UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

WORKING PAPERS SES

The finite sample performance of inference methods for propensity score matching and weighting estimators

Hugo Bodory,
Lorenzo Camponovo,
Martin Huber and
Michael Lechner

N. 466
I.2016

Faculté des sciences économiques et sociales
Wirtschafts- und sozialwissenschaftliche Fakultät

The finite sample performance of inference methods for propensity score matching and weighting estimators

Hugo Bodory*, Lorenzo Camponovo*, Martin Huber+, Michael Lechner*,**

* University of St. Gallen, Dept. of Economics

+ University of Fribourg, Dept. of Economics

Abstract: This paper investigates the finite sample properties of a range of inference methods for propensity score-based matching and weighting estimators frequently applied to evaluate the average treatment effect on the treated. We analyse both asymptotic approximations and bootstrap methods for computing variances and confidence intervals in our simulation design, which is based on large scale labor market data from Germany and varies w.r.t. treatment selectivity, effect heterogeneity, the share of treated, and the sample size. The results suggest that in general, the bootstrap procedures dominate the asymptotic ones in terms of size and power for both matching and weighting estimators. Furthermore, the results are qualitatively quite robust across the various simulation features.

Keywords: Inference, variance estimation, treatment effects, matching, inverse probability weighting.
JEL classification: C21

Addresses for correspondence: Hugo Bodory, University of St. Gallen, Varnbüelstrasse 14, CH-9000 St. Gallen, hugo.bodory@unisg.ch; Lorenzo Camponovo, University of St. Gallen, Bodanstrasse 6, CH-9000 St. Gallen, lorenzo.camponovo@unisg.ch; Martin Huber, University of Fribourg, Bd. de Pérolles 90, CH-1700 Fribourg, martin.huber@unifr.ch; Michael Lechner, University of St. Gallen, Varnbüelstrasse 14, CH-9000 St. Gallen, michael.lechner@unisg.ch. **: Michael Lechner is also affiliated with CEPR and PSI, London, CESIfo, Munich, IAB, Nuremberg, and IZA, Bonn.

1 Introduction

A large body of studies in empirical economics, political sciences, sociology, epidemiology, and other fields is devoted to the evaluation of the effect of some (binary) treatment (or intervention) under a 'selection-on-observables' or 'conditional independence' assumption, see for instance Imbens (2004) and Imbens and Wooldridge (2009). Researchers applying treatment effect estimators typically aim to assess the average causal effect of the intervention (e.g. assignment to a training program or a medical treatment) on some outcome variable (e.g. employment, earnings, or health), by controlling for differences in observed characteristics across treated and non-treated subsamples.[1] While some treatment effect estimators directly control for the observed covariates, most of them are based on conditioning on the treatment propensity score instead, i.e. the conditional probability to receive the treatment given the covariates, in order to avoid the 'curse of dimensionality' related to high dimensional covariates. This includes propensity score matching (see for instance Rosenbaum and Rubin (1985), Heckman, Ichimura, and Todd (1998), and Dehejia and Wahba (1999)) and inverse probability weighting (henceforth IPW, Horvitz and Thompson (1952) and Hirano, Imbens, and Ridder (2003)), which belong to the most popular methods among practitioners.[2]

Virtually all empirical implementations are semiparametric in the sense that parametric propensity score estimation (using logit or probit) is combined with nonparametric treatment effect estimation (using matching or weighting). To provide empiricists with some guidance about which approach may work well in practice, a growing number of simulation studies has investigated and compared the finite sample behavior of various point estimators, see Frölich (2004), Zhao (2004), Lunceford and Davidian (2004), Busso, DiNardo, and McCrary (2014), Huber, Lechner, and Wunsch (2013), and Frölich, Huber, and Wiesenfarth (2014).[3] While the behavior of the point estimators therefore appears to be comparably well studied, there exists, to the best of our knowledge, no comparably thorough simulation study on the performance of variance estimators in the context of treatment effect evaluation.[4] This is surprising, as the

[1] In general, treatment effect estimators may be applied to any issue in which the mean of some outcome across two subsamples should be evaluated net of differences due to observed variables, including wage gap decompositions (see for instance Frölich (2007) and Ñopo (2008)).

[2] However, there exist further classes of treatment effect estimators, see for instance Robins, Mark, and Newey (1992), Robins, Rotnitzky, and Zhao (1995), and Robins and Rotnitzky (1995), and Rothe and Firpo (2013) for so-called doubly robust estimators. Furthermore, the choice is ever increasing, see for instance Graham, Pinto, and Egel (2012), Hainmueller (2012), and Imai and Ratkovic (2014) for recent empirical likelihood and weighting approaches.

[3] As the studies differ in terms of model design, treatment selectivity, and comprehensiveness of estimators investigated, their cumulated results do not yield an unanimous ranking of estimators. They nevertheless give important insights on the robustness of various methods to problems like insufficient propensity score overlap across treatment states and on the effectiveness of trimming influential observations.

[4] Pingel (2015), for instance, focusses on the impact of tuning parameters on the accuracy of the variance

accuracy of inference appears equally important as the accuracy of point estimation.

