010) use simulated data coming from a structural model to evaluate the performance of treatment effect estimators.

Our EMCS approach is also closely related to the literature that examines the properties of estimators based on how capable they are of reproducing the results of an experimental control group, see for example LaLonde (1986), Heckman, Ichimura, Smith, and Todd

(1998), Dehejia and Wahba (1999, 2002), Smith and Todd (2005), Dehejia (2005), Zhao (2006), Flores and Mitnik (2009), and Jacob, Ludwig and Smith (2009). There are at least two important advantages of the EMCS compared to this approach if used for comparing estimators based on the same identifying assumptions. Firstly, because EMCS repeatedly draws subsamples from the population, it allows the distribution of the estimators to be fully recovered.[21] Secondly, probably the most important advantage of using EMCS is that it allows varying many parameters of the DGP, in particular the selection process and the sample size. Of course, if *large* experimental data is available, it could also fruitfully be used to implement an EMCS-type approach.

3.2 The population

In the next subsections we present the details of how the EMCS is implemented. We begin by describing the properties of the 'population' on which all our simulations are based.

3.2.1 Data

The data comprise a 2% random sample drawn of all German employees subject to social insurance.[22] They cover the period 1990-2006 and combine information from different administrative sources: (1) records provided by employers to the social insurance agency for each employee (1990-2006), (2) unemployment insurance records (1990-2006), (3) the programme participation register of the Public Employment Service (PES, 2000-2006) as well as (4) the jobseeker register of the PES (2000-2006). Finally, a variety of regional variables has been matched to the data using the official codes of the 439 German districts. These include information about migration and commuting, average earnings, unemployment rate, long-term

[21] When comparing an observational and an experimental control group for the US JTPA programme, Plesca and Smith (2007) obtain the distribution of their estimators in a related manner by bootstrap methods.

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

unemployment, welfare dependency rates, urbanisation codes, and measures of industry structure and public transport facilities.

For each individual the data comprise all aspects of their employment, earnings and UI history since 1990 including the beginning and end date of each spell, type of employment (full/part-time, high/low-skilled), occupation, earnings, type and amount of UI benefit, and remaining UI claim. Moreover, they cover all spells of participation in the major German labour market programmes from 2000 onwards with the exact start date, end date and type of programme as well as the planned end date for the training programmes. The jobseeker register contains a wealth of individual characteristics, including date of birth, gender, educational attainment, and marital status, number of children, age of youngest child, nationality, occupation, the presence of health impairments and disability status. With respect to job search the data contain the type of job looked for (full/part-time, high/low-skilled, occupation), whether the jobseeker is fully mobile within Germany and whether she has health impairments that affect employability.

This data was the basis of several evaluation studies thus far[23] and is fairly typical for the administrative data bases that are available in several European countries to evaluate the effects of active labour market policies.

3.2.2 Sample selection and treatment definition

As we are interested in evaluating typical labour market programmes in a representative industrialized economy we exclude East Germany and Berlin from the analysis since they are still affected by the aftermath of reunification. We start with a sample that covers all entries into unemployment in the period 2000-2003. Then, we exclude unemployment entries in

[23] See Hujer, Caliendo and Thomsen (2004), Hujer, Thomsen and Zeiss (2006), Caliendo, Hujer and Thomsen (2006, 2008a, b), Wunsch and Lechner (2008), Lechner and Wunsch (2009a), and Hujer and Thomsen (2010).

January-March 2000 because with programme information starting only in January 2000 we want to make sure that we do not accidentally classify entries from employment programmes (which we would consider as unemployed) as entries from unsubsidized employment because the accompanying programme spell is missing. Entries after 2003 are not considered to ensure that we have at least three years after starting unemployment to observe the outcomes.

We further restrict the analysis to the prime-age population aged 20-59 in order to limit the impact of schooling and (early) retirement decisions. To make our sample more homogeneous we also require that individuals were not unemployed or in any type of labour market programme (including subsidized employment) in the last 12 months before becoming unemployed. Finally, we exclude the very few cases whose last employment was any non-standard form of employment such as internships.

As in Lechner, Miquel and Wunsch (2011) and Lechner and Wunsch (2009b) we define participants (treated) as all of those individuals in our sample who start training courses that provide job-related vocational classroom training [24] within the first 12 months of unemployment. The non-treated are those who did not participate in any programme of the active labour market policy whatsoever in the same period. There are 3'266 treated and 114'349 controls.

3.2.3 Descriptive statistics

The upper part of Table 3.1 presents descriptive statistics for the two outcome variables we considered: average monthly earnings over the 3 years after entering unemployment, and an indicator whether there has been some (unsubsidized) employment in that period. This choice has been made to evaluate the estimators' performance with both a variable with only

[24] The programs we consider correspond to *general training* in Wunsch and Lechner (2008) and to *short and long training* in Lechner, Miquel and Wunsch (2011).

two support points and a semi-continuous variable (50% zeros). Furthermore, the table contains the descriptive statistics for the 38 confounders that are taken into consideration in the selection equation. Among those are also eight interaction terms, which will be used later on to judge the robustness of the estimators with respect to functional misspecification of the propensity score.

To describe selectivity, Table 3.1 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 that there is a substantial amount of selectivity that is, however, not captured by a single variable, but by several variables. This view is also confirmed by considering the last two lines of this table which display the normalized differences for the estimated propensity score as well as its linear index. Not surprisingly, those summary measures show much higher selectivity than the single variables, despite the low pseudo-R^2 of about 4%, which is, however, in the range common to such studies. [25]

[25] Table B.1 in Appendix B.1 shows the results of a probit and linear regression using, respectively, employment and earnings as dependent variables and the covariates as independent variables to confirm that the latter do not only determine selection but are also related to the outcomes in a way such that confounding takes place.

Table 3.1: Descriptive statistics of the 'population'

Variable	Treated		Control		Standardized difference	Probit est. of selection equat.	
	mean	std.	Mean	std.	in %	Marg.eff. in %	std. error
3 years since beginning of UE spell	.63	0.56	.48	0.50	9	-	-
some unsubsidized employ.							
av. monthly earnings (EUR)	1193	1041	1115	1152	9	-	-
Constant term	-	-	-	-	-	-	-
Age / 10	3.6	3.5	0.84	1.1	8	7.3	0.5
… squared / 1000	1.4	1.4	0.63	0.85	3	-9.1	0.6
20 - 25 years old	0.21	binary	0.41		22	0.9	0.2
Women	0.57	.46	0.50	0.50	15	-5.8	1.5
Not German	0.11	binary	0.31		16	-0.5	0.1
Secondary degree	0.32	binary	0.47		15	1.1	0.1
University entrance qualification	0.29	binary	0.45		15	1.0	0.1
No vocational degree	0.18	binary	0.39		26	-0.3	0.1
At least one child in household	0.42	binary	0.49		22	-0.2	0.1
Last occupation: Non-skilled worker	0.14	binary	0.35		13	0.3	0.1
Last occupation: Salaried worker	0.40	binary	0.49		29	1.8	0.2
Last occupation: Part time	0.22	binary	0.42		12	2.1	0.3
UI benefits: 0	0.33	binary	0.47		16	-0.6	0.1
> 650 EUR per month	0.26	binary	0.44		7	0.7	0.1
Last 10 years before UE: share empl.	0.49	0.46	0.34	0.35	8	-1.4	0.2
share unemployed	0.06	0.05	0.11	0.11	1	-2.5	0.5
share in programme	0.01	0.01	0.04	0.03	9	5.1	1.2
Last year before UE: share marg. em.*	0.07	0.03	0.23	0.14	15	-1.0	0.7
share part time	0.16	0.11	0.33	0.29	10	-1.0	0.2
share out-of-the labour force (OLF)	0.28	0.37	0.40	0.44	14	-1.3	0.2
Entering UE in 2000	0.26	binary	0.44		13	1.6	0.2
2001	0.29	binary	0.46		5	0.9	0.1
2003	0.20	binary	0.40		12	0.0	0.1
Share of pop. living in/ close to big city	0.76	0.73	0.35	0.37	6	0.4	0.1
Health restrictions	0.09	binary	0.29		13	-0.6	0.1
Never out of labour force	0.14	binary	0.34		6	0.6	0.2
Part time in last 10 years	0.35	binary	0.48		9	-0.5	0.1
Never employed	0.11	binary	0.31		17	-1.0	0.1
Duration of last employment > 1 year	0.41	binary	0.49		4	-0.6	0.1
Average earnings last 10 years when employed / 1000	0.59	0.52	0.41	0.40	13	-0.4	0.2
Women x age / 10	2.1	1.7	1.9	1.9	17	2.6	0.6
x squared / 1000	0.83	0.64	0.85	0.90	15	-2.6	0.8
x no vocational degree	0.09	binary	0.28		15	-0.9	0.1
x at least one child in household	0.32	binary	0.47		25	0.9	0.2
x share minor employment last year	0.06	0.02	0.22	0.13	16	3.2	0.7
x share OLF last year	0.19	0.18	0.36	0.35	3	1.0	0.2
x average earnings last 10 y. if empl.	0.26	0.19	0.34	0.30	16	-1.0	0.2
x entering UE in 2003	0.10	binary	0.30		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^2 in %	3266		114349			3.6	

Note: * Marg(inal) em(ployment) is 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 $\left(1-\sum_{i=1}^{N}[d_i-\hat{p}(x_i)]\Big/\sum_{i=1}^{N}\left[d_i-\sum_{i=1}^{N}(d_i)/N\right]\right)$. 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 effect based on discrete changes for binary variables and derivatives otherwise.

3.3 The simulations in detail

After having estimated the propensity in the full population (see Table 3.1), the treated are discarded and no longer play a role in the following simulations.[26] The next step is to draw the individual random sample of size *N* from the population of non-treated (independent draws with replacement). For the sample sizes we choose 300, 1'200, and 4'800. The motivation for the smallest sample size is that semiparametric methods are not expected to perform well (and are rarely used in applications) for small samples.[27] The choice of the largest sample size on the other hand is heavily influenced by the computational burden it creates, because several of the estimators used are computationally expensive.[28] Furthermore, the largest sample should be small compared to our population of 114'349 controls. If an estimator does not perform well with this comparatively large sample (much larger than in other Monte Carlo studies), a researcher planning to use this estimator might be worried anyway even if a larger sample would be available (as is the case in several recent labour market evaluations). One the other hand, if an estimator performs well for this sample size, i.e. is close to its asymptotic distribution, we expect it to perform similarly or even better for larger sample sizes. As all estimators are $\sqrt{N}-$convergent, increasing sample sizes by a factor of four should reduce the standard error by 50% (in large samples). Thus, this choice facilitates checking whether the estimators already attain this asymptotic convergence rate in finite samples.

[26] The GAUSS code used to generate the simulated data is available from the authors on request. The pseudo random number generator used in all simulations is the one implemented in Gauss 9.0.

[27] Note that the simulations in Busso, DiNardo, and McCrary (2009a, b) are based on sample sizes of 100 and 500, which is much more convenient with respect to computational burden. However, with the number of covariates usually found in applications using matching estimators, it is very difficult if not impossible to estimate the propensity score with 100 observations with some precision.

[28] Computation for one specification with the large sample size can take up to 3 weeks on a standard PC of 2010 vintage.

