DISCUSSION PAPER SERIES

IZA DP No. 5268

How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score

Martin Huber
Michael Lechner
Conny Wunsch

October 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score

Martin Huber
SEW, University of St. Gallen

Michael Lechner
SEW, University of St. Gallen,
ZEW, CEPR, PSI, CESifo, IAB and IZA

Conny Wunsch
SEW, University of St. Gallen,
CESifo and IZA

Discussion Paper No. 5268
October 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5268
October 2010

ABSTRACT

How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score*

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 observable 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 is important for all estimators. The key 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 is considered an important property.

JEL Classification: C21

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

Corresponding author:

Michael Lechner
Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen
Varnbüelstrasse 14
CH-9000 St. Gallen
Switzerland
E-mail: michael.lechner@unisg.ch

* 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.

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 observables.[1] Compared to (non-saturated) parametric regressions, they have the advantage to allow for effect heterogeneity and to include the covariates in a more flexible way without incurring a course-of-dimensionality problem. The latter problem, which is highly relevant due to the usually large number of covariates that should be adjusted for, is avoided by collapsing the covariate information into a single parametric function, the so-called propensity score, which is defined as the probability of being observed in one of two subsamples conditional on the covariates. 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.

Although many of these propensity-score-based methods are not asymptotically efficient (see for example Heckman, Ichimura, and Todd, 1998, and Hahn, 1998),[2] 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

[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] 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.

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 shall address this question and add further insights to it.

Broadly speaking, the popular estimators can be subdivided into five 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), direct matching estimators (Rubin, 1974, Rosenbaum and Rubin, 1983), and kernel matching estimators (Heckman, Ichimura, and Todd, 1998).[3] However, many variants of the estimators exist within each class and several methods are combining the principles underlying these main classes.

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.

[3] There exists 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).

The second strand of the literature provides Monte Carlo evidence. 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).[4] 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. Below, we argue that their work is 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 estimator based on adjusting covariate differences in the following way: First of all, we suggest a different approach of conducting simulations. This new approach is based on 'real' data. Therefore, we call our approach an 'Empirical Monte Carlo Study'. The basic idea is using 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 selection is based on 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 is mirroring specifications used in past applied

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

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.[5]

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 tuning 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. This issue has also been raised by Busso, DiNardo, and McCrary (2009a) to account for common support problems. However, they find that none of the remedies for poor support considered in their paper seems to work in a robust way, particularly in small samples. Therefore, we propose a different, data driven trimming rule that is (i) easy to implement, (ii) identical for all estimators, and (iii) avoids any asymptotic bias. We show that for all estimators considered, including the parametric ones, trimming based on this rule very effectively improves their performance (even when there is no common support problem).

[5] Our results are also robust to arbitrary effect heterogeneity.

Overall, we find that (i) trimming individual observations that have a 'too large' weight is important for all estimators (even without any common support problem); (ii) the choices of the various tuning parameters is important; (iii) simple matching estimators are inefficient and have considerable small sample bias; (iv) no estimator is superior in all designs; (v) particular bias-adjusted radius matching estimators perform best on average, but may have fat tails if the number of controls is not large enough; and finally, (vi) flexible, but simple parametric approaches do almost as well in the smaller samples, because their gain in precision frequently overcompensates for their larger bias which, however, dominates when samples become larger. One conclusion from these findings is that the choice of the broad class of estimators may be less important than using an optimised version.

The plan of the paper is as follows: In the next section we discuss the principles of 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. The main results are presented in Section 4, while the full set of results is given in Appendix C. Section 5 concludes. 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 as well as the Gauss and Stata code for the preferred estimators.[6]

[6] Until user friendly versions of the estimators are made available on the website, readers are invited to send us an email indicating their interest in either the Gauss or Stata versions. We will inform them when the respective versions become available.

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.[7]

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

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

If there are no other (perhaps unobservable) 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*.[8] The assumption required to interpret θ as a causal parameter is

---

[7] 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*.

[8] 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

called either unconfoundedness, conditional independence assumption (CIA) or selection on observables (e.g., Imbens, 2004). The plausibility of CIA depends on the particular empirical problem considered and on the richness of the data at hand. That is, 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, whether θ has a causal interpretation or not, does not matter for this paper. It is important to note that other semiparametric estimators also rely on propensity score based covariate adjustments, like, for example, the instrumental variable estimator proposed by Frölich (2007) and semi-parametric versions of the difference-in-difference estimator (e.g., Abadie, 2005, Blundell, Meghir, Costas Dias, and van Reenen, 2004, Lechner, 2010a).

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, $\hat{w}_i$.

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

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

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.

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,[9] which is the motivation to analyse 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 attains the semi-parametric efficiency bound derived by Hahn (1998) when using the estimated propensity score based on the correct parametric model.[10]

---

[9] 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.

[10] Hirano, Imbens, and Ridder (2003) also 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).

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:

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

The normalization $\sum_{j=1}^{N}\frac{(1-d_j)\cdot\hat{p}(x_j)}{1-\hat{p}(x_j)}$ ensures that the weights add up to one. This estimator directly reweights the non-treated outcomes to control for differences in the propensity scores between treated and non-treated observations. It is the estimator recommended by Busso, DiNardo, and McCrary (2009a).

Although this estimator is attractive from a computational as well as from an asymptotic efficiency point of view, 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, 2010).

2.4 Direct matching

Pair or one-to-one matching is considered to be the prototype of a matching estimator (with replacement)[11]. The pair matching estimator (PM) is defined as:

[11] 'With replacement' means that a control variable can be used many times as match, whereas in estimators 'without replacement' it is used 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

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

$\underline{1}(\cdot)$ 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 even if the misspecified propensity score model is a monotone transformation of the true model (see the simulation results in Drake, 1993, Zhao, 2008, Millimet and Tchernis, 2009, and Huber, 2010, suggesting some robustness of matching to over- and under-fitting 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 do not exist any 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.

---

matching without replacement, many more matching algorithms 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) for some discussion of these issues.

The third class of direct matching estimators considered is the one-to-many calliper matching algorithm as, for example, discussed by Rosenbaum and Rubin (1985) and used by Dehejia and Wahba (1999, 2002). Calliper or radius matching uses all comparison observations within a predefined distance around the propensity score of the respective treated. 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.[12] 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 remove the bias from the asymptotic distribution that may occur when *X* is more than one-dimensional.

An additional suggestion to improve 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 simulation results obtained by Zhao (2004) suggest that this idea works.

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

The estimator proposed by Lechner, Miquel, and Wunsch (2010) and used in several applications by these authors,[13] combines the features of calliper matching with additional predictors and linear or nonlinear regression adjustment. After the first step of distance-weighted calliper matching with predictors, this estimator uses the weights obtained from matching in a weighted linear or non-linear regression in order to remove any bias due to mismatches. The matching protocol of this estimator is shown in Appendix A.

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 $\hat{p}(X)$ as the treated observed:

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

where $\hat{m}(\cdot)$ denotes the nonparametrically estimated conditional expectation function. Heckman, Ichimura, and Todd (1998) are early examples for 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, the latter is of course not feasible 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

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

sample properties. Ridge regression may be considered as 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. 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 (2007). Note that the latter does not include a ridge term, which is not necessary because of the finite support of the expectation of the outcome variable (even under very large coefficients) due to the logit link function.

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 $m(\cdot)$, 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.[14] 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.[15] However, specifying a joint model for the treated and controls that just includes a treatment dummy is

[14] 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. Furthermore, the curse of dimensionality problem is less relevant in parametric regressions.