This paper is the first one to provide a comprehensive simulation study on various variance estimators of point estimators of the average treatment effect on the treated (ATET) and therefore fills an important gap in the literature on the finite sample behavior of treatment effect methods.[5] To this end, we focus on four ATET estimators: IPW, which was competitive in several simulation designs of Busso, DiNardo, and McCrary (2014), the prototypical propensity score pair matching estimator, and radius matching with and without linear bias adjustment (see Abadie and Imbens (2011)) as suggested in Lechner, Miquel, and Wunsch (2011) (which was the best performing estimator in Huber, Lechner, and Wunsch (2013)). Using the same trimming rule as Huber, Lechner, and Wunsch (2013), we discard observations with (too) large weights in ATET estimation in order to tackle potential common support problems. Our choice of IPW and matching is predominantly motivated by the popularity of these estimators in practice, but in the case of matching also by the theoretical finding of Abadie and Imbens (2008) suggesting that standard bootstrap inference is invalid for 'non-smooth' implementations of the estimator (such as pair matching) when there are continuous covariates. As the latter result is widely ignored by practitioners (who frequently apply the bootstrap in matching estimation), one interesting question is whether the theoretical inconsistency of the bootstrap entails biases that are large enough to be practically relevant.

In the light of the unsatisfactory result that the standard bootstrap is inconsistent for some matching algorithms, recent studies propose modified bootstrap procedures that are consistent even for non-smooth (pair or one-to-many) matching estimators with continuous covariates. For instance, Otsu and Rai (2015) introduce and prove the validity of a weighted bootstrap algorithm for particular classes of pair matching estimators that, however, do not include propensity score matching. Furthermore, Bodory, Camponovo, Huber, and Lechner (2016) generalize the approach of Otsu and Rai (2015) by introducing a wild bootstrap procedure that can also be applied to propensity score matching estimators. Unlike the standard bootstrap, this wild bootstrap algorithm does not construct bootstrap samples by randomly selecting with replacement from the original sample. Instead, it constructs wild bootstrap approximations based on the result of Abadie and Imbens (2012a) that matching estimators can be expressed as a sum of martingale processes. This novel approach is also included in our simulation study.

We investigate the finite sample performance of the following variance estimators: two-step

estimator of Abadie and Imbens (2012b), but does not compare several classes of variance estimators.

[5] It would have been interesting to also include estimators of the average treatment effect (ATE) in our analysis. However, due to almost prohibitive computation time, including the ATE would have forced us to investigate fewer variance estimators, which we would have considered a larger sacrifice than focussing on the ATET, the most frequently estimated causal parameter in the program evaluation literature.

GMM-based variance estimation (for IPW), approximations of the variances based on the weights nontreated receive in ATET estimation as in Lechner (2002) (for IPW and matching), and the variance formula of Abadie and Imbens (2006), which is based on the propensity score rather than estimation weights (for pair matching). As the latter two methods treat the propensity scores as fixed, they are (for matching only) also implemented with a variance correction that accounts for the estimation of the propensity score as suggested in Abadie and Imbens (2012b). Furthermore, we consider various implementations of both the standard bootstrap (considered for IPW and matching) and the wild bootstrap (considered for pair matching only): (i) bootstrapping the ATET estimates to compute confidence intervals and p-values based on either the asymptotic distribution of the t-statistic or on the quantiles of the effects (percentile method), and (ii) bootstrapping the (asymptotically pivotal) t-statistic and conducting inference based on its quantiles. For the latter approach we also consider kernel smoothing of the bootstrap distribution of the t-statistics as suggested by Racine and MacKinnon (2007) to improve accuracy of inference when the number of bootstrap replications is low.

In the spirit of Huber, Lechner, and Wunsch (2013) and Lechner and Wunsch (2013) (see also Frölich, Huber, and Wiesenfarth (2014)), we use an‘Empirical Monte Carlo Study’ (EMCS) approach to base our simulation design as much as possible on empirical data. Specifically, we use German labour market data for the evaluation of labor market programs to realistically simulate ‘placebo treatments’ among the non-treated, where the remaining non-treated without placebo treatment permit estimating the (known) non-treatment outcome of the ‘placebo-treated’. To this end, the treatment selection process is estimated from the data and the empirical relation between the outcome and the covariates is retained, rather than relying on an arbitrarily chosen model for the data generating process. We vary several empirically relevant design features in our simulations, namely the sample size, selection into treatment, share of treated, and effect heterogeneity.

The simulation results suggest that inference methods which are based on asymptotic approximations that ignore the estimation of the propensity score tend to be conservative, while accounting for propensity score estimation entails excessive size for some procedures applicable to matching estimators. GMM-based variance estimation of IPW is rather conservative, too, even though it accounts for the estimation of the propensity score. A further finding is that in general, the empirical size of the bootstrap procedures is more accurate than that of the asymptotic ones. For matching, the methods based on bootstrapping t-statistics which account for propensity score estimation generally come closest to the nominal size. For pair matching, the inconsistency of the standard bootstrap seems to have little practical relevance for most methods when accounting for propensity score estimation. Nevertheless, the wild bootstrap entails a more accurate size for several inference estimators than standard bootstrapping, in particular when propensity score estimation is ignored. Concerning power, none of the methods have severe lack-of-power issues, not even the conservative ones. Again, the bootstrap procedures frequently dominate the asymptotic approximations or are at least comparably powerful. Finally, the size and power properties of the different inference procedures are rather stable across the different simulation features like the distribution of the outcome variable, sample size, share of treated, and treatment selection.