Having drawn the sample, the next step consists of simulating treated observations in this sample. We base this simulation step on the propensity score that has been estimated in the population and can be computed for each individual as $\hat{p}_i(x_i) = \Phi(x_i\hat{\beta})$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, x_i is the observed covariate value of observation *i* (including the constant), and $\hat{\beta}$ are the estimated parameters. Our baseline specification is (almost) based on using $\hat{p}_i(x_i)$ for the simulation of the treatment.

However, there are at least two dimensions we want to influence because of their important heterogeneity in applications. First of all, the shares of treated observations are 10%, 50%, and 90%. The smallest share is much smaller than those usually found in Monte Carlo studies, but is chosen because small shares of treated frequently occur in applications.[29] The largest share, on the other hand, mimics the situation when the role of treated and controls is reversed as in the estimation of the average treatment effect on the non-treated. The second dimension that varies considerably among applications and may also have a great impact on the relative performance of the estimators is the magnitude of the selection, for example measured in terms of the pseudo-R^2 of the propensity score or its normalized difference (see Table 3.2). We consider (i) the benchmark case of random assignment, (ii) selection that corresponds roughly to the one in our 'population' and (iii) a case of very strong selection.

The resulting scenarios are implemented based on the following equation:

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

[29] Even our smallest share used in the simulations is larger than the share of treated observed in our population, which is just 3%. However, using 3% instead of 10% would have required a further increase in sample sizes and would have put too much additional demand on computation time.

where u_i denotes a standard normally distributed i.i.d. random number, λ is a parameter with three different values that determine the magnitude of selection, and the parameter α is chosen such that the expected number of treated equals 10%, 50%, or 90%, respectively.[30] Table 3.2 summarizes the 21 scenarios that are used in the EMCS and also gives summary statistics about the amount of selection implied by each scenario.[31]

Note that this simulation routine always ensures common support, at least in expectation. Table C.1 and Figures C.1 to C.18 (internet Appendix C) document the overlap in the 'population'. One result that follows from this table and the figures is that in particular 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.

Table 3.2: 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

Note: See note of Table 3.1.

In addition, note that it is not possible to combine the small sample size with the extreme shares of participants. This would frequently include cases in which the number of

[30] Note that the simulations are not conditional on *D*. Thus, the share of treated is random.

[31] 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 treated differs from the original share leads to different values of those statistics even in the case that mimics selection in the original population.

covariates exceeds the number of treated or non-treated observations, thus, posing numerical problems on the estimation of the propensity score. Hence, in the small sample the unconditional treatment probability is 0.5, which also makes small sample issues concerning the common support unproblematic.

Since the true effect is always zero, one might worry that our results are specific to the case of effect homogeneity which would be of less practical relevance. This is, however, not the case as we estimate the average treatment on the treated (ATET). The ATET has two components: the expected potential outcome of the treated under treatment and under no treatment. Here, the former is always estimated in the same way, namely as the mean observed outcome of the treated.[32] The non-treatment potential outcome of the treated is imputed by the different estimators we consider using the outcome from the non-treated only. Without loss of generality, we can model the potential treatment outcome as the sum of the potential non-treatment outcome and a possible heterogeneous effect that may or may not depend on the covariates. However, as this effect only concerns the observed outcomes of the treated, it will asymptotically not affect our relative comparison of estimators (see Tables D.1 to D.3 in internet Appendix D: even in finite samples the best estimator for a given DGP is in most cases the same with and without effect heterogeneity). The only exception is that the trimming rule used may be sensitive to this homogeneity assumption. However, finding the most effective trimming rule is beyond the scope of the paper.

Another parameter of the EMCS, as in any Monte Carlo study, is the number of replications. Ideally, one would choose a number as large as possible to minimize simulation noise. Simulation noise depends negatively on the number of replications and positively on

[32] For some parametric models, different regressions were run in both subsamples of treated and non-treated (instead of using the mean of the treated). However, they are almost identical to the version of taking means for the treated and running the regression for the non-treated only.

the variance of the estimators. Since the latter is doubled when the sample size is reduced by half, and since simulation noise is doubled when the number of replications is reduced by half (at least for averages over the i.i.d. simulations), we chose to make the number of replications proportional to the sample size. For the smallest sample, we use 16'000 replications, for the medium sample 4'000, and for the largest sample 1'000, as the latter is computationally most expensive and has the least variability of the results across different simulation samples.

4 Trimming

From equation (1) we see that all estimators can be written as the mean outcome of the treated minus the weighted outcome of the non-treated observations. By the nature of this estimation principle, the weights of the non-treated are not uniform (except in the case of random assignment in which they should be very similar even in the smallest sample). They depend on the covariates via the propensity score. If particular values of $p(x)$ are rare among the controls and common among the treated, such control observations receive a very large weight in all estimators of the ATET. Consider the extreme case that all treated observations have a value of $p(x) = 0.99$. However, there is only one non-treated observation with such a value (and no other 'similar' non-treated observations). For most of the semiparametric estimators this observation will receive a weight of one (or very close to one) and the remaining non-treated observations a weight of (almost) zero. Thus, such estimators have an infinite variance because they are based on the mean of effectively only one observation. As the sample grows, by the definition of the propensity score, there will be more non-treated observations with $p(x) = 0.99$ (on average, for every 99 additional treated with $p(x) = 0.99$, there will be one additional control with $p(x) = 0.99$) and the problem becomes less severe.

This suggests that the properties of the estimators deteriorate when single observations obtain 'too' large weights and start to dominate the estimator (and its variance). Indeed, the

Monte Carlo simulations strongly suggest that this intuition is correct.[33] However, removing such observations with a (non-normalized) weight larger than a given value (for example defined in terms of *p(x)*) comes at the cost of incurring potential asymptotic bias, if it does not disappear fast enough with increasing sample size. When treated observations are removed based on a fixed cut-off value of *p(x)*, the population underlying the definition of the ATET changes. When control observations are removed, we may not be able to reweight the controls successfully towards the distribution of the covariates observed for the treated. Therefore, we suggest setting all weights to zero if their share of the sum of all weights is larger than *t*%, i.e. $w_{i|d_i=0} = w_i \mathbb{1}\left[w_i \Big/ \sum_{j=1}^{N} (1-d_j) w_j \leq t\% \right]$. After this step, the remaining weights are normalized again.[34] This correction disappears as the sample increases as each sample unit has asymptotically no influence on the estimator (at least with discrete covariates). Indeed, such a suggestion was already made by Imbens (2004, p. 23) to account for common support problems. It is important to note that all estimators, whether they are parametric or semiparametric, are treated exactly the same way: Control observations are removed if their IPW weights are above the threshold (and treated are adjusted accordingly).

[33] We thank an anonymous referee for pointing out that this has been shown formally by Chen, Hong and Tarozzi (2008) in a related setting. These authors suggest a regression-like approach which is less likely to have extreme weights. Furthermore, note the similarity between our approach of removing observation with a 'too high' weight, and thus with the largest influence on the predicted mean potential non-treatment outcome, and the literature on robust statistics (e.g. Huber and Ronchetti, 2009) in which many approaches aim at reducing the impact of single observations when they become too influential. The fact that the approaches developed in this literature (for different estimation problems) are far more sophisticated than our naïve, but effective, approach, suggest that it should be possible to improve upon our approach in future work. This is however beyond the scope of this paper. A referee also pointed us to the work of Hill and Renault (2010) who use the same motivation to trim moment conditions in a GMM time series framework.

[34] To avoid a severely unbalanced sample induced by trimming, we also remove all treated observations with a value of *p(x)* larger then smallest value of *p(x)* among the control observations removed by this trimming rule (if such observations exist at all). Tables C.3 and C.4 in internet Appendix C.2 describe how many treated are affected for the particular samples. Not surprisingly the share is the larger the higher the trimming level, the smaller the sample, the larger the share of the treated and the stronger the selection process (see also the discussion in the next section). Table E.1 in internet Appendix E shows the effect of removing the treated on the estimators for the most critical DGP, the one with heavy selection and a large share of treated. We see that removing the treated together with the controls is indeed critical, as there might be a substantial bias for all estimators otherwise.

As suggested by the discussion above and in Imbens (2004), trimming is also relevant to the common support problem and the 'thin-support' problem recently looked at by Khan and Tamer (2009). The key conceptional difference to our trimming rule is that support issues are asymptotic problems while we are concerned with a small sample adjustment only.[35] The common support problem has been discussed by many authors (see the surveys by Heckman, LaLonde, and Smith, 1999, Imbens, 2004, and Imbens and Wooldridge, 2009). Recently, Crump, Hotz, Imbens, and Mitnik (2009) propose removing treated observations with 'extreme' values of the propensity score to improve the precision of the estimator (they recommend using only values of $p(x)$ below 0.9). Of course, at the same time this procedure increases the bias (or changes the estimated parameter by implicitly changing the reference population, which is the same), which will remain asymptotically. There have been different proposals in the literature on how to tackle the common support problem, but they all share the feature that they will lead to asymptotic bias,[36] or give up point identification (Lechner, 2008). In contrast, trimming based on our suggestion vanishes as the sample size increases such that the estimation is asymptotically unbiased.

---

[35] From a practical point of view, our analysis can be seen as a comparison of the performance of different estimators *after* the common support has been enforced. Investigating different ways of ensuring common support is beyond the scope of this paper. Tables C.2 and C.3 in internet Appendix C.2 give more details on maximum weights that follow from this procedure as well as about support issues in the samples used (by checking how many treated were located to the right of the largest control observation after enforcing the different trimming rules). The lower panel of Table E.2 in internet Appendix E shows the effect of using a stricter rule for the treated, namely removing all treated with a larger p-score than the largest p-score of any control remaining after trimming. However, when applying this stricter rule, the results do not change much.

[36] See the excellent discussion of this issue by Busso, DiNardo, and McCrary (2009a). They use four different trimming rules to improve e common support in their Monte Carlo study: the method proposed by Dehejia and Wahba (1999), which is based on comparing the maximum values of $p(x)$ among the treated and controls; the method proposed by Heckman, Ichimura, Smith, and Todd (1998), which is based on requiring a minimum density of $p(x)$; the method brought forward by Ho, Imai, King, and Stuart (2007) which defines the common support as the convex hull of $p(x)$ used by pair matching; and the proposal by Crump, Hotz, Imbens and Mitnik (2009) already mentioned. They conclude that none of the proposals works in the case of heterogeneous treatment effects. Some of them seem to work for some estimators in the case of homogeneous effects.

Khan and Tamer (2009) analyse the problems that may appear for estimators adjusting for covariate differences if identification requires estimation in what they call thin-support regions. Such regions of the covariate space could occur, for example, when one of the covariates has infinite support. This might result in very large (infinite) weights leading to a reduction of the convergence rates together with numerical instability in small samples. Khan and Tamer (2009) develop a new inference routine to account for this abnormal behaviour. Again, this is essentially an asymptotic problem. In contrast, trimming in our simulation merely tackles 'too' large weights in finite samples as there is no asymptotic support problem by the definition of the propensity score.

5 Results

In this section, we first discuss several issues concerning the implementation of the various estimators (5.1). After that, the results are discussed, beginning with issues that concern all estimators simultaneously, like the impact of different features of the data generating process, the specification of the propensity score and the trimming (5.2). Then, we analyse implementational issues that are specific to the particular classes of estimators considered (5.3). Finally, we compare the best estimators across the different classes to come to an overall conclusion (5.4).