[15] We also consider 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.

unnecessarily restrictive. I.e., it can lead to large biases 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 as double-robust regression (DR) in the (epidemiologic) literature. DR estimation follows in two steps. First, we run weighted regressions of Y on X in the pool of non-treated individuals, where the weights are proportional to $\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}$. This reweighs the controls according to the distribution of $\hat{p}(X)$ among the treated. 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.[16]

[16] 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).

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 all 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).[17] In that it is reported in the literature that the small sample behaviour of some of the estimators appears to be design dependent, we propose an alternative method that we call *Empirical Monte Carlo Study* (EMCS) from now on.

The idea of the 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 coming from an infinite population and it has to be relevant for the estimators under investiga-

[17] 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 the covariates considered.

tion. That is, it should be a typical data set in a field where the methods under investigation are commonly applied.

Since the estimators we consider are heavily used for the evaluation of active labour market programmes for unemployment 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 step 2 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. 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 feature 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 with regard to 'checking' the properties of estimators based on how capable they are of reproducing the results of an experimental control group (although in this case the true effect is not known but only estimated and contains sampling error), see for example LaLonde (1986), Heckman, Ichimura, Smith, 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, in that EMCS repeatedly draws subsamples from the population, it allows the distribution of the estimators to be recovered. In contrast, a comparison of one (noisy) experimental estimate with one (noisy) alternative estimate will at best give some (noisy) idea of the bias and cannot reveal anything about other aspects of the distribution of the estimators. Secondly, using EMCS it is possible to vary a couple of parameters of the DGP, in particular the selection process, and check how the performance of the estimators changes accordingly.

3.2 The population

In the next subsections we present the details of how 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.[18] 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 information has been matched to the data using the official codes of the 439 German districts. It contains migration and commuting streams, average earnings, unemployment rate, long-term unemployment, welfare dependency rates, urbanisation codes, industry structure and public transport facilities.

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

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, remaining UI claim. Moreover, they cover all spells of participation in the major German labour market programmes from 2000 onward with exact beginning, end 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, profession, 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[19] 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

In that 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 from a sample that covers all entries into unemployment in the period 2000-2003. Then, we exclude unemployment entries in 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 employ-

[19] 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).

ment 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 in order 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 avoid having to model educational choices or (early) retirement decisions. To make our sample 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 (2010) 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 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 employment in that period. This choice has been made to evaluate the estimators' performance with both a variable with only 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.

Table 3.1: Descriptive statistics of the 'population'

Variable	Treated		Control		Standardized difference	Probit estimation of selection equation	
	mean	std.	mean	std.	in %	coef.	std. error
Employed	.63	0.56	.48	0.50	9	-	-
Earnings in EUR	1193	1041	1115	1152	9	-	-
Constant term	-	-	-	-	-	-4.90	.22
Age / 10	3.6	3.5	.84	1.1	8	1.60	.11
… squared / 1000	1.4	1.4	.63	.85	3	2.01	-.13
20 - 25 years old	.21	binary	.41		22	.19	.04
Women	.57	.46	.50	.50	15	-1.16	.24
Not German	.11	binary	.31		16	-.13	.03
Secondary degree	.32	binary	.47		15	.21	.02
University entrance qualification	.29	binary	.45		15	.19	.02
No vocational degree	.18	binary	.39		26	-.07	.03
At least one child in household	.42	binary	.49		22	-.04	.03
Last occupation: Non-skilled worker	.14	binary	.35		13	.07	.03
Last occupation: Salaried worker	.40	binary	.49		29	.33	.03
Last occupation: Part time	.22	binary	.42		12	.36	.05
UI benefits: 0	.33	binary	.47		16	-.14	.02
> 650 EUR per month	.26	binary	.44		7	.13	.03
Last 10 years before UE:							
share employed	.49	.46	.34	.35	8	-.30	.04
share unemployed	.06	.05	.11	.11	1	-.55	.10
share in programme	.01	.01	.04	.03	9	1.12	.25
Last year before UE: share minor empl.	.07	.03	.23	.14	15	-.21	.17
share part time	.16	.11	.33	.29	10	-.22	.05
share out-of-the labour force (OLF)	.28	.37	.40	.44	14	-.30	.04
Entering UE in 2000	.26	binary	.44		13	.29	.02
2001	.29	binary	.46		5	.18	.02
2003	.20	binary	.40		12	.004	.03
Share of population close to big city	.76	.73	.35	.37	6	.09	.02
Health restrictions	.09	binary	.29		13	-.15	.03
Never out of labour force	.14	binary	.34		6	.12	.03
Part time in last 10 years	.35	binary	.48		9	-.12	.03
Never employed	.11	binary	.31		17	-.27	.05
Duration of last employment > 1 year	.41	binary	.49		4	-.13	.02
Average earnings last 10 years when employed / 1000	.59	.52	.41	.40	13	-.09	.04
Women x age / 10	2.1	1.7	1.9	1.9	17	.57	.13
x squared / 1000	.83	.64	.85	.90	15	-.57	.17
x no vocational degree	.09	binary	.28		15	-.23	.04
x at least one child in household	.32	binary	.47		25	.17	.04
x share minor employment last year	.06	.02	.22	.13	16	.71	.18
x share OLF last year	.19	.18	.36	.35	3	.22	.05
x average earnings last 10 y. if empl.	.26	.19	.34	.30	16	-.23	.06
x entering UE in 2003	.10	binary	.30		6	-.14	.04
$x_i\hat{\beta}$	-1.7	.42	-2.1	.42	68	-	-
$\Phi(x_i\hat{\beta})$	.06	.03	.05	.03	59	-	-
Number of obs., Pseudo-R^2 in %	3266		114349			3.6	

Note: 'binary': indicates a binary variable (where the standard deviation can be directly deduced from the mean). $\hat{\beta}$ denotes the estimated probit coefficients and $\Phi(a)$ is the cumulative distribution function of the standard normal distribution evaluate at *a*. *Pseudo-R²* is the so-called Efron's R^2 $(1-\sum_{i=1}^{N}[d_i-\hat{p}(x_i)]\Big/\sum_{i=1}^{N}\left[d_i-\sum_{i=1}^{N}(d_i)/N\right])$.

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).

Table 3.1 also contains the normalized differences between treated and controls as well as the coefficients for the respective covariates in the estimation of the true propensity score to describe selectivity. 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. [20]

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. 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 rarely used in applications) for much smaller samples.[21] 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.[22] 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

[20] Table B.1 in Appendix B.1 shows the results of a probit and tobit 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 also significantly explain the outcomes such that confounding takes place.

[21] Note that the simulations in Busso, DiNardo, and McCrary (2009a,b) are based on sample sizes of 100 and 500, which is of course 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.

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

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 the check whether the estimators attain this asymptotic convergence rate already in the finite samples.

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.[23]

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.[24] The largest share, on the other hand, mimics the situation when the role of treated and con-

[23] The pseudo random number generator used in all simulations is the one implemented in Gauss 9.0.

[24] 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.

trols 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:

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

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.[25] 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.[26]

Note that this simulation routine always ensures common support, at least in expectation, because the treatment probability given the covariates never exceeds 90%. In addition, note that it is not possible to combine the small sample with the extreme shares of participants. This would frequently include the case that the number of covariates exceeded the treated or non-treated observations thus, posing numerical problems on the estimation of the

[25] Note that the simulations are not conditional on *D*, and thus the share of treated is a random number.

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

propensity score. Thus, in the small sample the unconditional treatment probability is 50%, which also makes small sample issues concerning the common support unproblematic.

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 considered*
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 on Table 3.1.