The remainder of this paper is organized as follows. Section 2 introduces the ATET and the point estimators (IPW, pair matching, radius matching) and a trimming procedure to deal with problems of common support. Section 3 presents the variance estimators based on asymptotic approximations or various bootstrap implementations. Section 4 discusses our labor market data and the simulation design. Section 5 presents the results for various features of the simulations. Section 6 concludes.

2 Point estimation

We subsequently discuss the (identification of the) parameter of interest (ATET) and present the point estimators (IPW, matching) as well as the trimming rule for ensuring common support.

2.1 Identification of the ATET

Let D denote the binary treatment indicator (e.g. training participation), Y the outcome (e.g. earnings in some follow up period), and X a vector of observed covariates. Furthermore, let $Y(1)$, $Y(0)$ denote the potential outcomes under hypothetical treatment assignment 1 and 0, see Rubin (1974). The average treatment effect on the treated (ATET), denoted by θ, is defined as

$$\theta = E[Y(1) - Y(0)|D = 1], \tag{1}$$

and is identified under two conditions.[6] First, the so-called ‘selection on observables’ or ‘conditional independence’ assumption (CIA) (see for instance Imbens (2004) and Imbens and Wooldridge (2009)) has to be satisfied:

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

[6] In addition, the ‘Stable Unit Treatment Value Assumption’ (SUTVA) needs to hold, see for instance (Rubin 1990).

where '$\perp$' stands for statistical independence. This rules out the existence of (further) confounders that jointly influence the treatment and the potential outcome under non-treatment conditional on X. Second, it must hold that the conditional probability to receive the treatment given X, the so-called propensity score, is smaller than one:

$$\Pr(D = 1|X) < 1, \tag{3}$$

otherwise for (at least) some of the treated units, there exist no untreated units that are comparable in terms X. For ease of notation, let henceforth $p(X) = \Pr(D = 1|X)$.

Under (2) and (3), the ATET is identified by

$$\theta = E(Y|D = 1) - E[E(Y|D = 0, X)|D = 1]. \tag{4}$$

Note that rather than conditioning on X directly as in (4), it follows from Rosenbaum and Rubin (1983) that one may control for the propensity score, $p(X)$ instead, because it possesses the so-called 'balancing property'. That is, conditioning on the one-dimensional $p(X)$ equalizes the distribution of the (possibly high dimensional) covariates X across D, such that the ATET is also identified by

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

2.2 Estimation

As among others discussed in Smith and Todd (2005), a general representation of all treatment effect estimators adjusting for covariate differences is

$$\hat{\theta} = \frac{1}{n_1}\sum_{i=1}^{n} D_i \hat{W}_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - D_i)\hat{W}_i Y_i. \tag{6}$$

n denotes the size of an i.i.d. sample of realizations of $\{Y_i, D_i, X_i\}$ with any observation $i \in 1, ..., n$. $n_1 = \sum_{i=1}^{n} D_i$ is the size of the treated subsample, $n_0 = n - n_1$, and $\hat{W}_i$ are weights that may depend on $\hat{p}(X_i)$, an estimate of the propensity score $p(X_i)$. We specify the latter as a probit model. In our simulations, four different point estimators out of this general class of estimators are included: inverse probability weighting (IPW; an idea going back to Horvitz and Thompson (1952)), pair matching, and radius matching with and without bias correction.

ATET estimation based on IPW reweighs non-treated outcomes such that the distribution of the propensity score among the treated is matched, see Hirano, Imbens, and Ridder (2003) for a more detailed discussion. We consider the following normalized IPW estimator in our simulations, which performed well in several simulation designs considered in Busso, DiNardo, and McCrary (2014):

$$\hat{\theta}_{\text{IPW}} = \frac{1}{n_1} \sum_{i=1}^{n} D_i Y_i - \sum_{i=1}^{n} (1 - D_i) Y_i \left\{ \frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}}{\sum_{j=1}^{n} \frac{(1-D_j)\hat{p}(X_j)}{1-\hat{p}(X_j)}} \right\}. \tag{7}$$

The normalization $\sum_{j=1}^{n} \frac{(1-D_j)\hat{p}_j}{1-\hat{p}_j}$ makes the weights sum up to one, see Imbens (2004) for further discussion. It is easy to see that (7) corresponds to (6) when setting $\hat{W}_i$ in the latter to $D_i + (1 - D_i) n_0 \left\{ \frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}}{\sum_{j=1}^{n} \frac{(1-D_j)\hat{p}(X_j)}{1-\hat{p}(X_j)}} \right\}$. IPW possesses the desirable property that it can attain the semiparametric efficiency bound derived by Hahn (1998), if the propensity score is estimated nonparametrically (while this is generally not the case for parametric propensity scores). Furthermore, it is computationally inexpensive and easy to implement. However, IPW also has an important drawback: if the common support assumption (3) is close to being violated, estimation may be unstable and the variance may explode in finite samples, see Frölich (2004) and Khan and Tamer (2010).