When discussing the results, most of our conclusions come from analysing the root mean squared error (RMSE) of the estimators. Among other information, internet Appendix D contains additional information with respect to the absolute bias and the standard deviation of the estimators, which sometimes will be useful to better understand the effect on the RMSE. Since there might be a concern that in particular for small samples some of the estimators have no moments, we also verified our main results based on the mean absolute error. There were no substantial differences (for details, again, see internet Appendix D). Furthermore,

additional sensitivity checks with respect to trimming and specifications of the score are contained in internet Appendix E.

5.1 Implementation of estimators

While Section 2 contains the general principles underlying the different classes of estimators, we present the details for the particular versions of the estimators, as implemented in the simulations, in this section as well as in Appendix A.

5.1.1 All estimators

All estimators are based on (i) a correctly specified model for the propensity score and on (ii) a functionally misspecified model where all eight interaction terms and the two terms capturing non-linearities in age are omitted in the estimation. This is most likely a misspecification that frequently occurs in applications and some robustness in that direction is desirable. This specification problem is relevant as the variables are jointly highly significant in the propensity score as well as in the outcome equations based on Wald-statistics (see Table B.2 in Appendix B.2).

The same trimming rule is used for all estimators by setting t to 4%, and 6% (and 100% for the untrimmed case). This trimming rule is directly based on the propensity score, i.e. the weight that is used in the IPW estimator.[37] The main reason is computational speed, as estimator-specific rules would require additional computational steps in a simulation study that is already computationally extremely expensive. A further motivation is that this rule is very easy to implement in applications and that the weights used by the other (consistent) estimators should be at least asymptotically similar to the IPW weights.

[37] The rule is only applied once and not iteratively. Thus, in the trimmed sample the weights may be above the threshold. Table C.3 and C.4 in the internet Appendix C show the largest weights used in the particular setting after trimming. The largest weight in the untrimmed is around 21% and drops to 6% after trimming (heavy selection, large share of treated, smaller sample), but in most DGP's the levels and the differences of the weights are much smaller.

5.1.2 Inverse probability weighting and tilting

The IPW and IPT estimators described in Section 2 are implemented as stated (there are no tuning parameters to choose). The IPW is the version that performed well in Busso, DiNardo and McCrary (2009a, b).

5.1.3 Direct matching

We consider the following types of propensity score matching estimators: Pair-matching, radius matching and radius matching with linear and non-linear post-matching regressions. Before looking at these estimators in turn, let us discuss other features that have been varied but are common to all estimators: (i) To measure the distance between observations we consider the propensity score as well as its linear index (this monotone transformation may matter at the boundary of the propensity score where the c.d.f. is highly non-linear); (ii) We also use matching estimators that use a Mahalanobis matching framework in which the propensity score or its linear index is supplemented by two covariates, namely the indicator variable for being *female,* and *average earnings in the 10 years before becoming unemployed*. Both are good predictors of post-training earnings and employment as well as programme participation (they are jointly significant in the participation and both outcome equations based on Wald tests; see Table B.2 in Appendix B.2).

Radius matching requires defining a radius, or caliper size, in terms of the distance between treated and non-treated. Since no well established algorithm exists, we follow Lechner, Miquel and Wunsch (2011) who suggest defining the caliper size in terms of the largest distance calculated from pair-matching. Here, we use half that distance, as well as 1.5 and three times that distance. If a caliper is empty, which may happen only in the first case, the nearest neighbour is chosen. When computing the local mean of the outcome variables in a caliper, the observations within the caliper are weighted proportionally to the inverse of their distance to the respective treated they are matched to.

Finally, radius matching is combined with linear regression (both outcomes) or logit regression (employment only) to remove bias due to mismatch as explained above. See Appendix A.1 for all details. In total we consider 48 matching estimators for employment and 32 matching estimators for earnings.

The final remark concerns the use of matching estimators: to foster computational efficiency in a very demanding simulation exercise (in particular for the large sample size), we remove some variants that are clearly dominated by similar ones. To be specific, we discard all radius matching estimators matching only on the propensity score or its linear index, respectively, as they are always dominated by the Mahalanobis distance-based versions which additionally include the two covariates.

5.1.4 Kernel matching

The details on ridge regression matching are presented in Appendix A.2. The main feature we vary is the bandwidth. Starting with the value suggested by least squares cross-validation, we also take one third of and three times that value. Furthermore, we use a Silverman (1986) type rule of thumb for the Epanechnikov kernel. The reason for considering different values of the bandwidth is that, intuitively, the cross-validation bandwidth is optimal for the regression curve but not for the particular average of it that enters the ATET, see also the discussion in Frölich (2005) and Imbens and Wooldridge (2009). [38] Therefore, one would expect that some undersmoothing is optimal (although this turns out to rarely be the case in the simulation). In addition, it is interesting to see the sensitivity of the estimator with respect to the important bandwidth choice decision. Furthermore, for the binary outcome an estimator

[38] Frölich (2005) offers a plug-in method for optimal bandwidth selection in kernel matching based on an approximation of the mean squared error. However, based on his simulations he concludes that the approximation is not sufficiently accurate for the sample sizes he considered (200, 1000). For this reason, we do not implement his procedure for bandwidth selection. As an alternative, Frölich (2005) also considers conventional cross-validation and finds that it performs rather well in his simulations, even though asymptotically it does not provide the optimal bandwidth.

based on a local logit instead of a local linear specification is also used. In total we have eight estimators for the binary outcome and four estimators for the semi-continuous outcome.

5.1.5 Parametric models

The parametric models generally consist of two versions: one that is applied just to the non-treated (whereas for the treated, simply their sample average outcome is computed), and another one that also includes a separate parametric model for the treated. As expected, these two versions lead to almost identical results.

We consider several model choices. Firstly, a linear regression model is used for both the binary and the semi-continuous outcome variable even though this constitutes a misspecification in both cases (due to bounded theoretical support and a mass point at zero, respectively). Therefore, we also use a tobit model both estimated by maximum likelihood (henceforth simply referred to as tobit) as well as in its control function form (i.e., the heckit model; see Heckman, 1976) for earnings, as well as a probit model for the binary employment outcome. Finally, we use flexible data-driven OLS and probit estimation that selectively adds higher order and interaction terms to chose the optimal model with respect to minimization of the corrected Akaike information criterion (AIC). This basic implementation of sieve regression, see for instance Chen (2007), is outlined in Appendix E.

In total we use 7 estimators for employment and 8 estimators for earnings (two versions of OLS, probit or heckit and tobit, respectively, flexible probit/OLS estimation based on the corrected AIC, and DR estimation based on probit or heckit and OLS, respectively, that weights the regression by $\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}$).[39]

[39] Since heckit turned out to be very unstable for the smaller samples, DR with OLS was included for earnings as well, despite its misspecification for the semi-continuous outcome. For the latter reason DR based on OLS was not used for the

5.2 Results for features that concern all estimators

There is a wealth of information produced by the Monte Carlo study. For the employment and earnings outcomes we have more than 5700 and 3700 data points, respectively, for each measure of estimator quality we consider. Thus, we have to summarise this information. We do so by using linear regression analysis in which the features of the DGPs, the propensity score specifications, and the outcome variables used are coded as covariates (partially interacted). Due to the large expected heterogeneity and non-linearity, this analysis is conducted within strata defined by the sample size and classes of estimators.[40]

Table 5.1 (employment) and Table 5.2 (earnings) contain the coefficients of the regression results for the root mean squared error, whereas the results for the bias and the standard deviation are relegated to internet Appendix D.2 (Tables D.4 and D.5 for employment and Tables D.6 and D.7 for earnings).

5.2.1 Strength of selection and share of treated

The upper panels of those tables contain indicator variables for the magnitude of the selection and the share of the treated (the medium cases being the references).[41] We find that the RMSE increases in the strength of selection and the sources appear to be both the bias and the precision of the estimators. When looking at the 10% and 50% shares of treated, this result is mainly driven by precision, while the impact of the strength of selection on the bias increases when the number of control observations is reduced further.

---

binary outcome, for which DR with probit works fine. Finally, as both the maximum likelihood and heckit estimators of the tobit model appear to be uncompetitive compared to OLS (see Table D.8 in internet Appendix D), a DR version of the maximum likelihood version was not computed.

[40] This approach is very similar to ideas underlying meta-analysis which uses regression techniques to summarize the results of different studies (for a recent application in programme evaluation see Card, Kluve and Weber, 2010).

[41] The tobit and heckit estimators turned out to be highly unstable for the earnings outcome for the small and intermediate sample sizes. Therefore, these estimators are excluded from the regressions.

Considering the influence of the share of the treated, the results are again clear-cut: a balanced sample leads to the lowest RMSE. In particular for the sample with very few control observations, there is a significant small sample bias for all types of estimators.

Table 5.1: Features of the estimators by OLS regression for employment outcome

Variables (all indicators)		IPW			Kernel			Matching			Parametric		
Sample Size		300	1200	4800	300	1200	4800	300	1200	4800	300	1200	4800
Constant		5.9	3.0	1.0	7.1	3.1	1.3	7.1	3.6	2.0	7.1	4.1	1.3
Features of the data generating process													
Selection:	Random	(-1.3)	-1.0	(-0.7)	-0.8	-0.8	-0.9	-1.0	-0.9	-0.8	-1.0	(-1.0)	-0.7
	Observed	0	0	0	0	0	0	0	0	0	0	0	0
	Strong	3.4	2.8	2.8	2.8	2.7	2.5	2.5	2.5	2.4	2.3	2.8	1.9
Share treated:	10%	-	1.0	(0.5)	-	1.3	0.6	-	1.7	0.6	-	1.8	0.5
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	3.8	2.3	-	3.3	2.1	-	4.4	2.1	-	4.2	1.7
Features of the estimators													
Misspecified p-score		(0.7)	(0.6)	1.1	-0.8	0.2	1.1	-0.8	(-0.2)	0.8	-0.3	(-0.3)	0.9
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		(1.5)	(0.5)	0.8	-0.5	-0.3	(-0.0)	-0.8	-0.9	-0.1	-0.7	-1.3	(-0.1)
Trimming max 4%		(-0.9)	(-0.8)	(-0.3)	-0.7	-0.4	(-0.1)	-1.0	-1.1	-0.2	-0.9	-1.4	(-0.2)
Inverse probability tilting		(1.3)	(-0.1)	(-0.4)									
Bandwidth:	Low				(0.2)	(0.2)	(0.0)						
	Cross validation				0	0	0						
	High				-0.6	(-0.2)	(-0.0)						
	Rule of thumb				(0.3)	(0.1)	(0.0)						
	Local logit				0.9	0.3	(-0.1)						
Nearest neighbour								2.7	1.6	0.2			
Radius matching:	Radius low							0.8	(0.3)	(-0.1)			
	medium							0	0	0			
	large							(-0.0)	(0.1)	(0.1)			
No adjustment								0	0	0			
Regression adjustment								0.7	0.5	-0.9			
Logit adjustment								-0.8	-1.0	(0.1)			
PScore instead of linear index								(0.1)	(0.1)	(0.1)			
Regression for treated											(-0.0)	(0.4)	(0.1)
Robust											0.4	(-0.5)	(0.1)
Probit											(-0.0)	(-0.8)	(-0.2)
Statistics													
R^2 (in %)		54	67	58	74	82	74	73	59	70	88	33	75
Number of observations		36	108	108	144	432	432	540	1620	1620	108	324	324

Note: Dependent variable: *R*oot*M*ean*S*quared*E*rror. The two larger samples also contain additional data generating processes. The largest sample is based on a reduced number of estimators. All coefficients are in %. Coefficients that are not significant at the 5% level (conventional OLS standard errors) appear in parentheses.