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 outcome of the treated under treatment and under no treatment. The former is always estimated in the same way, namely as a simple average outcome of the treated. We only vary the estimator of the counterfactual non-treatment outcome of the treated. Due to using only true non-treated individuals, by construction the true effect in the EMCS is zero and homogeneous. Therefore, any kind of effect heterogeneity has to be simulated by changing the outcome of some simulated treated but not that of the non-treated because otherwise the stable unit treatment value assumption, which is implicit in our framework, would be violated. Consequently, the relative performance of the estimators remains unchanged as they only differ in how they estimate the counterfactual no treatment outcome, which is by construction unaffected by any kind of simulated effect heterogeneity.

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

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 observations receive a very large weight in all estimators. 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 estimators this observation will receive a weight of one and the remaining non-treated observations a weight of zero. Thus, such estimators have an infinite variance because they are based on the mean of 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. 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.[27] Therefore, we suggest setting all weights to zero if their share of the sum of all weights is larger than *t*%, i.e. $w_{i|d_i=0} = w_i \underline{1}\left[w_i \Big/ \sum_{j=1}^{N} (1-d_j) w_j \leq t\% \right]$. After this step, the remaining weights are normalized again. This correction disappears when 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. However, note that our trimming rule is not concerned with common support and affects only the non-treated while the treated are left untouched.

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 is that support issues are asymptotic problems. 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 to remove 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) and that bias will remain asymptotically. There have been different proposals in the literature on how to tackle the common support problem, but they

[27] When treated observations are removed, the population underlying the definition of the ATET changes. When control observations are removed, we may not be able reweight the controls successfully towards the distribution of the covariates observed for the treated.

all share the feature that they will lead to asymptotic bias,[28] or give up point identification (Lechner, 2010b). In contrast, trimming based on our suggestion vanishes as the sample size increases such that the estimation is asymptotically unbiased.

Khan and Tamer (2009) analyse the problems that may appear for estimators adjusting for covariate differences if identification requires estimation in regions of the covariate space which they call thin-support regions. Such regions 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 as well as the trimming rules. 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. Then, we analyse implementational issues that are specific to the particular classes of

[28] See the excellent discussion of this issue by Busso, DiNardo, and McCrary (2009a). They use four different trimming rules to improve 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. As our trimming rule - in contrast to those rules - concerns only non-treated observations, its performance is of course independent of whether there is effect heterogeneity or not.

estimators considered. Finally, we compare the best estimators across the different classes to come to an overall conclusion.

Before discussing the results, two general remarks are in order. Firstly, most of our conclusions come from analysing the root mean squared error (RMSE) of the estimators. Appendix C contains additional information with respect to the absolute bias and the standard deviation of the estimators, which will sometimes 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.

The second remark concerns the wealth of information produced by the Monte Carlo study. For the employment outcome we have about 5700 data points and for the earnings outcome about 3700 data points for each measure of estimator quality we consider. Thus, we have to summarise this information. In the first parts of this section, 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 class of estimator.

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 (i) on a correctly specified model for the propensity score and (ii) on a functionally misspecified model where all eight interaction terms and the two terms

capturing non-linearities in age are omitted from 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%, 5%, 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.[29] 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.

5.1.2 Inverse probability weighting

The estimator described in Section 2 is directly implemented. It is the version that also 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

[29] The rule is only applied once and not iteratively.

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 calliper size, in terms of the distance between treated and non-treated. Since no well established algorithm exists, we follow Lechner, Miquel and Wunsch (2010) who suggest defining the calliper 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 calliper 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 calliper, the observations within the calliper 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 under 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, as the cross-validation bandwidth is optimal for the regression curve but not for the average of it that enters the ATET, one would expect that some undersmoothing is optimal (although this turns out rarely to 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 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, both 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 truncation at zero, respectively). Therefore, we also use a Tobit model 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. In total we use 6 estimators for employment and 7 estimators for earn-

ings (two versions of OLS, probit or heckit, respectively, and DR estimation based on probit or heckit and OLS, respectively, that weights the regression by $\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}$).[30]

5.2 Results for features that concern all estimators

Table 5.1 (employment) and Table 5.2 (earnings) contain the regression results for the root mean squared error, whereas the results for the bias and the standard deviation are relegated to Appendix C.2 (Tables C.2 and C.3 for employment and Tables C.4 and C.5 for earnings). [31]

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).[32] 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.

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.

---

[30] Since heckit turned out to be very unstable for the smaller samples, DR with OLS was included for earnings as well, although it is a misspecification for the semi-continuous outcome. For the latter reason DR based on OLS was not used for the binary outcome, for which DR with probit works fine.

[31] Furthermore, Tables C.6 and C.7 in Appendix C.3 contain similar results for the subset of estimators that are analysed in detail in Tables 5.3 to 5.5 below. This subset excludes estimators that can be seen as extreme and therefore might be expected to give more reliable results. However, these results seem generally in accordance with the results obtained from the tables in the main part of the text.

[32] Some of estimators based on the parametric models are highly unstable for the earnings outcome under the small and intermediate sample sizes. We do not report the regressions in this case.

5.2.2 Functional misspecification of the propensity score

A misspecification 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.

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		7.4	3.6	1.5	7.1	3.2	1.3	7.1	3.7	2.0	7.1	4.0	1.3
Features of the data generating process													
Selection:	Random	(-1.0)	-1.0	(-0.9)	-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	3.0	3.3	3.1	2.8	2.9	2.5	2.5	2.7	2.4	2.3	3.1	1.9
Share treated:	10%	-	(0.9)	(0.2)	-	1.2	0.6	-	1.5	0.6	-	1.8	0.5
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	3.5	3.5	-	3.2	2.1	-	4.4	2.1	-	4.4	1.7
Features of the estimators													
Misspecified p-score		(-0.1)	(0.4)	1.3	-0.8	(0.1)	1.1	-0.8	-0.3	0.8	-0.3	(-0.5)	0.9
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		-1.3	-0.9	(-0.4)	-0.5	-0.3	(-0.0)	-0.8	-0.9	-0.1	-0.7	-1.3	(-0.1)
Trimming max 4%		-1.5	(-0.9)	(-0.5)	-0.7	-0.3	(-0.1)	-1.0	-0.9	-0.2	-0.9	-1.2	(-0.2)
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 %)		85	76	73	74	81	74	73	59	69	88	34	74
Number of observations		18	54	54	144	432	432	540	1620	1620	108	324	324

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. All coefficients are in %. Coefficients that are not significant at the 5% level (conventional OLS standard errors), appear in brackets.

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 Table C.1 in Appendix C.

Table 5.2: Analysis of features of matching estimators by OLS regression: Earnings (sample sizes)

Variables (all indicators)		IPW			Kernel			Matching			Parametric*)		
Sample Size		300	1200	4800	300	1200	4800	300	1200	4800	300	1200	4800
Constant		191	96	40	171	82	34	180	103	62	171	113	33
Features of the data generating process													
Selection:	Random	-36	-34	-27	-32	-29	-27	-36	-39	-30	-34	(-30)	-25
	Observed	0	0	0	0	0	0	0	0	0	0	0	0
	Strong	83	81	76	65	73	72	63	69	69	82	144	63
Share treated:	10%	-	(20)	(2)	-	31	13	-	39	12	-	(31)	(8)
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	84	59	-	73	55	-	103	50	-	174	42
Features of the estimators													
Misspecified p-score		(-1)	(11)	29	(-3)	9	23	-9	(3)	32	(-13)	(-40)	33
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		-42	-27	(-12)	(-8)	-9	(-2)	-24	-26	(-4)	-29	-89	(-4)
Trimming max 4%		-48	-27	(-14)	-15	-9	(-3)	-32	-28	-6	-36	-85	(-5)
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			
No adjustment								0	0	0			
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 %)		82	79	71	88	85	73	70	61	72	70	23	76
Number of observations		18	54	54	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 brackets.
*): Heckit estimates are very unstable and therefore excluded from the regressions presented in this table.