Propensity score matching is based on assigning (matching) to each treated observation one or more non-treated units with comparable propensity scores to estimate the ATET by the average difference in the outcomes of the treated and the (appropriately weighted) non-treated matches. All matching estimators have the following general form:

$$\hat{\theta}_{\text{match}} = \frac{1}{n_1} \sum_{i:D_i=1} \left(Y_i - \sum_{j:D_j=0} \varpi_{i,j} Y_j \right), \tag{8}$$

where $\varpi_{i,j}$ is the weight of the outcome of non-treated observation j when matched to a treated unit i. Pair (or one-to-one) matching with replacement,[7] see for instance Rubin (1973), matches to each treated observation exactly the non-treated observation with the most similar propensity score. This implies the following weights in (8):

$$\varpi_{i,j} = \mathbb{I} \left\{ |\hat{p}(X_j) - \hat{p}(X_i)| = \min_{l:D_l=0} |\hat{p}(X_l) - \hat{p}(X_i)| \right\}, \tag{9}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function which is one if its argument is true and zero otherwise.

[7] 'With replacement' means that a non-treated observation may by matched several times, whereas estimation 'without replacement' requires that it is matched at most once. The latter approach is only feasible when there are substantially more non-treated than treated observations. It is not frequently applied in econometrics.

Therefore, all weights are zero except for that observation j that has the smallest distance to i in terms of the estimated propensity score and receives a weight of one. Because only one non-treated observation is matched to each treated unit irrespective of the sample size and the potential availability of several 'good' matches with similar propensity scores, pair matching is not efficient. On the other hand, it is likely more robust to propensity score misspecification than IPW (in particular if the misspecified propensity score model is only a monotone transformation of the true model), see for instance Zhao (2008), Millimet and Tchernis (2009), Waernbaum (2012), and Huber, Lechner, and Wunsch (2013).

Radius matching (see for instance Rosenbaum and Rubin (1985) and Dehejia and Wahba (1999)) uses *all* non-treated observations with propensity scores within a predefined radius around that of the treated reference unit, which trades off some bias in order to increase efficiency. It is expected to work particularly well if several good potential matches are available. In the simulations, we consider the radius matching algorithm of Lechner, Miquel, and Wunsch (2011), which performed well in Huber, Lechner, and Wunsch (2013). The estimator combines distance-weighted radius matching (i.e. non-treated units within the radius are weighted proportionally to the inverse of their distance to the treated observation) with an OLS regression adjustment for bias correction (see Rubin (1979) and Abadie and Imbens (2011)) to remove small and large sample bias due to mismatches. See Huber, Lechner, and Steinmayr (2014) for a detailed description of the (algorithm of the) estimator. As in Lechner, Miquel, and Wunsch (2011), the radius size in our simulations is defined as a function of the distribution of distances between treated and matched non-treated observations in pair matching. Namely, it is set to either 1.5 or 3 times the maximum pair matching distance. Note that we include radius matching both with and without bias correction in our simulations. All in all, this entails six estimators: IPW, pair matching, and radius matching with and without bias adjustment, each with two different radius sizes.[8]

2.3 Trimming

A practically relevant issue of treatment effect methods is thin or lacking common support (or overlap) in the propensity score across treatment states, which may compromise estimation due to a non-comparability of treated and non-treated observations, see the discussion in Imbens (2004), Imbens and Wooldridge (2009), and Lechner and Strittmatter (2014). If specific propensity score values among the treated are either very rare (thin common support) or absent (lack of common support) among the non-treated, as it may occur in particular close to the boundary

[8] Although it would be interesting to extend our analysis to other estimators as well, computation costs become prohibitive.

of 1, non-treated units with such or similar values receive a large weight $\hat{W}_i$. In the case of thin common support, these observations could dominate the estimator of the ATET which may entail a possible explosion of the variance. In the case of lacking common support, this even introduces asymptotic bias by giving a large weight to non-treated observations that are not comparable to the treated in terms of the propensity score.

Huber, Lechner, and Wunsch (2013) suggest using a trimming procedure first discussed in Imbens (2004), which is asymptotically unbiased if common support holds asymptotically.[9] It is based on setting the weights of those non-treated observations to zero whose relative share of all weights in (6) exceeds a particular threshold value in % (denoted by t):

$$\hat{W}_{i|D_i=0} = \hat{W}_i \mathbb{I}\left\{ \frac{\hat{W}_i}{\sum_{j=1}^{n}(1-D_j)\hat{W}_j} \leq t\% \right\} \tag{10}$$

As in Huber, Lechner, and Wunsch (2013), we trim observations based on the weights of normalized IPW, see (7), irrespective of the point estimator considered. In order to not create an unbalanced sample by trimming the non-treated observations only, any treated with propensity scores larger than the largest value among the remaining non-treated are discarded, too (if such observations exist). Strictly speaking, this (in finite samples) changes the target parameter due to discarding extreme support areas, but ensures common support prior to estimation. As also considered in Huber, Lechner, and Wunsch (2013), we set $t = 4\%$. Note that among the variance estimators discussed in Section 3, only the bootstrap approaches of Sections 3.4 and 3.5 account for the stochastic nature of trimming, while the other procedures outlined in Sections 3.1, 3.2, and 3.3 treat trimming as fixed.