5.2.2 Functional misspecification of the propensity score

A functional misspecification which leads to an inconsistent estimation of the propensity score leads to an increase of the bias (at least for the larger samples) and to a reduction of the variance (probably because the misspecified propensity score depends on fewer variables and may thus be more precisely estimated) of the estimators. Considering the joint impact on the RMSE, we find that in the smallest sample the gain in precision due to the misspecification dominates, while in the largest sample the bias dominates. In the final section, we discuss

this issue again to see whether the different estimators are affected differently by this kind of misspecification.

5.2.3 Trimming

Before presenting the results of the different estimators for different levels of trimming, it seems worth investigating how many observations are trimmed depending on the features of the DGPs and the levels of trimming. The details are provided in Tables C.2 to C.4 in internet Appendix C.

Table 5.2: Analysis of features of matching estimators by OLS regression: Earnings

Variables (all indicators)		IPW			Kernel			Matching			Parametric*)		
Sample Size		300	1200	4800	300	1200	4800	300	1200	4800	300	1200	4800
Constant		(168)	(60)	28	171	83	34	180	104	62	171	117	33
		Features of the data generating process											
Selection:	Random	(-95)	(-79)	-21	-32	-29	-27	-36	-39	-30	-34	(-30)	-25
	Observed	0	0	0	0	0	0	0	0	0	0	0	0
	Strong	(155)	105	73	65	69	72	63	63	69	82	134	63
Share treated:	10%	-	(25)	(11)	-	31	13	-	40	12	-	(30)	(8)
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	233	58	-	70	55	-	99	50	-	163	42
		Features of the estimators											
Misspecified p-score		165	(65)	23	(-3)	11	23	-9	7	32	(-13)	(-33)	33
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		(-141)	(-88)	(17)	(-8)	-9	(-2)	-24	-26	(-4)	-29	-89	(-4)
Trimming max 4%		(-143)	(-78)	(-9)	-15	-13	(-3)	-32	-34	-6	-36	-95	(-5)
Inverse probability tilting		204	95	(-9)									
Bandwidth: Low					(-7)	(-0)	(3)						
Cross validation					0	0	0						
High					-13	(-4)	(0)						
Rule of thumb					(2)	(0)	(1)						
Nearest neighbour								56	29	-12			
Radius matching:	Radius low							16	(5)	(-3)			
	medium							0	0	0			
	large							(-1)	(4)	7			
Regression adjustment								(-0)	(-6)	-36			
PScore instead of linear index								(3)	(2)	(-1)			
Regression for treated											(-0)	(1)	(0)
Robust											37	(29)	(2)
		Statistics											
R^2 (in %)		43	37	60	89	83	73	70	60	72	73	26	78
Number of observations		36	108	108	72	216	216	360	1080	1080	72	216	216

Note: Dependent variable: RMSE. The two larger samples also contain additional data generating processes. The largest sample is based on a reduced number of estimators. Coefficients that are not significant at the 5% level (conventional OLS standard errors) appear in parentheses.
*): Heckit and Tobit estimates are very unstable and therefore excluded from the regressions presented in this table (see Table D.8 in internet Appendix D for details).

By construction, the number of trimmed observations decreases with an increasing level of the threshold. However, even for a level of 4%, in the worst case no more than 4.3 control observations are trimmed on average. In all other cases, this number is considerably

lower. Thus, very few control observations are trimmed by this rule, but of course these are the control observations with the largest influence on the final estimate.

Although only few control observations are trimmed, the regressions suggest that moving from no trimming to discarding all observations with weights larger than 6% leads to a considerable reduction in the RMSE. A trimming rule with a lower maximum weight (4%) still decreases the RMSE, but only by a small amount. The RMSE reduction is driven by a reduction in the small sample bias and in the variance.

The effects of trimming are very much DGP dependent. Under those features of the DGP that entail the largest deletion of observations (strong selection and small share of controls), the effects of trimming seem to be unambiguously positive and large in that both bias and variance are reduced. In the other cases (in which trimming really does not change much as extreme weights rarely occur), these findings hold only for the smallest sample (if at all). We conclude that trimming in the proposed way seems to be very effective in cases where it is most needed, while it does not hurt much in the other scenarios. This issue of trimming will be taken up again when considering selected single estimators in detail in section 5.4.

5.3 Estimator-specific issues

5.3.1 Inverse probability weighting and tilting

Tables 5.1 and 5.2 reveal that in small samples, inverse probability weighting (IPW) seems to dominate inverse probability tilting (IPT) in terms of RMSE because of its better bias and variance properties (see Tables D.4-D.7 in Appendix D). For large samples, however, the converse holds due to the lower finite sample bias of IPT. These features will become more obvious (and more differentiated) from the direct comparison of these two estimators that is contained in section 5.4.

5.3.2 Direct matching

When comparing nearest neighbour matching to the other direct matching estimators we replicate the result frequently found in the literature: although being the least biased for all sample sizes nearest neighbour matching is not competitive in terms of RMSE, because of its substantially larger variability. Yet, for the largest sample, which has a sample size that was not considered in other relevant studies, we obtain a surprising result: as the absolute difference in precision is reduced due to the general decrease of the variance with increasing sample size (all variances tend to zero asymptotically), the better bias properties become more important as the bias become more relevant. Despite this feature, the results later on will show that nearest neighbour matching is still dominated by other matching methods.

Considering the caliper size for radius matching, the findings are again in line with our expectations: The smaller the caliper, the larger the variance and the smaller the bias. With respect to the post-matching regression adjustment, we observe a similar phenomenon: the bias is reduced but the variance increases and the regression adjustments become more attractive as the sample gets larger. For the binary outcome, the logit adjustment is superior to the linear regression adjustment, at least for the smaller samples.

The results concerning the inclusion of additional covariates in Mahalanobis matching are similar in the sense that the variance is reduced and the bias (somewhat) increased. In our simulations the gains in precision dominate.[42] Finally, using the linear index instead of the propensity score does not have much of an effect at all.

[42] To save computation time the matching estimators without including additional covariates in a Mahalanobis metric have only been computed for the small and medium sized samples. In those simulations they have always been dominated by the versions that include the covariates. Therefore, the former are not considered in the tables of this section that are only based on estimators computed for all sample sizes.

5.3.3 Kernel matching

Although the results for the different bandwidths are not really clear-cut, on average choosing the largest bandwidth (here, three times of what is suggested by least squares cross-validation) seems to be the dominant strategy. We will take up that issue again in the next section. Concerning the issue whether to use the local logit or local linear regression for the binary outcome, the results suggest that local logit performs only slightly better in the larger sample, whereas local linear regression dominates over all. In conclusion, this estimator does not appear to be sensitive to reasonably chosen smoothing parameters.

5.3.4 Parametric models

Among the parametric models, standard probit and OLS are the preferred choices for the employment and earnings outcome, respectively, in terms of the RMSE. It may seem surprising that OLS is superior to the tobit and heckit estimators in the earnings regressions despite the mass point at zero and that both OLS and probit generally outperform DR procedures. A closer inspection of the results shows that the disappointing performance of the DR and tobit/heckit estimators is rooted in their comparably large variances in the small and medium samples, in particular when the share of treated is high (and thus the number of controls on which these regressions are based is very small). In particular the heckit-based DR estimators seem to suffer from numerical instabilities when the number of observations is too small. This is also suggested by its 'non-normal' empirical distribution. Therefore, the tobit/heckit estimators are not considered in the regressions presented in Table 5.2, [43] but their standard versions are compared to OLS in Table D.8 of the internet Appendix D. Even

[43] See also Kang and Schafer (2007) who examine the finite sample behaviour of DR estimators in a missing data context using up to 1000 observations. None of the investigated DR methods outperform the simple regression-based prediction of the missing values. Therefore, the authors conclude that using two incorrect models in DR estimation is not necessarily better than a regression based on just one wrong specification.

without heckit, DR does not appear attractive because of its larger variability compared to standard regression (or IPW, see below). A further result that may come as a surprise is that standard OLS and probit perform clearly better in terms of RMSE than flexible estimation based on the corrected AIC (the results of which are not reported, but available on request). Finally, estimating an additional model for the treated, too, does not change the results in any relevant way.

5.4 Comparisons across different classes of estimators

Having compared the different features of the estimators and the DGPs within classes of estimators, we now move to comparisons across classes. The aim is to come to a final conclusion about which estimator appears to be most suitable for particular applications. Therefore, for a selected group of estimators Tables 5.3 to 5.5 present the difference in %-points of RMSE relative to the best estimator (which is marked 'B' if it is part of the group of estimators considered in the table), as well as the bias, the standard deviation, the skewness and the kurtosis of the estimators' Monte Carlo distributions. The latter two statistics are included to see whether there are any important deviations from normality which may cause problems for inference. Large values of the kurtosis are also a good indicator of estimator instability leading to important outliers, which is very undesirable in empirical applications.

To be able to present the results in a concise way, we have selected estimators that dominate their respective class of estimators. Dominance is judged on the basis of the RMSE and is defined in a two-step procedure within the class of estimators (direct matching / IPW-IPT / kernel / parametric). First, a minimum requirement is imposed: For each scenario the best estimator is determined and estimators are grouped according to the distance to that estimator (0-25%, 25%-100%, > 100%). To be considered further, estimators have to be in the

best group in at least half of the cases and never be in the worst group. Among that group, we choose the best estimators in terms of average RMSE.[44]

The dominant estimator is regression-adjusted radius matching (using linear regression for earnings and logit for employment) with additional predictors based on the linear index and using the large radius. Even though it is not competitive, we also consider simple pair-matching based on the propensity score, in that it represents a benchmark frequently used in practice. Concerning the class of kernel matching estimators, there was no clear-cut winner with respect to the bandwidth selection rules. Therefore, we present the results for the estimators with the largest and the smallest bandwidth to be able to consider the sensitivity in that respect in greater detail. As local linear regression is (slightly) superior to local logit for the employment outcome, all results in the tables refer to the former method. Among the parametric methods, the non-weighted OLS and probit estimators are the best, closely followed by the probit and OLS DR-versions that are presented as well.

The comparison across classes starts by taking up the issue of trimming again. Table 5.3 shows the results without trimming as well as for two different levels of trimming, averaged over all DGPs separately for the correctly and incorrectly specified propensity score. The relative RMSEs refer to the best estimator under any trimming rule.

Trimming is indeed important for the correctly specified as well as the misspecified model. On average, all estimators but IPT unambiguously benefit from trimming in terms of bias, precision, skewness and kurtosis, particularly in the case of the semi-continuous outcome. When moving from no trimming to 6% the gains appear fairly large, while trimming

[44] Obviously, these criteria are arbitrary, but they insure that estimators perform reasonably in a large group of DGP's and specifications. The final conclusions are not very sensitive to how the respective groups are formed and which exact shares are imposed. For IPT the criteria have been weakened as there are not so many versions of weighting estimators.

further observations using the 4% cut-off value only leads to small additional gains (with the exception of IPT).