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

Thus, very few observations are trimmed by this trimming rule, but of course these are those observations with the largest influence on the final estimate. Although only few 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 admissible 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 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 precision loss declines due to the general decrease of the variance with increasing sample size, the bias reduction increases in relative terms, such that both effects almost cancel out. Despite this

feature, the results later on will show that nearest neighbour matching is still dominated by other matching methods.

Considering the calliper size for radius matching, the findings are again in line with our expectations: The smaller the calliper, 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.[33] Finally, using the linear index instead of the propensity score does not have much of an effect at all.

5.3.2 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.

[33] To save computation time the matching estimators without including additional covariates in a Mahalanobis metric have only be 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 based only on estimators computed for all sample sizes.

5.3.3 Parametric models

Among the parametric models, 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 heckit estimator in the earnings regressions despite censoring 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 heckit estimators is rooted in their comparably large variances in the small and medium samples. In particular the heckit-based DR estimators seem to suffer from numerical instabilities when the number of observations is too small as also their 'non-normal' distribution suggests. Therefore, the heckit estimator is not considered in the regressions presented in Table 5.2.[34] Even without heckit, DR does not appear attractive because of its larger variability compared to standard regression (or IPW, see below). Finally, estimating an additional model for the treated as well 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, Tables 5.3 and 5.4 present for a selected group of estimators 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

[34] 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.

and the kurtosis. The latter two are included to see whether there is any important deviation from normality which may cause problems for inference.

To be able to present the results in a concise way, we selected estimators that dominated 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 / 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.[35]

Among the matching estimators, regression-adjusted radius matching (using linear regression for earnings and logit for employment) with additional predictors based on the linear index with a large radius is dominant. 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 take a look at the sensitivity 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.

[35] Obviously, these criteria are arbitrary, but they insure that estimators perform reasonably in a large group of DGP's and specification. The final conclusions are not very sensitive to how the respective groups are formed and which shares are exactly imposed.

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.

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

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
Propensity score correctly specified														
Without trimming														
RelRMSE	39	16	26	16	93	10	28	46	16	35	36	201	62	144
Bias	0.5	1.0	1.5	0.9	0.3	0.9	0.9	10	29	39	23	5	29	9
Std. dev.	5.1	4.1	4.2	4.1	7.1	3.9	4.6	129	93	109	117	178	137	216
Skew.	0.1	0.0	0.0	0.1	0.2	0.0	0.0	-0.4	0.0	-0.2	-0.4	-0.2	-2.8	-2.8
Kurtosis	3.4	3.0	3.7	3.0	3.2	3.0	3.2	4.7	3.1	3.6	7.0	3.4	172	174
Trimming level 6%														
RelRMSE	11	9	16	9	70	2	10	12	7	22	12	73	3	21
Bias	0.3	0.7	1.2	0.8	0.2	0.6	0.5	7	21	30	10	4	23	6
Std. dev.	4.1	3.9	4.0	3.9	6.2	3.7	4.0	99	90	100	98	153	86	108
Skew.	0.0	0.0	0.0	0.1	0.2	0.0	0.0	-0.1	0.0	-0.1	-0.3	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.4	3.0	3.3	3.0	3.0	3.1	3.1	3.5	7.5	3.6	5.2	7.4
Trimming level 4%														
RelRMSE	7	7	14	7	61	B	7	7	4	18	4	63	B	15
Bias	0.2	0.7	1.0	0.7	0.1	0.5	0.5	6	19	27	8	3	22	5
Std. dev.	3.9	3.9	3.9	3.9	5.9	3.6	3.9	94	88	96	92	145	84	101
Skew.	0.0	0.0	0.0	0.0	0.2	0.0	0.0	-0.1	0.0	-0.1	-0.2	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.4	3.0	3.4	3.0	3.1	3.1	3.1	3.4	8.2	3.7	4.5	7.0
Propensity score misspecified														
Without trimming														
RelRMSE	35	20	13	2	62	10	19	26	16	10	9	51	16	22
Bias	3.0	2.8	2.4	1.4	2.9	2.3	2.4	71	76	65	52	68	75	66
Std. dev.	4.3	3.5	3.6	3.8	5.7	3.6	4.0	109	84	88	98	141	88	107
Skew.	0.1	0.0	0.0	0.1	0.1	0.1	0.1	-0.4	0.0	-0.1	-0.6	-0.2	-0.2	-0.1
Kurtosis	3.3	3.0	3.1	3.0	2.9	3.0	3.1	5.4	3.1	3.2	14.7	3.1	7.2	5.9
Trimming level 6%														
RelRMSE	22	24	17	1	55	6	10	13	15	9	3	43	10	13
Bias	2.8	3.0	2.5	1.4	2.7	2.2	2.2	68	75	64	53	65	71	63
Std. dev.	3.7	3.6	3.7	3.7	5.4	3.5	3.7	92	86	89	90	134	83	97
Skew.	0.0	0.0	0.0	0.1	0.1	0.0	0.1	-0.1	0.0	-0.1	-0.5	-0.2	-0.1	-0.1
Kurtosis	3.0	3.0	3.1	3.0	2.9	3.0	3.0	3.1	3.1	3.2	13.9	3.2	4.7	6.2
Trimming level 4%														
RelRMSE	18	22	14	B	51	5	7	9	13	7	B	39	8	9
Bias	2.7	2.9	2.4	1.5	2.6	2.1	2.1	66	73	62	53	63	70	62
Std. dev.	3.6	3.6	3.6	3.7	5.3	3.5	3.6	88	85	87	87	129	81	93
Skew.	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.5	-0.2	-0.1	-0.1
Kurtosis	3.0	3.0	3.0	3.0	2.9	3.0	3.0	3.1	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 is given in %-points. DR: Double robust (weighted) version of estimator.

Trimming is indeed important for the correctly specified as well as the misspecified model. On average, all estimators benefit from trimming in terms of bias, precision, skewness and kurtosis. When moving from no trimming to 6% the gains appear fairly large, while trimming further observations using the 4% cut-off value leads to only small additional gains. All estimators, including the parametric ones, benefit from trimming, particularly in the case of the semi-continuous outcome. As already discussed, most of the gains originate in 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 lead to a better prediction and, thus, an increased likelihood of weights 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. It is also worth noting that the trimming level does not appear to have any relevant impact on the ordering of the respective estimators.

Comparing the estimators to each other shows that they all 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. 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, logit regression adjusted radius matching (for employment) and the OLS regression adjusted radius matching (for earnings) are the best.

Furthermore, note that the distributional properties of the estimators are dependent on the outcome considered. For the binary employment outcome, the best performing 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 somewhat different. Although the same classes of estima-

tors (OLS adjusted radius matching and OLS) are preferred on RMSE grounds, they have fat tails despite the trimming. Since all other estimators show reasonable tail behaviour, they may be preferred (with the exception of pair matching) despite their slightly higher RMSE.

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, Tables 5.4 and 5.5 present different subsets of the results. As trimming improves any method to some extent 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. For the earnings outcome, fat tails are present for radius matching and OLS while the other estimators do not have this problem and are (apart from pair matching) very close in terms of the RMSE. For the largest sample these tail problems disappear and OLS adjusted radius matching dominates all other estimators.