3 Inference

This section presents the inference methods considered in the simulations for the IPW and matching estimators. As in (6), we subsequently denote by $\hat{\theta}$ a general ATET estimator, indicating that the discussion refers to any of the methods, while adding a subscript (like '$_{\text{IPW}}$') implies that the attention is restricted to a particular method.

For IPW, the following variance estimators are investigated: asymptotic variance approximation based on GMM (Section 3.1), variance estimation conditional on the weights in the estimation of the counterfactuals (Section 3.3), bootstrapping the ATET estimates to perform

[9] Other proposals sugested in the literature include Heckman, Ichimura, Smith, and Todd (1998), Dehejia and Wahba (1999), Ho, Imai, King, and Stuart (2007), and Crump, Hotz, Imbens, and Mitnik (2009). However, they all introduce asymptotic bias.

inference based on either the asymptotic distribution of the t-statistic or on the quantiles of the effects (Section 3.4), and bootstrapping the t-statistic, which is computed using either the analytic variance expressions of Sections 3.1 or 3.3, to perform inference based on its quantiles (Section 3.4). For the latter approach we also consider kernel smoothing of the bootstrap distribution of the t-statistics as suggested by Racine and MacKinnon (2007) to improve accuracy of inference when the number of bootstrap replications is low. For pair matching, the asymptotic variance formula of Abadie and Imbens (2006) as well as the propensity score-adjusted version of Abadie and Imbens (2012b) (Section 3.2), variance estimation conditional on matching weights (Section 3.3), and bootstrapping the ATET or the t-statistics with and without kernel smoothing (Section 3.4) are considered. In addition, we also investigate the wild bootstrap procedure introduced in Bodory, Camponovo, Huber, and Lechner (2016), see Section 3.5. For any (standard or wild) bootstrap procedure based on the t-statistic, the latter is computed using the analytic variance expressions of Sections 3.2 or 3.3 and again, the procedures are assessed with and without kernel smoothing. For radius matching with and without bias adjustment, we assess inference based on variance estimation conditional on matching weights (Section 3.3), and on bootstrapping the ATET or the t-statistic (Section 3.4), where the latter is obtained using the analytic expressions in Section 3.3 and implemented with and without kernel smoothing.

3.1 GMM-based asymptotic approximation of the IPW variance

To derive the asymptotic approximation for the variance of IPW based on GMM, we first rewrite (7) as follows:

$$\hat{\theta}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \omega_i(D_i, X_i, \hat{\beta})Y_i, \tag{11}$$

where the weights ω_i for the outcomes Y_i depend on the individual treatment state D_i, covariates X_i and the maximum likelihood estimate $\hat{\beta}$ of the parameter vector of the probit model for the propensity score in the following way:

$$\begin{aligned}
w_i &= n\tilde{w}_i(D_i, X_i, \hat{\beta}),\\
\tilde{w}_i(D_i, X_i, \hat{\beta}) &= D_i\tilde{w}_i(1, X_i, \hat{\beta}) - (1-D_i)\tilde{w}_i(0, X_i, \hat{\beta}),\\
\tilde{w}_i(1, X_i, \hat{\beta}) &= \frac{1}{n_1}, \quad \tilde{w}_i(0, X_i, \hat{\beta}) = \frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}}{\sum_{j=1}^{n}\frac{\hat{p}(X_j)}{1-\hat{p}(X_j)}}.
\end{aligned}$$

Note that by the probit specification of the propensity score, $\hat{p}(X_i) = \Phi(X_i\hat{\beta})$ with Φ denoting the cumulative distribution function (c.d.f.) of the standard normal distribution. Following Newey

(1984), the estimator in (11) can be considered as a two step (or sequential) GMM estimator. In the first step, the score functions of the propensity score model leads to the following $P+1$ moment conditions, where P is the dimension of X:

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i, \hat{\beta}) = 0,$$

where g is the score function, i.e. the first derivative of the log-likelihood of the probit model. In the second step, the estimation of the ATET yields a further moment condition:

$$\frac{1}{n}\sum_{i=1}^{n} h(Y_i, X_i, \hat{\beta}, \hat{\theta}_{\text{IPW}}) = 0,$$

with the moment function $h(Y_i, X_i, \beta, \theta) = \theta - w_i(X_i, \beta)Y_i$ being the difference between the true ATET and the weighted outcomes. If these conditions hold, the resulting GMM estimator is consistent and asymptotically normal under standard regularity conditions.[10]