Table 5.3: Comparison of the properties of the selected estimators: trimming

	Employment								Earnings							
			Kernel		Matching		Probit				Kernel		Matching		OLS	
	IPW	IPT	high	low	logit	pair		DR	IPW	IPT	high	low	OLS	pair		DR
Propensity score correctly specified																
Without trimming																
RelRMSE	39	12	16	26	16	93	10	28	46	80	16	35	36	101	62	144
Bias	0.5	0.9	1.0	1.5	0.9	0.3	0.9	0.9	10	27	29	39	23	5	29	9
Std. dev.	5.1	3.9	4.1	4.2	4.1	7.1	3.9	4.6	129	154	93	109	117	178	137	216
Skew.	0.1	0.0	0.0	0.0	0.1	0.2	0.0	0.0	-0.4	-3.5	0.0	-0.2	-0.4	-0.2	-2.8	-2.8
Kurtosis	3.4	3.0	3.0	3.7	3.0	3.2	3.0	3.2	4.7	218	3.1	3.6	7.0	3.4	172	173
Trimming level 6%																
RelRMSE	11	40	9	16	9	70	2	10	12	49	7	22	12	73	3	21
Bias	0.3	0.5	0.7	1.1	0.8	0.2	0.6	0.5	7	9	21	30	10	4	23	6
Std. dev.	4.1	5.1	3.9	4.0	3.9	6.2	3.7	4.0	99	131	90	100	98	153	86	108
Skew.	0.0	0.1	0.0	0.0	0.1	0.2	0.0	0.0	-0.1	-0.4	0.0	-0.1	-0.3	-0.2	0.0	-0.1
Kurtosis	3.0	3.4	3.0	3.4	3.0	3.3	3.0	3.0	3.1	5.2	3.1	3.5	7.5	3.6	5.2	7.4
Trimming level 4%																
RelRMSE	7	2	7	14	7	61	B	7	7	15	4	18	4	63	B	15
Bias	0.2	0.6	0.7	1.0	0.7	0.1	0.5	0.5	6	21	19	27	8	3	22	5
Std. dev.	3.9	3.7	3.9	3.9	3.9	5.9	3.6	3.9	94	97	88	96	92	145	84	101
Skew.	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	-0.1	-3.0	0.0	-0.1	-0.2	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.0	3.4	3.0	3.4	3.0	3.1	3.1	476	3.1	3.4	8.2	3.7	4.5	7.0
Propensity score misspecified																
Without trimming																
RelRMSE	45	5	29	21	10	74	18	28	26	192	16	10	9	51	16	22
Bias	3.0	0.9	2.8	2.4	1.4	2.9	2.3	2.4	71	27	76	65	52	68	75	66
Std. dev.	4.3	3.9	3.5	3.6	3.8	5.6	3.6	4.0	109	314	84	88	98	141	88	107
Skew.	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.1	-0.4	-1.7	0.0	-0.1	-0.6	-0.2	-0.2	-0.1
Kurtosis	3.3	3.0	3.0	3.1	3.0	2.9	3.0	3.1	5.4	721	3.1	3.2	14.7	3.1	7.2	5.9
Trimming level 6%																
RelRMSE	31	97	33	25	8	67	14	18	13	73	15	9	3	43	10	13
Bias	2.8	3.3	3.0	2.5	1.5	2.7	2.2	2.2	68	64	75	64	53	65	71	63
Std. dev.	3.7	6.3	3.6	3.7	3.7	5.4	3.5	3.7	92	168	86	89	90	134	83	97
Skew.	0.0	0.2	0.0	0.0	0.1	0.1	0.0	0.1	-0.1	-1.1	0.0	-0.1	-0.5	-0.2	-0.1	-0.1
Kurtosis	3.0	4.8	3.0	3.1	3.0	2.9	3.0	3.0	3.1	12	3.1	3.2	13.9	3.2	4.7	6.2
Trimming level 4%																
RelRMSE	27	B	31	23	7	62	13	15	9	83	13	7	B	39	8	9
Bias	2.7	0.8	2.9	2.4	1.5	2.6	2.1	2.1	66	25	73	62	53	63	70	62
Std. dev.	3.6	3.7	3.6	3.6	3.7	5.3	3.5	3.6	88	194	85	87	87	129	81	93
Skew.	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	-1.9	0.0	-0.1	-0.5	-0.2	-0.1	-0.1
Kurtosis	3.0	3.0	3.0	3.1	3.0	2.9	3.0	3.0	3.1	524	3.1	3.2	13.5	3.2	4.6	5.5

Note: RelRMSE: Difference in relative root mean squared error in % compared to best estimator, marked as 'B'. Bias and standard deviation for employment are given in %-points. DR: Double robust (weighted) version of estimator.

As already discussed, most of the gains originate from the DGPs with heavy selection and few controls. The gains are probably larger for the correctly specified model because the

propensity score of this model contains additional interaction terms that should lead to a 'better' individual prediction. Since such a prediction is likely to increase the (unconditional) variance of the propensity score, it becomes more likely that the weights are above the threshold.

The upper part of Table 5.3 that relates to the correctly specified model sheds light on the potential threat that trimming might lead to a bias of the estimators. If anything, the (small sample) bias is reduced, but certainly not increased. Furthermore, the trimming level does not appear to have any relevant impact on the ordering of the respective estimators with the exception of IPT, which, in the binary outcome case, fairs better under no trimming than for 6% trimming, but best for 4% trimming.

Comparing the estimators to each other shows that most appear to lie within a reasonable distance to the respective best estimator, with the exception of pair matching, which is never competitive in terms of the RMSE due to its large variance. Moreover, when the propensity score is misspecified, IPT does often worse than pair matching and again, its RMSE is not a monotonic function of trimming. For the case of a correctly specified model, probit and OLS appear to be the best estimators in terms of RMSE, while for the functionally misspecified propensity score, IPT (for employment) and the OLS regression adjusted radius matching (for earnings) are the best.

Note that the distributional properties of the estimators are dependent on the outcome considered. For the binary employment outcome, the best performing IPT, logit adjusted radius matching and the probit estimators also have 'good' higher order moments. All the other estimators appear to have reasonable properties as well. For the semi-continuous earnings outcome, the results look strikingly different. Although the same classes of estimators (OLS adjusted radius matching and OLS) are preferred on RMSE grounds, they have fat tails despite the trimming (but only in the small and medium sized samples as can be seen in Table

5.4). A particularly surprising feature is exhibited by IPT, which is among the best estimators for the binary outcome but performs strikingly bad for the continuous one.

It is likely that this ranking based on averaging across DGP features and propensity score specifications is subject to some heterogeneity. To investigate this issue further, Tables 5.4 and 5.5 present different subsets of the results. As 4%-trimming improves any method to some extent (even IPT) all results in these tables refer to the 4%-trimmed versions of the estimators only.

Table 5.4 is concerned with variations in the sample size. Looking at the upper three blocks of the table, the average results for the employment outcome shown in Table 5.3 are confirmed. Note, however, that despite its good performance for the binary outcomes, IPT performs very badly for the earnings outcome in the small samples. Even for the medium sized sample its performance is very unsatisfactory if the propensity score is misspecified. For the large sample it appears however to be the best estimator for both outcome variables.[45] Furthermore, note that fat tails are also present for radius matching, and OLS, while the other estimators do not have this problem and are (apart from pair matching and IPT) very close in terms of the RMSE. For the largest sample these tail problems disappear and OLS-adjusted radius matching dominates all other estimators for earnings.

It may seem surprising that IPW does not outperform the other estimators in the large sample despite the property that it is asymptotically efficient when the propensity score is non-parametrically estimated. The reasons for this could be that (i) the propensity score is parametrically estimated, or / and that (ii) the score is not re-estimated after trimming, which might lead to some improvement (but could also lead to new problems of very large weights).

[45] Note that although this is true on average over all different DGPs considered for the large sample, the radius matching estimator with regression adjustment performs better for the correctly specified scores.

Note that changing the sample sizes in our comparisons goes along with changing other DGP features for the smaller sample sizes: for the smallest sample we only consider the case of 50% treated, while the larger samples also contain the more problematic DGP's with 10% and 90% treated. Furthermore, note that because specifications with incorrectly specified propensity scores are also included, they are not expected to be unbiased. Therefore, to study the pure effect of the sample size in settings where the estimators are consistent, the lower three blocks of Table 5.4 only consider cases with 50% treated and a correct specification of the propensity score.

Before comparing the relative performance of the estimators, a few general observations concerning all estimators are in order. Firstly, compared to the standard deviation the bias is small when the model is correctly specified. There are however important differences between the two outcomes: While the bias shrinks with sample size for the employment outcome, this is not necessarily the case for the earnings outcome. For example, the bias of IPT and OLS for earnings seems to be independent of the sample size, while the bias of IPT and the probit disappears for the employment outcome. The performance of OLS suggests that the linear model is not flexible enough and thus misspecified (leading to an asymptotic bias), while the probit seems to be a good approximation of the conditional expectation of the outcome variable. A similar phenomenon occurs for the kernel estimators, but the level of the bias is smaller in this case. Secondly, the standard deviations are approximately reduced by half when quadrupling the sample size. Again, this is more obvious for the binary outcome than for the semi-continuous outcome. It is also interesting to note that while the relative differences in the RMSE of the estimators are moderate in the smallest sample (of course with the exceptions already discussed), they become more pronounced when the sample size increases.

Table 5.4: Comparison of the properties of the selected estimators: sample size

	Employment								Earnings							
			Kernel		Matching		Probit				Kernel		Matching		OLS	
	IPW	IPT	high	low	logit	pair		DR	IPW	IPT	high	low	OLS	pair		DR
							N = 300									
RelRMSE	1	5	2	4	B	50	2	6	4	99	3	6	0.3	53	B	15
Bias	1.3	1.8	1.3	1.4	0.9	1.1	2.0	2.0	36	27	33	34	27	30	42	31
Std. dev.	6.2	6.5	6.4	6.4	6.4	9.5	6.2	6.5	148	299	148	152	148	226	140	167
Skew.	0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.0	-0.1	-22.7	0.0	-0.1	-0.3	-0.4	-0.1	-0.1
Kurtosis	3.0	3.0	3.1	3.1	3.0	4.3	3.0	2.9	3.1	2910	3.1	3.2	7.5	4.6	3.7	15.9
							N = 1200									
RelRMSE	12	B	13	12	5	56	3	7	3	68	3	5	1	45	B	7
Bias	1.5	0.7	1.7	1.7	1.1	1.3	1.3	1.3	36	24	44	45	30	32	46	33
Std. dev.	4.4	4.3	4.4	4.4	4.5	6.5	4.2	4.3	106	192	102	105	108	160	98	112
Skew.	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	-0.1	1.9	0.0	-0.1	-0.6	-0.2	-0.1	-0.2
Kurtosis	3.0	3.0	3.0	3.2	3.0	2.9	3.0	3.0	3.1	194	3.1	3.5	19	3.3	6.3	5.9
							N = 4800									
RelRMSE	46	B	52	48	20	87	20	27	37	B	43	47	20	78	32	34
Bias	1.5	0.3	2.0	1.8	1.2	1.5	1.1	1.1	36	21	51	48	32	35	47	35
Std. dev.	2.3	2.1	2.2	2.2	2.2	3.4	2.0	2.3	58	49	51	58	51	85	47	59
Skew.	0.0	0.0	0.0	-0.1	0.0	0.1	0.0	0.0	-0.1	0.0	0.0	-0.1	0.0	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.0	3.3	3.0	3.0	3.0	3.1	3.1	3.1	3.1	3.2	3.2	3.2	3.1	3.4
							N = 300 (correctly specified score; 50% treated)									
RelRMSE	B	8	0.02	3	3	58	4	9	5	44	2	8	5	65	B	24
Bias	0.3	2.0	0.4	0.6	1.0	0.1	1.7	1.7	9	21	8	9	13	3	22	3
Std. dev.	6.5	6.8	6.5	6.7	6.6	10.3	6.5	6.8	153	208	149	157	153	240	143	180
Skew.	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	-0.1	-27.6	0.0	0.0	-0.3	-0.4	0.0	0.1
Kurtosis	3.0	2.9	3.1	3.1	3.1	5.7	2.9	2.8	3.1	3018	3.1	3.3	9.0	5.9	4.0	24.0
							N = 1200 (correctly specified score; 50% treated)									
RelRMSE	14	1	14	25	11	67	B	13	15	1	9	23	1	77	B	11
Bias	0.2	0.2	0.3	0.8	0.7	0.2	0.2	0.1	5	18	7	15	5	1	18	5
Std. dev.	3.3	2.9	3.3	3.5	3.1	4.8	2.9	3.3	81	67	76	85	71	125	66	78
Skew.	0.1	0.1	0.0	0.3	0.1	0.2	0.1	0.1	-0.1	-0.1	0.0	0.1	0.0	-0.2	0.0	0.0
Kurtosis	3.0	2.9	3.0	3.5	2.9	3.0	3.0	3.0	3.0	3.5	3.0	3.5	2.9	3.3	3.0	3.1
							N = 4800 (correctly specified score; 50% treated)									
RelRMSE	22	0.2	25	40	15	74	B	19	29	15	15	48	0.1*	87	17	28
Bias	0.1	0.1	0.5	0.2	0.5	0.1	0.1	0.1	1	20	10	11	3	3	21	8
Std. dev.	1.7	1.4	1.7	2.0	1.5	2.5	1.4	1.7	46	33	40	52	36	67	33	45
Skew.	0.0	0.0	0.0	0.1	0.0	0.2	0.0	0.0	-0.1	0.0	0.0	0.1	0.0	-0.1	0.0	0.0
Kurtosis	3.0	3.0	2.9	2.9	3.0	3.0	2.9	2.9	3.1	3.1	3.1	3.0	3.2	3.0	3.1	3.4