Note that changing the sample sizes in our comparisons goes along with changing other DGP features: 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, since also specifications with incorrectly specified propensity scores are included, they are not expected to be unbiased. Therefore, to study the pure effect of the sample size in settings where the estimator 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, the bias goes to zero for most of the estimators. There are however important exceptions: the bias of OLS for earnings seems to be independent of the sample size, while the bias of the probit for employment disappears. A similar phenomenon

occurs for the kernel matching estimators, in particular when using the large bandwidth, but the level of the bias is small in this case.

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

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
N = 300*														
RelRMSE	2	3	4	1	51	3	7	4	3	6	0.4	53	0.1	15
Bias	1.3	1.3	1.4	0.9	1.1	2.0	2.0	36	33	34	27	30	42	31
Std. dev.	6.2	6.4	6.4	6.4	9.5	6.2	6.5	148	148	152	148	226	140	167
Skew.	0.1	0.1	0.1	0.1	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.3	-0.4	-0.1	-0.1
Kurtosis	3.0	3.1	3.1	3.0	4.3	3.0	2.9	3.1	3.1	3.2	7.5	4.6	3.7	15.9
N = 1200														
RelRMSE	9	10	9	2	52	B	4	3	3	5	1	45	B	7
Bias	1.5	1.7	1.7	1.1	1.3	1.3	1.3	36	44	45	30	32	46	33
Std. dev.	4.4	4.4	4.4	4.5	6.5	4.2	4.3	106	102	105	108	160	98	112
Skew.	0.0	0.0	0.1	0.1	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.6	-0.2	-0.1	-0.2
Kurtosis	3.0	3.0	3.2	3.0	2.9	3.0	3.0	3.1	3.1	3.5	19.6	3.3	6.3	5.9
N = 4800														
RelRMSE	22	27	24	1	56	B	6	15	19	23	B	49	10	12
Bias	1.5	2.0	1.8	1.2	1.5	1.1	1.1	36	51	48	32	35	47	35
Std. dev.	2.3	2.2	2.2	2.2	3.4	2.0	2.3	58	51	58	51	85	47	59
Skew.	0.0	0.0	-0.1	0.0	0.1	0.0	0.0	-0.1	0.0	-0.1	0.0	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.3	3.0	3.0	3.0	3.1	3.1	3.1	3.2	3.2	3.2	3.1	3.4
N = 300 (correctly specified score; 50% treated)**														
RelRMSE	1	2	2	5	62	7	10	2	1	3	6	60	0.1	12
Bias	0.3	0.2	0.2	0.6	0.1	1.4	1.4	6	3	2	10	3	9	1
Std. dev.	5.8	5.8	5.9	5.9	9.3	6.0	6.1	136	135	139	142	214	134	150
Skew.	0.1	0.0	0.1	0.1	0.0	0.0	0.0	-0.1	0.0	-0.1	-0.5	-0.5	-0.1	0.2
Kurtosis	3.0	3.1	3.1	3.0	6.6	2.9	2.9	3.0	3.1	3.2	11.9	6.2	4.3	10.9
N = 1200 (correctly specified score; 50% treated)														
RelRMSE	14	14	25	11	67	B	13	15	9	23	1	77	B	11
Bias	0.2	0.3	0.8	0.7	0.2	0.2	0.1	5	7	15	5	1	18	5
Std. dev.	3.3	3.3	3.5	3.1	4.8	2.9	3.3	81	76	85	71	125	66	78
Skew.	0.1	0.0	0.3	0.1	0.2	0.1	0.1	-0.1	0.0	0.1	0.0	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.5	2.9	3.0	3.0	3.0	3.0	3.0	3.5	2.9	3.3	3.0	3.1
N = 4800 (correctly specified score; 50% treated)														
RelRMSE	22	25	40	15	74	B	19	29	15	48	B	87	17	28
Bias	0.1	0.5	0.2	0.5	0.1	0.1	0.1	1	10	11	3	3	21	8
Std. dev.	1.7	1.7	2.0	1.5	2.5	1.4	1.7	46	40	52	36	67	33	45
Skew.	0.0	0.0	0.1	0.0	0.2	0.0	0.0	-0.1	0.0	0.1	0.0	-0.1	0.0	0.0
Kurtosis	3.0	2.9	2.9	3.0	3.0	2.9	2.9	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 for employment is radius matching (r=150) with the index and additional *X* as matching variables. The best estimator for earnings is radius matching (r=300) with the p-score and additional X as matching variables and regression adjustment. **The best estimator for employment is radius matching (r=150) with the index and additional X as matching variables. The best estimator for earnings is radius matching (r=150) with the p-score and additional X as matching variables.

Secondly, the standard deviation is approximately reduced by half when quadrupling the sample size. It is also interesting to note that while the relative differences in the RMSE of

the estimators are moderate in the smallest sample, they become more pronounced when the sample size increases. Under the larger sample sizes the probit dominates for the employment outcome while regression-adjusted radius matching is in second 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 bias of OLS, which does not decrease in the sample size, is starting to dominate the RMSE, while in the medium sample they perform similarly well (because OLS always has a larger bias but a smaller variance). In the smallest sample both estimators are, as before, fat-tailed. 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 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.

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.[36] Under misspecification, the regression-adjusted radius matching estimators dominate as they have the smallest bias which points to a desirable robustness property.

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. 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

[36] 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.

some bias, is present for the 'normal' selection process. For cases with strong selection, it is again the radius matching estimators as well as the probit and OLS which dominate, with radius matching being least biased, and the parametric estimators being the most precise.

Finally, consider a variation in the percentage of treated. First of all, we observe that independent of the share of the treated, the parametric estimate is either the best or close to being the best. The same holds true (maybe with some reservations for the case of 10% selection for the employment outcome) for regression-adjusted radius matching. When the number of controls is reduced, the differences between the estimators become somewhat more pronounced. The results suggest that the fat-tail problems observed for OLS and OLS adjusted radius matching are related to the lack of control observations, as they are confined to the smaller samples in the scenario with 90% treated.

As mentioned before, our results are somewhat at odds with Frölich (2004) and Busso, DiNardo, and McCrary (2009a,b), as regression-adjusted radius matching (and parametric regression) 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. 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, therefore, are rarely found in empirical applications, we do not examine this case.