Using the results of Newey (1984), the asymptotic variance of $\hat{\theta}_{\text{IPW}}$, denoted by $asV\left[\sqrt{n}\hat{\theta}_{\text{IPW}}\right]$, is given by the following expression:

$$\begin{aligned} asV\left[\sqrt{n}\hat{\theta}_{\text{IPW}}\right] &= n^2 Var\left[\tilde{w}_i Y_i\right] \\ &= H_{\theta_{\text{IPW}}}^{-1}(V_{hh} + H_\beta G_\beta^{-1} V_{gg} G_\beta^{-1\prime} H_\beta{}' - H_\beta G_\beta^{-1} V_{gh} - V_{hg} G_\beta^{-1} H_\beta{}') H_{\theta_{\text{IPW}}}^{-1}{}'. \end{aligned}$$

This variance formula shows that $asV\left[\sqrt{n}\hat{\theta}_{\text{IPW}}\right]$ can be expressed as the variance of the weighted outcomes adjusted by terms that depend on the two sets of moment conditions. The components are:

$$H_{\theta_{\text{IPW}}} = E[\partial h(.)/\partial\theta_{\text{IPW}}] = 1, \quad V_{hh} = E[h(.)^2] = Var[n\tilde{w}_i Y_i], \quad H_\beta(d=1) = E[\partial h(.)/\partial\beta] = 0,$$

$$H_\beta(d=0) = E[\partial h(.)/\partial\beta] = E\left[n\frac{\frac{X_i\phi_i}{(1-p(X_i))^2}\sum_{i=1}^n \frac{p(X_i)}{1-p(X_i)} - \frac{p(X_i)}{1-p(X_i)}\sum_{i=1}^n \frac{X_i\phi_i}{(1-p(X_i))^2}}{\left(\sum_{i=1}^n \frac{p(X_i)}{1-p(X_i)}\right)^2} Y_i\right],$$

$$G_\beta = E[\partial g(.)/\partial\beta], \quad V_{gg} = E[g(.)g(.)'], \quad V_{gh} = E[g(.)h(.)], \quad V_{hg} = V_{gh}'.$$

The functions $p(X_i) = \Phi(X_i\beta)$ and $\phi_i = \phi(X_i\beta)$ denote the c.d.f. and the probability density

[10] In particular, the data must be generated from stationary and ergodic processes, the moment functions and the respective derivatives must exist and must be measurable and continuous, the parameters must be finite and not at the boundary of the parameter space, and the derivatives of the moment conditions w.r.t. the parameters must have full rank. Furthermore, the sample moments must converge to their population counterparts with decreasing variances and to uniquely identified values of the unknown parameters.

function (p.d.f.) of the standard normal distribution, respectively, evaluated at $X_i\beta$. The variance of $\hat{\theta}_{\text{IPW}}$ can be consistently estimated by replacing β and θ by their estimates $\hat{\beta}$ and $\hat{\theta}_{\text{IPW}}$ everywhere.[11]

3.2 Asymptotic variance approximations of Abadie and Imbens

Abadie and Imbens (2006) derive the large sample variance of pair and one-to-many matching estimators when matching directly on control variables, based on a decomposition of the total variance into the expectation of the conditional variance and the variance of the conditional expectation given the matching variables. To review their results, we introduce some further notation: let K_i denote the overall number of times a (non-treated) unit i is used as match for any treated observation and $\sigma^2(p(X_i), D_i) = V(Y_i|p(X_i), D_i)$ the conditional variance of the outcome given the (true) propensity score and the treatment. Assuming that the true propensity score is known (rather than estimated), the variance of the pair matching estimator, denoted by $V(\hat{\theta}_{\text{pm, true ps}})$, is given by

$$V(\hat{\theta}_{\text{pm, true ps}}) = \frac{1}{n_1}\left\{E\left[(\theta(X_i)-\theta)^2\,|D_i=1\right] + E\left[\frac{1}{n_1}\sum_{i=1}^{n}(D_i-(1-D_i)K_i)^2\,\sigma^2(p(X_i),D_i)\right]\right\}.(12)$$

Furthermore, let $\hat{\sigma}^2(p(X_i), D_i) = V(Y_i|p(X_i), D_i)$ denote an an asymptotically unbiased estimator of $\sigma^2(p(X_i), D_i) = V(Y_i|p(X_i), D_i)$. Abadie and Imbens (2006) show that $V(\hat{\theta}_{\text{pm, true ps}})$ can be consistently estimated by

$$\hat{V}(\hat{\theta}_{\text{pm, true ps}}) = \frac{n}{n_1^2}\sum_{i=1}^{n} D_i\left(Y_i - \sum_{j:D_j=0}\varpi_{i,j}Y_j - \hat{\theta}_{\text{pm}}\right)^2 + \frac{n}{n_1^2}\sum_{i=1}^{n}(1-D_i)K_i(K_i-1)\hat{\sigma}^2(p(X_i),D_i),(13)$$

where $\varpi_{i,j}$ is defined in (9). In applications, the true propensity score is usually unknown and needs to be estimated, for instance based on the probit model $\hat{p}(X_i) = \Phi(X_i\hat{\beta})$, implying that $\hat{\sigma}^2(p(X_i), D_i)$ in (13) is in fact $\hat{\sigma}^2(\hat{p}(X_i), D_i)$. As this affects the large sample distribution of matching estimators, the variance is in this case different to (12), a fact frequently ignored among practitioners. We therefore consider estimator (13) for pair matching inference in our simulations, to investigate whether its inconsistency is practically relevant.[12] For the estimation

[11] These results are a special case of the variance estimator proposed by Lechner (2009) for the dynamic treatment model when only the first period is considered.