Note: RelRMSE: Difference in relative root mean squared error in % compared to best estimator. Bias and standard deviation for employment is given in %. All results based on relative trimming level of 4%. *The best estimator is this estimator with 6% trimming.

With larger sample sizes (for the correct specification of the score and 50% treated) probit dominates for the employment outcome with IPT a close second while regression-adjusted radius matching is in third place with a RMSE that is 15% higher than the one of the probit. For the earnings outcome, this order is reversed for the largest sample size because the biases of OLS and IPT, which do not decrease with the sample size, are starting to dominate

the RMSE, while in the medium sample all three estimators perform similarly well (because OLS and IPT always have a larger bias but a smaller variance). In the smallest sample all three estimators are, as before, fat-tailed, but the tails of IPT are extreme (and it also has a large variance). Note that the double-robust version of OLS (and probit) does not have the bias problem, but is not precise enough to dominate the other estimators. It is worth mentioning that these results are somewhat contrary to the findings by Busso, DiNardo and McCrary (2009a, b) and Frölich (2004) which favour IPW and kernel matching, respectively. Although those estimators do not perform badly, they are nowhere near the top, with the exception of the smallest sample (see the further discussions of these differences below).

The upper two blocks of Table 5.5 report the results using a correctly and an incorrectly specified propensity score. While for the correctly specified propensity score all estimators appear to be close, except for pair matching, the parametric ones are the best.[46] Note, however, the very fat tails of IPT for earnings. Under misspecification, IPT and regression-adjusted radius matching dominate in the binary and semi-continuous outcome case, respectively, as these estimators have the smallest bias, which points to a desirable robustness property. However, the variance of IPT is again comparably high (and tails very fat) such that it is not competitive in terms of the RMSE.[47]

Next, different magnitudes of selection are evaluated. In the case of random selection all estimators are almost unbiased and perform well apart from pair matching and, when the

[46] As the parametric models mirror the specification of the propensity score, the model with the correctly specified score implies that the parametric models contain these interaction term as well and are, thus, more flexibly specified than those with an incorrectly specified score.

[47] For completeness, a case of over-specification of the propensity score has been considered as well by additionally including squares of the seven continuous variables. For the medium and the large sample size the results are stable (standard errors increase slightly), while for the small sample the model is now clearly too flexible. The details are presented in Table E.2 in Internet Appendix E.3.

earnings outcome is considered, also apart from IPT. Surprisingly, the fat-tail problem observed before is particularly acute for this most innocuous case, where the propensity score should play no role in the adjustment. A similar result, but now with some bias, is present for the 'normal' selection process. For cases with strong selection, it is again IPT and radius matching which dominate for employment and earnings, respectively, due to being least biased. In terms of RMSE, probit and OLS do not lack far behind as they are most precise. It is important to realize that when increasing the strength of selection from random to strong the support issue becomes more acute, which may then explain why the relative performance of some estimators deteriorates rapidly in the case of strong selection (which is then based on fewer observations), although the order of the relative performance of the estimators is not much affected by this.

Finally, consider variation in the percentage of the treated. For the share of treated being just 10%, we observe that IPT dominates under both outcomes, which suggests that this estimator is particularly attractive when the number of available controls is large. As the share of controls increases, the attractiveness of IPT in terms of RMSE decreases drastically when considering the earnings outcome. When the treated share is 90% (incidentally this is also a case with more limited overlap of the finite sample support of the propensity score), radius matching has the smallest RSME under the earnings outcomes, while being a close second after IPT for the binary outcome case. Also note that the parametric methods are among the most competitive estimators independent of the share of treated. Furthermore, the results suggest that the fat-tail problems observed for IPT, OLS and OLS adjusted radius matching are related to the lack of ‘enough’ control observations, as they are confined to the smaller samples in the scenario with 90% treated. For IPT, however, these problems occur even in cases with 50% treated.

Considering the results over all outcomes and DGPs, the following picture emerges in our view: IPT is attractive for the binary outcome but performs very poorly for the semi-continuous outcome. IPW and kernel regression have a reasonable performance but are in many cases dominated by the parametric and the adjusted radius-matching. Comparing the latter two, it is obvious that the probit and OLS estimators are computationally much more inexpensive than the matching methods (and their standard errors are easier to compute). [48] Moreover, parametric models as well as IPW and IPT do not require the choice of tuning parameters. However, given that the parametric estimators are biased in the case of misspecification and this bias does not disappear with sample size, matching still is preferable.

As mentioned before, our results are somewhat at odds with Frölich (2004) and Busso, DiNardo, and McCrary (2009a, b), as regression-adjusted radius matching, parametric regression, and for the binary outcome also IPT on average outperform any other method including kernel-ridge matching and IPW. The different findings may be due to the fact that the previous studies did not consider all classes and implementations of estimators considered in this paper, in particular not those with the best properties in terms of the RMSE.

[48] With respect to computation time, OLS/probit, DR, and IPT are all very fast, while the kernel regressions and matching are considerably slower. The speed differences increase faster than the sample size. The exact amount of speed differences are very sensitive to how the various tuning parameters are chosen.

Table 5.5: Comparison of the properties of the selected estimators: other features

	Employment								Earnings							
			Kernel		Matching		Probit				Kernel		Matching		OLS	
	IPW	IPT	high	low	logit	pair		DR	IPW	IPT	high	low	OLS	pair		DR
Correctly specified propensity score																
RelRMSE	7	2	7	14	7	61	B	7	7	15	4	18	4	63	B	15
Bias	0.2	0.6	0.7	1.0	0.7	0.1	0.5	0.5	6	21	19	27	8	3	22	5
Std. dev.	3.9	3.7	3.9	3.9	3.9	5.9	3.6	3.9	94	97	88	96	92	145	84	101
Skew.	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	-0.1	-3.0	0.0	-0.1	-0.2	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.0	3.4	3.0	3.4	3.0	3.1	3.1	476	3.1	3.4	8.2	3.7	4.5	7.0
Misspecified propensity score																
RelRMSE	27	B	31	23	7	62	13	15	9	83	13	7	B	39	8	9
Bias	2.7	0.8	2.9	2.4	1.5	2.6	2.1	2.1	66	25	73	62	53	63	70	62
Std. dev.	3.6	3.7	3.6	3.6	3.7	5.3	3.5	3.6	88	194	85	87	87	129	81	93
Skew.	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	-1.9	0.0	-0.1	-0.5	-0.2	-0.1	-0.1
Kurtosis	3.0	3.0	3.0	3.1	3.0	2.9	3.0	3.0	3.1	524	3.1	3.2	13.5	3.2	4.6	5.5
Selection: Normal																
RelRMSE	8	B	7	8	2	49	4	5	4	54	3	5	B	45	1	6
Bias	1.3	0.6	1.6	1.4	0.9	1.2	1.2	1.2	31	15	43	37	26	28	39	31
Std. dev.	3.5	3.6	3.4	3.5	3.6	5.1	3.5	3.5	88	143	82	87	87	129	82	90
Skew.	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	-9.7	0.0	-0.1	-0.6	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.0	3.1	3.0	3.2	3.1	3.0	3.0	936	3.0	3.3	12.3	3.3	3.3	4.0
Selection: Random																
RelRMSE	0.5	3	4	0.1*	7	48	2	2	0.5	31	3	0.1*	11	49	3	5
Bias	0.0	0.4	0.1	0.1	0.4	0.1	0.3	0.2	1	1	2	2	4	2	1	1
Std. dev.	3.1	3.1	3.2	3.1	3.2	4.5	3.1	3.1	70	92	72	70	78	104	72	74
Skew.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-2.3	0.0	0.0	-0.3	-0.1	-0.2	-0.1
Kurtosis	3.0	3.0	3.0	3.0	3.0	3.3	3.0	3.0	3.1	294	3.1	3.1	17.1	3.4	5.7	6.3
Selection: Strong																
RelRMSE	37	B	40	39	12	81	12	22	17	66	19	25	B	57	10	21
Bias	3.1	1.0	3.6	3.7	2.0	2.8	2.5	2.4	76	54	92	95	62	69	98	69
Std. dev.	4.7	4.4	4.5	4.7	4.5	7.1	4.1	4.6	116	203	106	119	103	178	93	128
Skew.	0.1	0.0	0.0	0.0	0.1	0.2	0.1	0.1	-0.2	4.6	0.0	-0.2	-0.1	-0.4	0.0	-0.2
Kurtosis	3.0	3.0	3.0	3.6	3.0	3.0	3.0	3.1	3.2	272	3.1	3.6	3.2	3.8	4.6	8.5
Share of treated: 10%																
RelRMSE	19	0.1*	30	19	18	66	8	14	14	0.2*	26	21	18	66	14	19
Bias	1.1	0.2	1.7	1.3	1.0	1.1	0.8	0.8	24	14	39	36	27	24	38	27
Std. dev.	3.4	3.2	3.5	3.4	3.5	4.9	3.2	3.3	83	77	87	84	85	126	77	85
Skew.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0
Kurtosis	3.0	3.0	3.0	3.0	2.9	3.0	3.0	3.0	3.0	3.1	3.0	3.0	3.0	3.0	3.1	3.1
Share of treated: 50%																
RelRMSE	15	B	15	18	4	59	6	11	11	47	7	12	B	54	6	15
Bias	1.3	0.7	1.3	1.4	1.0	1.3	1.3	1.3	33	21	33	35	26	30	43	31
Std. dev.	3.7	3.6	3.7	3.8	3.6	5.4	3.5	3.7	90	133	88	92	84	134	80	96
Skew.	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.1	-0.1	-7.6	0.0	0.0	-0.1	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.0	3.1	3.0	3.4	3.0	2.9	3.1	972	3.1	3.2	4.5	3.6	3.3	7.5
Share of treated: 90%																
RelRMSE	17	B	15	16	1	58	4	7	8	105	6	12	B	44	3	10
Bias	2.1	1.1	2.5	2.7	1.4	1.8	1.9	1.8	52	36	72	67	39	46	58	43
Std. dev.	4.4	4.3	3.9	4.1	4.2	6.5	4.0	4.2	102	234	84	99	102	152	92	111
Skew.	0.0	-0.1	0.0	-0.2	0.0	0.2	0.0	0.0	-0.1	2.9	-0.1	-0.4	-1.0	-0.4	-0.1	-0.3
Kurtosis	3.1	3.1	3.0	3.6	3.0	2.9	3.2	3.2	3.2	291	3.1	3.7	28.2	3.7	8.0	7.5

Note: RelRMSE: Difference in relative root mean squared error in % compared to best estimator. Bias and standard deviation for employment are given in %. All results based on relative trimming level of 4%. *The best estimator is this estimator without trimming.