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

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
Correctly specified propensity score														
RelRMSE	7	7	14	7	61	B	7	7	4	18	4	63	B	15
Bias	0.2	0.7	1.0	0.7	0.1	0.5	0.5	6	19	27	8	3	22	5
Std. dev.	3.9	3.9	3.9	3.9	5.9	3.6	3.9	94	88	96	92	145	84	101
Skew.	0.0	0.0	0.0	0.0	0.2	0.0	0.0	-0.1	0.0	-0.1	-0.2	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.4	3.0	3.4	3.0	3.1	3.1	3.1	3.4	8.2	3.7	4.5	7.0
Misspecified propensity score														
RelRMSE	18	22	14	B	51	5	7	9	13	7	B	39	8	9
Bias	2.7	2.9	2.4	1.5	2.6	2.1	2.1	66	73	62	53	63	70	62
Std. dev.	3.6	3.6	3.6	3.7	5.3	3.5	3.6	88	85	87	87	129	81	93
Skew.	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.5	-0.2	-0.1	-0.1
Kurtosis	3.0	3.0	3.0	3.0	2.9	3.0	3.0	3.1	3.1	3.2	13.5	3.2	4.6	5.5
Selection: Random														
RelRMSE	0.5	4	0.1*	7	48	2	2	0.5	3	0.1*	11	49	3	5
Bias	0.0	0.1	0.1	0.4	0.1	0.3	0.2	1	2	2	4	2	1	1
Std. dev.	3.1	3.2	3.1	3.2	4.5	3.1	3.1	70	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	0.0	-0.3	-0.1	-0.2	-0.1
Kurtosis	3.0	3.0	3.0	3.0	3.3	3.0	3.0	3.1	3.1	3.1	17.1	3.4	5.7	6.3
Selection: Normal														
RelRMSE	6	5	5	B	46	1	3	4	3	5	B	45	1	6
Bias	1.3	1.6	1.4	0.9	1.2	1.2	1.2	31	43	37	26	28	39	31
Std. dev.	3.5	3.4	3.5	3.6	5.1	3.5	3.5	88	82	87	87	129	82	90
Skew.	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.6	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.1	3.0	3.2	3.1	3.0	3.0	3.0	3.3	12.3	3.3	3.3	4.0
Selection: Strong														
RelRMSE	22	25	24	B	62	0.2	9	17	19	25	B	57	10	21
Bias	3.1	3.6	3.7	2.0	2.8	2.5	2.4	76	92	95	62	69	98	69
Std. dev.	4.7	4.5	4.7	4.5	7.1	4.1	4.6	116	106	119	103	178	93	128
Skew.	0.1	0.0	0.0	0.1	0.2	0.1	0.1	-0.2	0.0	-0.2	-0.1	-0.4	0.0	-0.2
Kurtosis	3.0	3.0	3.6	3.0	3.0	3.0	3.1	3.2	3.1	3.6	3.2	3.8	4.6	8.5
Share of treated: 10%														
RelRMSE	10	20	10	10	53	0.1*	5	0.1	10	6	4	45	0.2*	4
Bias	1.1	1.7	1.3	1.0	1.1	0.8	0.8	24	39	36	27	24	38	27
Std. dev.	3.4	3.5	3.4	3.5	4.9	3.2	3.3	83	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	-0.1	0.0	0.0
Kurtosis	3.0	3.0	3.0	2.9	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.1	3.1
Share of treated: 50%														
RelRMSE	12	12	15	B	53	2	7	11	8	13	B	53	7	13
Bias	1.4	1.4	1.4	0.9	1.3	1.3	1.2	34	34	35	27	32	43	33
Std. dev.	3.4	3.5	3.6	3.4	5.0	3.2	3.4	84	82	86	79	125	75	87
Skew.	0.0	0.1	0.1	0.1	0.1	0.0	0.1	-0.1	0.0	0.0	-0.1	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.1	3.0	3.4	3.0	3.0	3.1	3.1	3.2	4.6	3.5	3.3	5.0
Share of treated: 90%														
RelRMSE	18	16	18	B	56	6	9	8	8	12	B	41	3	7
Bias	2.2	2.6	2.7	1.4	2.0	2.0	1.9	55	74	69	41	50	59	46
Std. dev.	4.1	3.6	3.8	3.9	6.0	3.8	4.1	97	80	95	98	143	87	103
Skew.	0.0	0.0	-0.2	0.0	0.2	0.0	0.0	-0.1	0.0	-0.4	-1.1	-0.3	-0.2	-0.3
Kurtosis	3.1	3.0	3.6	3.1	2.9	3.2	3.2	3.1	3.1	3.5	30.4	3.6	6.6	6.9

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 without trimming.

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 as a function of a variety of tuning parameters. Both features make this study the most comprehensive one in the field.

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 simulation 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, our design allows varying several DGP features 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 discarding control observations whose relative weight are larger than a particular threshold rather than a fixed threshold value of the propensity score. In contrast to many other trimming rules considered in the matching literature, it does not entail asymptotic bias.

Our results suggest that when averaging over all DGPs, trimming 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 and parametric regression (probit for the binary and OLS for the semi-continuous outcome) perform with respect to the RMSE. However, the latter may be subject to substantial bias that dominates the RMSE in larger samples, while the former may be subject to fat-tail behaviour when the control observations are too few.

Bias-adjusted radius matching also appears to be the most robust method 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. Thus, our results do not confirm some of the results of Frölich (2004) and Busso, DiNardo, and McCrary (2009a,b) who conclude that kernel ridge matching and inverse probability weighting perform best, respectively. Although those estimators do not perform badly, they are nowhere near the top in most cases.

The differences in the conclusions 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 bias-adjusted radius matching methods, which performed well. However, the previous studies also differ in other respects that may drive the results, e.g., the features of the DGPs, the implementation of the trimming rules, and the sample sizes considered, the smallest of which (100 observations) appears to be inappropriate for the evaluation of propensity score methods.

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.

References

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

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 (2009): "Matching on the Estimated Propensity Score", NBER Working Paper 15301.

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.

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

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

Galdo, J., J. Smith, and D. Black (2007): "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data", IZA Discussion Paper 3095.

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.

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

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): "Characterisation 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.

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, M. (2010): "Testing for covariate balance using quantile regression and resampling methods", University of St. Gallen, Dept. of Economics Discussion Paper No. 2010-18.

Huber, M., M. Lechner, and C. Wunsch (2010): "Does Leaving Welfare Improve Health? Evidence for Germany", forthcoming in *Health Economics*.

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.

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.

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

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

Lechner, M. (2010b): "A note on the common support problem in applied evaluation studies", *Annales d'Économie et de Statistique*, 91-92, 217-234.

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

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., R. Miquel, and C. Wunsch (2010): "Long-Run Effects of Public Sector Sponsored Training in West Germany", forthcoming in the *Journal of the European Economic Association*.

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.

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.

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 $p(x), \tilde{x}$. '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 from this estimate to get $\widehat{E(Y^0 \mid D=1)}$.

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

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

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.[37]

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

[37] We thank Markus Frölich for a fruitful discussion on this topic. If the bandwidth goes to zero with an increasing sample size as it is the case 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 to set the bandwidth to $2.34 \cdot \sigma \cdot N_0^{-1/5}$, where n is the sample size among 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 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 (2007), which, however, increases computational burden.