[12] For the average treatment effect (ATE), the asymptotic variance of matching on the known propensity score is at least as large as that of matching on the estimated propensity score, see Abadie and Imbens (2012b). Therefore, the variance estimator of Abadie and Imbens (2006) is conservative. An analogous result applies to IPW, see Hirano, Imbens, and Ridder (2003). For the ATET, however, the ordering of the variances is ambiguous

of $\sigma^2(p(X_i), D_i)$, we use pair matching on the propensity score within the same treatment group as outlined in Abadie and Imbens (2006), which is unbiased (but not consistent):

$$\hat{\sigma}^2(\hat{p}(X_i), D_i) = \left[Y_i - \sum_{j:D_j=D_i} \mathbb{I}\left\{ |\hat{p}(X_j) - \hat{p}(X_i)| = \min_{l:D_l=0} |\hat{p}(X_l) - \hat{p}(X_i)| \right\} Y_j \right]^2 \Big/ 2. \tag{14}$$

In a different paper, Abadie and Imbens (2012b) propose a correction to (12) such that uncertainty w.r.t. propensity score estimation is accounted for in the variance, now denoted by $V(\hat{\theta}_{\text{pm, est. ps}})$. We therefore also consider corrected variance estimators for all matching procedures with inference either relying on Abadie and Imbens (2006) (pair matching), or the variance estimator proposed in Section 3.3 (pair matching and radius matching with and without adjustment). Introducing additional notation, let $\mu(X_i, D_i) = E[Y_i|X_i, D_i]$ and $\mu(p(X_i), D_i) = E[Y_i|p(X_i), D_i]$ denote the conditional means of the outcome given X_i, D_i and $p(X_i)$, D_i, respectively, and $cov(X_i, \mu(X_i, D_i)|p(X_i))$ the covariance between X_i and $\mu(X_i, D_i)$ conditional on $p(X_i)$. Abadie and Imbens (2012b) show that

$$V(\hat{\theta}_{\text{pm, est. ps}}) = V(\hat{\theta}_{\text{pm, true ps}}) - c' I^{-1} c + \frac{\partial \theta}{\partial \beta}' I^{-1} \frac{\partial \theta}{\partial \beta}, \tag{15}$$

with the Fisher information matrix $I = -G_\beta$ and

$$\begin{aligned} c &= \frac{1}{E[p(X)]} E[X\phi(X\beta)(\mu(p(X),1) - \mu(p(X),0) - \theta)] \\ &+ \frac{1}{E[p(X)]} E\left[\left(cov(X, \mu(X,1)|p(X)) + \frac{p(X)}{1-p(X)} cov(X, \mu(X,0)|p(X)) \right) \phi(X\beta) \right], \\ \frac{\partial \theta}{\partial \beta} &= \frac{1}{E[p(X)]} E[X\phi(X\beta)(\mu(X,1) - \mu(X,0) - \theta)]. \end{aligned}$$

$cov(X, \mu(X, D)$ (which can be shown to equal $cov(X, Y|p(X), D)$), $\mu(p(X), D)$, and $\mu(X, D)$, which enter the correction terms in (15), may be estimated by pair matching within or across treatment groups, as we use do in our simulations, see Abadie and Imbens (2012b) for further details. Note that the adjustment term may increase or decrease the variance estimate of the ATET.[13]

and data-dependent.

[13] In some of the simulation draws (in particular when the sample size is small), it occurs that the estimated correction terms are larger than the uncorrected variance. In these cases, the correction is omitted.

3.3 Variance approximation based on weights

Under i.i.d. sampling, the asymptotic variance of the ATET estimator corresponds to the sum of the variances of the estimators of the treated population's mean potential outcomes under treatment and non-treatment, denoted by $\hat{E}[Y_i(1)|D_i = 1]$ and $\hat{E}[Y_i(0)|D_i = 1]$, respectively (ignoring any correlation that may occur due to the estimation of the propensity score):

$$V(\hat{\theta}) = V\left\{\hat{E}[Y_i(1)|D_i = 1]\right\} + V\left\{\hat{E}[Y_i(0)|D_i = 1]\right\}.$$

V and $\hat{V}$ denote the variance and its estimate throughout our discussion. As $E[Y(1)|D = 1] = E(Y|D = 1)$, it follows that $\hat{E}[Y_i(1)|D_i = 1] = \frac{1}{n_1}\sum_{i:D_i=1}^{n_1} Y_i$ such that the standard variance estimator for means of random variables can be applied:

$$\hat{V}\left\{\hat{E}[Y_i(1)|D_i = 1]\right\} = \frac{1}{n_1(n_1-1)}\sum_{i=1}^{n} D_i\left(Y_i - \frac{1}{n_1}\sum_{i=1}^{n} D_iY_i\right)^2.$$