However, the previous studies also differ in other respects that may drive the results, e.g. the nature of their (non-empirical) DGPs and the application of trimming rules. It is particularly worth noting that both Frölich (2004) and Busso, DiNardo and McCrary (2009a, b) consider much smaller sample sizes and less rich specifications than we do. It may well be that the relative performance of the estimators is reversed in very small samples. However, as samples with, for example, 100 observations appear to be inappropriate for a sound application of semi-parametric propensity score methods, and are therefore rarely found in empirical applications, we do not examine this case.

6 Conclusion

This paper investigates the finite sample properties of all major classes of propensity-score-based estimators of the average treatment effect on the treated (ATET) that are used in applications. Moreover, within each class of estimators we investigate the performance of the estimators for a variety of possible versions and various values of the tuning parameters. Both features make this study the most comprehensive one in the field so far.

We propose a way to overcome one of the main criticisms of Monte Carlo simulations, namely that of unrealistic, artificially and arbitrarily chosen DGPs. The key feature of our approach is that we base the simulations on real data, and hence real selection problems and dependencies between treatment and outcomes, but still know the true value of the parameter of interest. Moreover, to improve the external validity of our results that, strictly speaking, apply only to labour market evaluations, we vary several features of the DGP's such as the sample size, the magnitude of selection into the treatment, the share of treated observations, and the outcome. As a further contribution, we consider a simple trimming rule not investigated before that is based on identifying control observations whose relative weights

are larger than a particular threshold rather than imposing a fixed threshold value of the propensity score. This rule does not entail asymptotic bias.

Our results suggest that when averaging over all DGPs, trimming observations with a weight larger than 4% reduces the root-mean-squared-error (RMSE) of all estimators substantially. Among the best trimmed estimators of each class, we find that overall bias-adjusted radius matching, parametric regression (probit for the binary and OLS for the semi-continuous outcome), and - for the binary outcome case only - inverse probability tilting (IPT) perform best with respect to the RMSE. However, the parametric estimators may be subject to substantial bias that dominates the RMSE in larger samples than considered in this paper, while radius matching and IPT may be subject to fat-tail behaviour when there are too few control observations.

Bias-adjusted radius matching and, for binary outcomes only, IPT also appear to be the most robust methods when the propensity score is functionally misspecified. Yet, all other estimators (which are among the best within their class of estimators) are within a reasonable distance in terms of the RMSE.

Having understood the performance of the available estimators for covariate adjustment in an (almost) real application situation, future research targeted at identifying appropriate estimators in practice might also address the question of finding reliable inference procedures for these estimators. Furthermore, an Empirical Monte Carlo study has also the important limitation that it may not necessarily be valid in a different environment (although the advantage is that it is valid in at least one relevant environment). We leave it for future research to better understand the general external validity of the results presented in this paper. Future research might also investigate other dimensions of the methods important to empirical work, for example how to determine the common support or how to choose the particular

specification of the propensity score. Finally, the surprising behaviour of IPT in the case of the semi-continuous outcome is worth further investigation.

References

Abadie, A. (2005): "Semiparametric Difference-in-Difference Estimators", *Review of Economic Studies*, 72, 1-19.

Abadie, A., and G. W. Imbens (2002): "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," NBER Technical Working Paper 283.

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica*, 74, 235-267.

Abadie, A., and G. W. Imbens (2008): "On The Failure Of The Bootstrap For Matching Estimators", *Econometrica*, 76, 1537–1557.

Abadie, A., and G. W. Imbens (2009): "Matching on the Estimated Propensity Score", NBER Working Paper 15301.

Angrist, J. D. (1998): "Estimating the Labor Market Effects of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249-288.

Angrist, J. D., and J. Hahn (2004): "When to control for covariates? Panel-Asymptotic Results for Estimates of Treatment Effects", *Review of Economics and Statistics*, 86, 58-72.

Angrist, J. D., and S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricists' Companion*, Princeton, NJ: Princeton University Press.

Augurzky, B., and J. Kluve (2007): "Assessing the Performance of Matching Algorithms when Selection into Treatment is Strong", *Journal of Applied Econometrics*, 22, 533-557.

Bang H., and J. M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models", *Biometrics*, 61, 962–972.

Behncke, S., M. Frölich and M. Lechner (2010a): "Unemployed and their Case Workers: Should they be friends or foes?", *The Journal of the Royal Statistical Society - Series A*, 173, 67-92.

Behncke, S., M. Frölich and M. Lechner (2010b): "A caseworker like me - does the similarity between unemployed and caseworker increase job placements?", forthcoming in *The Economic Journal*.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates", *Quarterly Journal of Economics*, 249-275.

Blundell, R., and M. Costa Dias (2009): "Alternative Approaches to Evaluation in Empirical Microeconomics", *Journal of Human Resources*, 44, 565-640.

Blundell, R., C. Meghir, M. Costa Dias, and J. van Reenen (2004): "Evaluating the Employment Impact of a Mandatory Job Search Program", *Journal of the European Economic Association*, 2, 569-606.

Busso, M., J. DiNardo, and J. McCrary (2009a): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects", forthcoming in the *Journal of Business and Economic Statistics*.

Busso, M., J. DiNardo, and J. McCrary (2009b): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators", IZA discussion paper, 3998.

Caliendo, M., R. Hujer, and S. Thomsen (2006): "Sectoral Heterogeneity in the Employment Effects of Job Creation Schemes in Germany", *Journal of Economics and Statistics*, 226/2, 139-179.

Caliendo, M., R. Hujer, and S. Thomsen (2008a): "The Employment Effects of Job Creation Schemes in Germany - A Microeconometric Evaluation", in: Millimet, D., Smith, J. and Vytlacil, E. (eds.), *Advances in Econometrics, Volume 21: Estimating and Evaluating Treatment Effects in Econometrics*, 383-430.

Caliendo, M., R. Hujer, and S. Thomsen (2008b): "Identifying Effect Heterogeneity to Improve the Efficiency of Job Creation Schemes in Germany", *Applied Economics*, 40, 1101-1122.

Card, D, J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *Economic Journal*, 120, F452–F477.

Chen, X. (2007): "Large Sample Sieve Estimation Of Semi-Nonparametric Models," *Handbook of Econometrics*, Volume 6B, Elsevier B.V. Chapter 76, 5549-5632.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," Cowles Foundation Discussion Paper No. 1644; shorter version published as "Semiparametric Efficiency in GMM Models with Auxiliary Data, *The Annals of Statistics*, 36, 808-843.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects", *Biometrika*, 96, 187–199.

Dehejia, R. H. (2005): "Practical Propensity Score Estimation: a Reply to Smith and Todd", *Journal of Econometrics*, 125, 355-364.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes", *Journal of the American Statistical Association*, 94, 1053-1062.

Dehejia, R. H., and S. Wahba (2002): "Propensity Score- Matching Methods for Nonexperimental Causal Studies", *Review of Economics and Statistics*, 84, 151-161.

Diamond, A., and J. S. Sekhon (2008): "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies", mimeo.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach", *Econometrica*, 64, 1001-1044.

Drake, C. (1993): "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect", *Biometrics*, 49, 1231-1236.

Fan, J. (1992): "Design-adaptive Nonparametric Regression", *Journal of the American Statistical Association*, 87, 998–1004.

Flores, C. A., and O. A. Mitnik (2009): "Evaluating Nonexperimental Estimators for Multiple Treatments: Evidence from Experimental Data", mimeo.

Frölich, M. (2004): "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators", *Review of Economics and Statistics*, 86, 77–90.

Frölich, M. (2005): “Matching estimators and optimal bandwidth choice”, *Statistics and Computing* 15, 197-215.

Frölich, M. (2007a): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139, 35-75.

Frölich, M. (2007b): "Nonparametric regression for binary dependent variables", *Econometrics Journal*, 9, 511-540.

Galdo, J., J. Smith, and D. Black (2008): "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data", *Annales d'Économie et de Statistique*, 91-92, 189-216.

Gerfin, M., and M. Lechner (2002): "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland", *The Economic Journal*, 112, 854-893.

Glynn, A. N., and K. M. Quinn (2010): "An Introduction to the Augmented Inverse Propensity Weighted Estimator", *Political Analysis*, 18:36–56, doi:10.1093/pan/mpp036.

Graham, B. S., C. Pinto, and D. Egel. (2010). "Inverse probability tilting for moment condition models with missing data," forthcoming in the *Review of Economic Studies*.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66, 315-331.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Methods of Moments Estimators", *Econometrica*, 50, 1029-1055.

Hansen, B. (2004): “Full Matching in an Observational Study of Coaching for the SAT,” *Journal of the American Statistical Association*, 99 (467), 609-618.

Hall, P., J. Racine, and Q. Li (2004): "Cross-Validation and the Estimation of Conditional Probability Densities", *Journal of the American Statistical Association*, 99, 1015-1026.

Heckman, J. J. (1976): "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models", *Annals of Economic and Social Measurement*, 5, 475-492.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data", *Econometrica*, 66, 1017-1098.

Heckman, J. J., R. LaLonde, and J. A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs", in: O. Ashenfelter and D. Card (eds.), *Handbook of Labour Economics*, Vol. 3, 1865-2097, Amsterdam: North-Holland.

Hill, J. B. and E. Renault (2010): “Generalized Method of Moments with Tail Trimming,” mimeo, Dept. of Economics, University of North Carolina - Chapel Hill.

Hirano, K., and G. W. Imbens (2001): "Estimation of Causal Effects Using Propensity Score Weighting: An Application of Data on Right Ear Catheterization", *Health Services and Outcomes Research Methodology*, 2, 259–278.

Hirano, K., G.W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica*, 2003, 1161-1189.

Ho, D., K. Imai, G. King, and E. Stuart (2007): "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference", *Political Analysis*, August, 15, 199-236.

Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Population", *Journal of the American Statistical Association*, 47, 663-685.

Huber, P. J., and E. M. Ronchetti (2009), *Robust Statistics*, 2nd edition, Hoboken: Wiley.

Huber, M. (2011): "Testing for covariate balance using quantile regression and resampling methods", *Journal of Applied Statistics*, 38, 2881-2899.