Appendix B: Additional details on the Monte Carlo design

Appendix B.1 Probit and Tobit results for the outcome equations

Table B.1: Probit and Tobit results for the outcome equations

Dependent variable:	Employment (Probit)				Earnings (Tobit)			
	Participants		Nonparticipants		Participants		Nonparticipants	
Independent variables	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Constant term	.610	.772	-.356	.107	1171	540	345	102
Age / 10	.099	.407	1.02	.053	130	310	453	55
... squared / 1000	-.442	.510	-1.72	.064	-217	378	-741	79
20 - 25 years old	.135	.118	.119	.021	-9	86	129	14
Woman	-1.59	.818	-.624	.105	-282	699	-213	82
Not German	-.074	.080	-.025	.013	78	61	-19	9
Secondary degree	-.044	.061	.013	.012	16	41	68	8
University entrance qualification	-.227	.066	-.078	.013	331	72	389	10
No vocational degree	-.125	.090	-.116	.014	-170	69	-262	10
At least one child in household	.183	.086	.151	.015	-15	89	25	11
Last occupation: Non-skilled worker	-.085	.094	-.128	.014	-217	64	-151	9
Last occupation: Salaried worker	-.022	.083	-.124	.015	94	59	37	10
Last occupation: Part time	.003	.137	-.036	.025	-87	101	-82	16
UI benefits: 0	.075	.064	.082	.012	70	48	50	8
> 650 EUR per month	.112	.072	.021	.014	89	58	53	9
Last 10 years before UE: share empl.	.146	.120	.062	.022	-152	102	50	15
share unemployed	-1.45	.345	-.541	.053	-285	309	-335	37
share in programme	1.84	.801	.761	.162	-18	593	213	105
Last y. before UE: share in minor empl	.939	.591	.682	.090	-235	365	-264	45
share part-time employed	-.174	.150	-.242	.028	122	131	-44	20
share out-of-the labour force	-.160	.120	-.460	.019	90	111	57	22
Entering UE in 2000	.269	.072	.175	.013	316	61	210	10
2001	.159	.068	.131	.012	139	53	158	9
2003	-.355	.091	-.181	.015	10	110	-143	11
Share of population close to big city	.042	.067	-.107	.012	-12	49	3	9
Health impairments	-.123	.083	-.510	.013	78	84	-149	20
Never out of labour force	-.074	.085	-.024	.015	198	73	-17	11
Part time in last 10 years	.022	.075	.062	.014	38	53	37	9
Never employed	-.846	.137	-1.78	.024	778	283	879	92
Duration of last employment > 1 year	.035	.052	.074	.010	-12	36	39	7
Average earnings last 10 years when employed / 1000	-.090	.126	-.098	.024	1057	99	1466	18
Women x age / 10	1.09	.465	.347	.061	48	427	-36	49
x squared / 1000	-1.34	.617	-.371	.079	-145	540	138	65
x no vocational degree	-.095	.122	-.010	.019	101	100	145	14
x at least one child in household	-.149	.118	-.292	.022	-275	106	-304	17
x share minor employ. last year	-.677	.600	-.393	.096	21	344	-60	45
x share OLF last year	-.202	.151	-.152	.025	-200	115	-100	20
x av. earn in last 10 y. if employed	-.382	.160	-.219	.031	-38	149	-232	24
x entering UE in 2003	-.243	.117	-.029	.020	-4	94	49	14
Selection term in tobit specification					-714	486	-349	71
Number of observations	3266		114349		3266		114349	
(Pseudo) R^2	0.13		0.36		0.36		0.41	

Appendix B.2: Wald tests

Table B.2 Wald tests

	Additional matching variables		Nonlinear and interaction terms	
	Test statistic (df)	p-value in %	Test statistic (df)	p-value in %
	Participation equation(Probit)			
Full sample	27 (2)	0	671 (10)	0
	Outcome equation employment (Probit)			
Participants	4 (2)	11	23 (10)	1
Nonparticipants	53 (2)	0	1944 (10)	0
	Outcome equation earnings (Tobit)			
Participants	114 (2)	0	22 (10)	2
Nonparticipants	6834 (2)	0	755 (10)	0

Note: Test statistic is distributed as $\chi^2(df)$ with df degrees of freedom.

Appendix C: Further results from the simulations

Appendix C.1: Trimming

Table C.1: Number of deleted non-treated observations for different levels of trimming and different DGP's

Data generating processes			Trimming levels						
Magnitude of selection	Share of treated in %	Sample size	4%	5%	6%	7%	8%	9%	10%
Correct specification of the propensity score									
Random	10	1200	-	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	50	300	0.48	0.21	0.10	0.06	0.03	0.02	0.01
		1200	-	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	90	1200	0.40	0.17	0.09	0.05	0.03	0.02	0.01
		4800	-	-	-	-	-	-	-
Observed	10	1200	0.00	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	50	300	1.66	0.94	0.58	0.37	0.26	0.19	0.13
		1200	0.01	0.00	0.00	0.00	0.00	-	-
		4800	-	-	-	-	-	-	-
	90	1200	2.39	1.49	1.00	0.72	0.53	0.40	0.30
		4800	0.14	0.06	0.03	0.02	0.01	0.01	0.00
Strong	10	1200	0.73	0.41	0.25	0.17	0.11	0.08	0.06
		4800	0.02	0.01	0.01	0.00	0.00	0.00	-
	50	300	3.99	2.86	2.17	1.71	1.40	1.16	0.98
		1200	1.45	0.97	0.68	0.51	0.39	0.30	0.24
		4800	0.34	0.21	0.14	0.10	0.08	0.06	0.05
	90	1200	4.30	3.26	2.57	2.08	1.74	1.48	1.27
		4800	2.34	1.69	1.31	1.04	0.86	0.74	0.62
Functional misspecification of the propensity score									
Random	10	1200	-	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	50	300	0.21	0.09	0.04	0.02	0.01	0.01	0.01
		1200	-	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	90	1200	0.20	0.08	0.04	0.02	0.01	0.01	0.00
		4800	-	-	-	-	-	-	-
Observed	10	1200	-	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	50	300	0.92	0.47	0.26	0.16	0.10	0.06	0.04
		1200	0.00	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	90	1200	1.57	0.90	0.57	0.38	0.26	1.57	0.90
		4800	0.03	0.01	0.01	0.00	0.00	0.00	0.00
Strong	10	1200	0.15	0.06	0.03	0.02	0.01	0.15	0.06
		4800	-	-	-	-	-	-	-
	50	300	2.93	1.96	1.38	1.03	0.78	0.61	0.48
		1200	0.49	0.29	0.19	0.13	0.09	0.07	0.05
		4800	0.06	0.03	0.02	0.01	0.00	0.00	0.00
	90	1200	3.44	2.44	1.82	1.42	1.15	0.95	0.79
		4800	1.26	0.85	0.62	0.48	0.39	0.31	0.27

Note: See also note on Table 3.1. '-': no observations are removed. '0.00': average number of observations removed < 0.005.

Appendix C.2: Absolute bias and standard deviation of estimators

Table C.2: Features of the matching estimators by OLS regression for the employment outcome - absolute bias

Variables (all indicators)		IPW			Kernel			Matching			Parametric		
Sample Size		300	1200	4800	300	1200	4800	300	1200	4800	300	1200	4800
Constant		5.8	2.8	1.1	5.6	2.5	1.0	5.6	3.0	1.8	5.5	2.4	1.1
Features of the data generating process													
Selection:	Random	(-0.8)	-0.8	(-0.8)	-0.6	-0.6	-0.8	-0.7	-0.7	-0.7	-0.8	-0.7	-0.6
	Observed	0	0	0	0	0	0	0	0	0	0	0	0
	Strong	2.4	2.7	2.8	2.2	2.5	2.4	2.1	2.3	2.1	1.8	1.9	1.7
Share treated:	10%	-	(0.7)	(0.1)	-	0.9	0.4	-	1.2	0.4	-	1.4	0.3
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	2.8	1.8	-	2.7	1.8	-	3.2	1.6	-	2.8	1.3
Features of the estimators													
Misspecified p-score		(-0.1)	(0.4)	1.4	-0.6	(0.2)	1.1	-0.5	-0.2	0.8	(-0.2)	(0.1)	1.0
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		-1.0	(-0.7)	(-0.3)	-0.4	-0.3	(-0.0)	-0.6	-0.7	-0.1	-0.5	-0.5	(-0.1)
Trimming max 4%		-1.1	(-0.7)	(-0.4)	-0.6	-0.2	(-0.1)	-0.7	-0.6	-0.2	-0.6	-0.4	(-0.1)
Bandwidth: Low					(0.2)	(0.1)	(-0.0)						
Cross validation					0	0	0						
High					-0.5	(-0.1)	(0.0)						
Rule of thumb					(0.2)	(0.1)	(0.0)						
Local logit					0.7	0.2	(-0.1)						
Nearest neighbour								0.6	(0.2)	(-0.1)			
Radius matching:	Radius low							0.4	(-0.1)	-0.9			
	medium							0	0	0			
	large							-0.4	-0.4	(0.1)			
No adjustment								0	0	0			
Regression adjustment								0.4	(-0.1)	-0.9			
Logit adjustment								-0.4	-0.4	(0.1)			
PScore instead of linear index								(0.1)	(0.1)	(0.1)			
Regression for treated											(-0.1)	0.3	(0.0)
Robust											0.4	(0.2)	(-0.0)
Probit											(0.1)	(-0.0)	-0.2
Statistics													
R^2 (in %)		77	70	67	76	67	65	62	57	63	75	60	65
Number of observations		12	36	36	96	288	288	576	1728	1152	96	288	288

Note: Dependent variable: Bias. 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 brackets.

Table C.3: Features of the matching estimators by OLS regression for the employment outcome - standard deviation

Variables (all indicators)		IPW			Kernel			Matching			Parametric		
Sample Size		300	1200	4800	300	1200	4800	300	1200	4800	300	1200	4800
Constant		7.5	3.9	1.8	7.1	3.3	1.6	6.6	2.9	1.2	7.1	4.2	1.3
Features of the data generating process													
Selection:	Random	(-0.8)	(-0.7)	(-0.4)	-0.6	-0.5	-0.3	-0.8	-0.6	-0.3	-0.8	(-0.7)	-0.3
	Observed	0	0	0	0	0	0	0	0	0	0	0	0
	Strong	2.4	2.2	1.6	2.4	1.9	1.0	2.2	2.0	1.2	1.7	2.2	0.9
Share treated:	10%	-	1.1	(0.5)	-	1.2	0.6	-	1.7	0.9	-	1.8	0.8
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	3.2	2.0	-	2.6	1.3	-	4.2	1.8	-	4.0	1.5
Features of the estimators													
Misspecified p-score		(-0.8)	(-0.6)	(-0.3)	-1.2	-0.8	-0.4	-1.0	-0.9	-0.3	-0.8	-1.1	-0.2
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		-1.3	-0.9	(-0.5)	-0.5	(-0.2)	(-0.0)	-0.7	-0.8	-0.1	-0.7	-1.2	(-0.1)
Trimming max 4%		-1.4	-0.9	(-0.6)	-0.7	(-0.1)	(-0.1)	-0.9	-0.8	-0.2	-0.8	-1.1	-0.2
Bandwidth: Low					(0.2)	(0.2)	(0.1)						
Cross validation					0	0	0						
High					-0.6	-0.3	-0.1						
Rule of thumb					(0.3)	(0.1)	(0.1)						
Local logit					0.9	0.4	(0.1)						
Nearest neighbour								3.1	2.3	1.1			
Radius matching:	Radius low							0.9	0.5	0.2			
	medium							0	0	0			
	large							(-0.2)	(-0.1)	(-0.0)			
No adjustment								0	0	0			
Regression adjustment								1.2	1.3	0.2			
Logit adjustment								-0.9	-1.1	-0.1			
PScore instead of linear index								(0.1)	(0.1)	(-0.0)			
Regression for treated											(0.1)	(0.3)	(0.1)
Robust											0.3	(-0.6)	0.4
Probit											(-0.1)	(-0.8)	(0.1)
Statistics													
R^2 (in %)		79	71	63	68	72	72	71	56	62	82	27	70
Number of observations		18	54	54	144	432	432	540	1620	1620	108	324	324

Note: Dependent variable: Standard deviation. 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 brackets.

Table C.4: Features of the matching estimators by OLS regression for the earnings outcome - absolute bias

Variables (all indicators)		IPW			Kernel			Matching			Parametric*)		
Sample Size		300	1200	4800	300	1200	4800	300	1200	4800	300	1200	4800
Constant		147	73	29	134	63	24	141	86	57	129	63	27
Features of the data generating process													
Selection:	Random	-28	-27	-23	-25	-24	-24	-28	-30	-27	-25	-25	-22
	Observed	0	0	0	0	0	0	0	0	0	0	0	0
	Strong	65	67	66	51	65	68	54	64	64	63	62	57
Share treated:	10%	-	(15)	(-0)	-	24	(10)	-	29	7	-	20	(4)
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	68	48	-	63	50	-	70	42	-	74	35
Features of the estimators													
Misspecified p-score		(3)	(13)	31	(-1)	11	25	(-1)	8	33	(-3)	10	33
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		-29	-19	(-8)	(-6)	-9	(-2)	-17	-18	(-4)	-19	-18	(-3)
Trimming max 4%		-34	-20	(-10)	-11	-9	(-3)	-23	-19	-6	-24	-17	(-4)
Bandwidth: Low					(-6)	(-1)	(4)						
Cross validation					0	0	0						
High					-9	(-3)	(2)						
Rule of thumb					(0)	(0)	(1)						
Nearest neighbour								36	12	-18			
Radius matching:	Radius low							11	(2)	(-4)			
	medium							0	0	0			
	large							(1)	(4)	8			
No adjustment								0	0	0			
Regression adjustment								-7	-18	-37			
PScore instead of linear index								(2)	(0)	(-1)			
Regression for treated											(-0)	(0)	(0)
Robust											23	10	(-1)
Statistics													
R^2 (in %)		76	68	65	78	67	66	68	65	68	30	16	67
Number of observations		12	36	36	48	144	144	384	1152	768	96	288	288

Dependent variable: Bias. 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 brackets.

*): Heckit estimates are very unstable and therefore excluded from the regressions presented in this table.

Table C.5: Features of the matching estimators by OLS regression for the earnings outcome - standard deviation

Variables (all indicators)		IPW			Kernel			Matching			Parametric*)		
Sample Size		300	1200	4800	300	1200	4800	300	1200	4800	300	1200	4800
Constant		196	102	49	174	88	44	154	69	29	172	113	30
Features of the data generating process													
Selection:	Random	(-31)	-26	(-15)	-28	-19	-11	-26	-25	-10	-29	(-22)	-9
	Observed	0	0	0	0	0	0	0	0	0	0	0	0
	Strong	66	55	42	51	38	26	48	40	28	63	117	28
Share treated:	10%	-	26	(10)	-	33	14	-	46	22	-	(37)	15
	50%	0	0	0	0	0	0	0	0	0	0	0	0
	90%	-	73	49	-	48	29	-	96	38	-	171	32
Features of the estimators													
Misspecified p-score		(-18)	(-14)	(-8)	-17	-13	-9	-23	-19	-6	-29	-64	(-3)
No trimming		0	0	0	0	0	0	0	0	0	0	0	0
Trimming max 6%		-44	-28	(-14)	(-9)	(-4)	(-0)	-19	-21	(-2)	-29	-89	(-3)
Trimming max 4%		-50	-28	(-16)	-15	(-4)	(-1)	-26	-23	-4	-36	-85	(-4)
Bandwidth: Low					(-7)	(-2)	(1)						
Cross validation					0	0	0						
High					-13	-8	-8						
Rule of thumb					(2)	(-0)	(-0)						
Nearest neighbour								81	63	28			
Radius matching:	Radius low							21	12	5			
	medium							0	0	0			
	large							(-7)	(-4)	(-2)			
No adjustment								0	0	0			
Regression adjustment								27	29	5			
PScore instead of linear index								(5)	(3)	(0)			
Regression for treated											(-0)	(1)	(0)
Robust											42	(35)	14
Statistics													
R^2 (in %)		75	73	62	82	83	75	66	56	61	65	24	69
Number of observations		18	54	54	72	216	216	360	1080	1080	72	216	216

Dependent variable: Standard deviation. 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 brackets.

*): Heckit estimates are very unstable and therefore excluded from the regressions presented in this table.