Huber, M., M. Lechner, and C. Wunsch (2011): "Does Leaving Welfare Improve Health? Evidence for Germany", *Health Economics*, 20, 484-504.

Hujer, R., and S. Thomsen (2010): "How Do Employment Effects of Job Creation Schemes Differ with Respect to the Foregoing Unemployment Duration?", *Labour Economics*, 17, 38-51.

Hujer, R., M. Caliendo, and S. Thomsen (2004): "New Evidence on the Effects of Job Creation Schemes in Germany - A Matching Approach with Threefold Heterogeneity", *Research in Economics* 58, 257-302.

Hujer, R., S. Thomsen, and C. Zeiss (2006): "The Effects of Vocational Training Programmes on the Duration of Unemployment in Eastern Germany", *AStA Advances in Advances in Statistical Analysis, 90/2*, 299-322.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review", *Review of Economics and Statistics*, 86, 4-29.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47, 5–86.

Imbens, G. W., W. Newey, and G. Ridder (2006): "Mean-squared-error Calculations for Average Treatment Effects", IRP discussion paper.

Jacob, B. A., J. Ludwig, and J. Smith (2009): "Estimating Neighborhood Effects on Low-Income Youth", mimeo.

Kahn, S., and E. Tamer (2009): "Irregular Identification, Support Conditions, and Inverse Weight Estimation", mimeo.

Kang, J. D., and J. L. Schafer (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data", *Statistical Science*, 22, 523-539.

King, Gary, R. Nielsen, C. Coberley, J. E. Pope, and Aaron Wells (2011): "Comparative Effectiveness of Matching Methods for Causal Inference," mimeo, Harvard University.

Khwaja, A., G. P. M. Salm, and J. G. Trogdon (2010): "A Comparison of Treatment Effects Estimators Using a Structural Model of AMI Treatment Choices and Severity of Illness Information from Hospital Charts," *Journal of Applied Econometrics*, published online, doi: 10.1002/Jae.1181.

Kline, Patrick (2011): "Oaxaca-Blinder as a Reweighting Estimator," *American Economic Review: Papers & Proceedings*, 101 (3), 532–537.

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

Lechner, M. (1999): "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification", *Journal of Business & Economic Statistics*, 17, 74-90.

Lechner, M. (2000): "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany", *The Journal of Human Resources*, 35, 347-375.

Lechner, M. (2008): "A note on the common support problem in applied evaluation studies", *Annales d'Économie et de Statistique*, 91-92, 217-234.

Lechner, M. (2009): "Long-run labour market and health effects of individual sports activities", *The Journal of Health Economics*, 28, 839-854.

Lechner, M. (2010): "The Estimation of Causal Effects by Difference-in-Difference Methods", Discussion paper, Economics Department, University of St. Gallen.

Lechner, M., and C. Wunsch (2009a): "Active Labour Market Policy in East Germany: Waiting for the Economy to Take Off", *Economics of Transition*, 17, 661-702.

Lechner, M., and C. Wunsch (2009b): "Are Training Programs More Effective When Unemployment is High?", *Journal of Labor Economics*, 27, 653-692.

Lechner, M., R. Miquel, and C. Wunsch (2011): "Long-Run Effects of Public Sector Sponsored Training in West Germany", *Journal of the European Economic Association*, 9, 742-784.

Lee, S., and Y-J. Whang (2009): "Nonparametric Tests of Conditional Treatment Effects", Cowles Foundation Discussion Paper 1740.

Lunceford, J., and Davidian, M. (2004): "Stratification and weighting via the propensity score in estimation of causal treatment effects", *Statistics in Medicine*, 23, 2937–2960.

Millimet, D. L., and R. Tchernis (2009): "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies", *Journal of Business & Economic Statistics*, 27, 297-315.

Newey, W. K. (1984): "A Method of Moments Interpretation of Sequential Estimators", *Economics Letters*, 14, 201-206.

Plesca, M., and J. Smith (2007): "Evaluating Multi-Treatment Programs: Theory and Evidence from the U.S. Job Training Partnership Act", *Empirical Economics* 32, 491-528.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed", *Journal of the American Statistical Association*, 89, 846-866.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data", *Journal of the American Statistical Association*, 90, 122-129.

Robins, J. M., S. D. Mark, and W. K. Newey (1992): "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders", *Biometrics*, 48, 479-495.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rosenbaum, P. R., and D. B. Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", *The American Statistician*, 39, 33-38.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66, 688-701.

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

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999): "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models", *Journal of the American Statistical Association*, 94, 1096-1120.

Seifert, B., and T. Gasser (1996): "Finite-Sample Variance of Local Polynomials: Analysis and Solutions", *Journal of American Statistical Association*, 91, 267–275.

Seifert, B., and T. Gasser (2000): "Data Adaptive Ridging in Local Polynomial Regression", *Journal of Computational and Graphical Statistics*, 9, 338–360.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Smith, J., and P. Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?", *Journal of Econometrics*, 125, 305-353.

Stigler, S. M. (1977): "Do Robust Estimators Work with *Real* Data", *Annals of Statistics*, 5, 1055-1098.

Wooldridge, J. M. (2007): "Inverse probability weighted estimation for general missing data problems", *Journal of Econometrics*, 141, 1281–1301.

Wunsch, C., and M. Lechner (2008): "What Did All the Money Do? On the General Ineffectiveness of Recent West German Labour Market Programmes", *Kyklos*, 61, 134-174.

Zhao, Z. (2004): "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metric and a Monte Carlo Study", *The Review of Economics and Statistics*, 86, 91-107.

Zhao, Z. (2006): "Matching Estimators and the Data from the National Supported Work Demonstration Again", IZA Discussion Paper No. 2375.

Zhao, Z. (2008): "Sensitivity of Propensity Score Methods to the Specifications", *Economics Letters*, 98, 309–319.

Appendix A: More details on the estimators

Appendix A.1: Matching

Table A.1 describes the baseline matching protocol of all direct matching estimators.

Table A.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 $p(x)$, or x. 'Closeness' is based on the Mahalanobis distance.
Step C-1	Repeat A-1) and B-1) until no observation with *d*=1 is left.
Step D-1	Compute the maximum distance (*dist*) obtained for any comparison between a member of the reference distribution and matched comparison observations.
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 $R * dist$ to the one chosen in step A-2). 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. 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	D-2) For any potential comparison observation, add the weights obtained in A-2) and B-2).
Step E	Using the weights $w(x_i)$ obtained in D-2), run a weighted linear regression of the outcome variable on the variables used to define the distance (and an intercept).
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 $E(Y^0 \mid D=1)$ 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*. Add the bias to this estimate to get $\overline{E(Y^0 \mid D=1)}$.

Note: In the Monte Carlo study *R* is set to 50%, 150%, and 300%.

Appendix A.2: Kernel-ridge regression matching

Let $m(\rho)$ denote $E[Y \mid D=0, p(X)=\rho]$, the mean outcome in the control population conditional on the propensity score. The kernel matching estimator of the ATET is defined as

$$\hat{\theta}_{\text{ker}nel} = \frac{1}{N_1}\sum_{i=1}^{N} d_i \cdot \left[y_i - \hat{m}(\hat{p}(x_i))\right],$$

where $\hat{m}(\hat{p}(x_i))$ is the estimated conditional mean outcome among controls given the estimated propensity score $\hat{p}(x_i)$. The Seifert and Gasser (1996, 2000) ridge kernel regression estimator for the counterfactual outcome evaluated at $\rho = \hat{p}(x_i)$ is

$$\hat{m}_0(\hat{p}(x_i)) = \frac{A_0(\hat{p}(x_i))}{B_0(\hat{p}(x_i))} + \frac{A_1(\hat{p}(x_i)) \cdot (\hat{p}(x_j) - \overline{p}(x_i))}{B_1(\hat{p}(x_i)) + r \cdot h \mid \hat{p}(x_j) - \overline{p}(x_i) \mid},$$

where

$$A_a(\hat{p}(x_i)) = \sum_{j:d_j=0}^{N} y_j \cdot \left(\hat{p}(x_j) - \overline{p}(x_i)\right)^a \cdot K\left(\frac{\hat{p}(x_j) - \hat{p}(x_i)}{h}\right),$$

$$B_a(\hat{p}(x_i)) = \sum_{j:d_j=0}^{N} \left(\hat{p}(x_j) - \overline{p}(x_i)\right)^a \cdot K\left(\frac{\hat{p}(x_j) - \hat{p}(x_i)}{h}\right),$$

and

$$\overline{p}(x_i) = \sum_{j:d_j=0}^{N} \hat{p}(x_j) \cdot K\left(\frac{\hat{p}(x_j) - \hat{p}(x_i)}{h}\right) \Bigg/ \sum_{j:d_j=0}^{N} K\left(\frac{\hat{p}(x_j) - \hat{p}(x_i)}{h}\right)$$

$K(\cdot)$ denotes the kernel function and h the bandwidth operator that goes to zero as the sample size increases. r is the ridge term ensuring non-zero denominators that should be set to 0.3125 for the Epanechnikov kernel, which we use in the simulations, according to the rule of thumb of Seifert and Gasser (2000). That is, the ridge term is proportional to the bandwidth in finite samples given that the bandwidth is not too large (which is a case not considered by Seifert and Gasser, 2000). It should be zero if either the sample size or the bandwidth approaches infinity.[49]

Concerning the choice of h, we use both the rule of thumb, see Silverman (1986), as well as least squares cross validation, see for instance Hall, Racine, and Li (2004). For the

[49] We thank Markus Frölich for a fruitful discussion on this topic. If the bandwidth goes to zero with an increasing sample size, as it does in Seifert and Gasser (2000), the ridge term vanishes naturally. However, it should also go to zero for a bandwidth going to infinity, otherwise one would incorrectly estimate a global constant instead of a global linear model. Therefore, the ridge term should only be proportional to the bandwidth if the latter is not 'very large' and should be set to zero otherwise. Furthermore, we thank Markus Frölich for providing us with the GAUSS code of the estimator as well as the cross validation procedure.

Epanechnikov kernel, the rule of thumb suggests setting the bandwidth to $2.34 \cdot \sigma \cdot N_0^{-1/5}$, where n is the sample size among the non-treated and σ is the minimum of the standard deviation and the interquartile range divided by 1.349. The cross-validation bandwidth is chosen by

$$h^{CV} = \arg\min_h \sum_{i:d_i=0} [Y_i - \hat{m}_{-i}(p_i)]^2,$$

where $\hat{m}_{-i}(\rho)$ is the estimate of the conditional mean at propensity score ρ with observation i removed from the sample. This procedure chooses the bandwidth such that the expected value of the squared difference between the estimated and true regression function is minimized, where the expectation is taken with respect to the propensity score distribution among the controls. The bandwidth is (asymptotically) optimal for the estimation of the regression function $\hat{m}(\cdot)$, but not necessarily for the kernel matching estimator of the ATET; see also the discussion in Frölich (2005) and Imbens and Wooldridge (2009). Therefore, we consider $3 \cdot h^{CV}$ and $h^{CV}/3$ as additional bandwidths. Against the theoretical intuition which suggests that undersmoothing should dominate, it is the largest bandwidth $3 \cdot h^{CV}$ that works best on average in our simulations. As a final remark, note that we only consider global bandwidth choices as this is standard in empirical applications. Future work might investigate the usefulness of local bandwidth selection and/or weighted cross validation (where the weights refer to the mass of treated observations given a particular propensity score); see, for instance, Galdo, Smith, and Black (2008), which, however, increases computational burden.