Concerning the variance of the treated population's estimated mean potential outcome under non-treatment, first note that the estimated mean potential outcome under non-treatment of the treated can be expressed as a weighted sum of non-treated outcomes, with the (normalized) non-treated weights ($\tilde{W}_i$) summing up to one: $\hat{E}[Y_i(0)|D_i = 1] = \sum_{i=1}^{n}(1-D_i)Y_i\tilde{W}_i$. For instance, for the IPW estimator (7) $\tilde{W}_i = \left\{\frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}}{\sum_{j=1}^{n}\frac{(1-D_j)\hat{p}(X_j)}{1-\hat{p}(X_j)}}\right\}$. One simple approximation to the variance $V\left\{\hat{E}[Y_i(0)|D_i = 1]\right\}$ is the unconditional variance of $Y_i\tilde{W}_i$:

$$\hat{V}\left\{\hat{E}[Y_i(0)|D_i = 1]\right\} = \frac{1}{n_0-1}\sum_{i=1}^{n}(1-D_i)\left(Y_i\tilde{W}_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)Y_i\tilde{W}_i\right)^2. \quad (16)$$

This assumes homoscedasticity in $\tilde{W}_i$. To allow for heteroscedasticity in the weights when estimating the variance, we consider the following variance decomposition into the expectation of the conditional variance and the variance of the conditional expectation given the weights:

$$\begin{aligned} V\left\{\hat{E}[Y_i(0)|D_i = 1]\right\} &= V\left(\sum_{i=1}^{n}(1-D_i)Y_i\tilde{W}_i\right) \\ &= \underbrace{E\left\{V\left[\sum_{i=1}^{n}(1-D_i)Y_i\tilde{W}_i\middle|\tilde{W}_i\right]\right\}}_{A} + \underbrace{V\left\{E\left[\sum_{i=1}^{n}(1-D_i)Y_i\tilde{W}_i\middle|\tilde{W}_i\right]\right\}}_{B}. \quad (17)\end{aligned}$$

Note that

$$
\begin{aligned}
A &= E\left\{\sum_{i=1}^{n}(1-D_i)\tilde{W}_i^2\sigma^2(\tilde{W}_i, D_i=0)\right\}, && (18)\\
B &= V\left\{\sum_{i=1}^{n}(1-D_i)\tilde{W}_i E[Y_i|\tilde{W}_i]\right\}, && (19)
\end{aligned}
$$

with $\sigma^2(\tilde{W}_i, 0) = V(Y|\tilde{W}_i, D_i = 0)$ being the conditional variance of the outcome given the weight among the non-treated. Under the assumption that $\tilde{W}_i E[Y_i|\tilde{W}_i]$ is uncorrelated across i,[14] the variance of the sum equals n_0 times the variance:

$$
V\left\{\sum_{i=1}^{n}(1-D_i)\tilde{W}_i E[Y_i|\tilde{W}_i]\right\} = n_0 V\left\{\tilde{W}_i\mu(\tilde{W}_i, D_i=0),\right\}, \tag{20}
$$

where $\mu(\tilde{W}_i, 0) = E[Y_i|\tilde{W}_i, D_i = 0]$ is the conditional mean of the outcome given the weight among the non-treated. Basing variance estimation on the decomposition in (17) therefore requires estimates of $\mu(\tilde{W}_i, 0) = E[Y_i|\tilde{W}_i, D_i = 0]$ and $\sigma^2(\tilde{W}_i, 0) = E[(Y_i - \mu(\tilde{W}_i, D_i = 0))^2|\tilde{W}_i, D_i = 0]$, which we denote by $\hat{\mu}(\tilde{W}_i, 0)$ and $\hat{\sigma}^2(\tilde{W}_i, 0) = E[(Y_i - \hat{\mu}(\tilde{W}_i, D_i = 0))^2|\tilde{W}_i, D_i = 0]$. To estimate either parameter, we apply a particular one-to-many (nearest neighbor) matching algorithm, which computes the conditional mean and variance of some reference observation using a set of closest units in terms of weight $\tilde{W}_i$ that are in the same treatment state ($D_i = 0$).

Specifically, let $\mathcal{S}_M(i)$ denote the set of M matches for reference unit i among the units with the same treatment for an odd integer $M \geq 3$. The set includes (i) unit i itself, (ii) the $(M-1)/2$ nearest neighbors (in terms of weights) with a weight smaller or equal to $\tilde{W}_i$, and (iii) the $(M-1)/2$ nearest neighbors with a weight larger than $\tilde{W}_i$:

$$
\begin{aligned}
\mathcal{S}_M(i) &= \left\{ j = 1, \ldots, n : D_j = D_i, \left(\sum_{k:D_k=D_i, \tilde{W}_i - \tilde{W}_k \geq 0} \mathbb{I}\{\tilde{W}_i - \tilde{W}_k \leq \tilde{W}_i - \tilde{W}_j\} \right) \leq (M+1)/2 \right\}\\
&\cup \left\{ j = 1, \ldots, n : D_j = D_i, \left(\sum_{l:D_l=D_i, \tilde{W}_l - \tilde{W}_i > 0} \mathbb{I}\{\tilde{W}_l - \tilde{W}_i \leq \tilde{W}_j - \tilde{W}_i\} \right) \leq (M-1)/2 \right\}.
\end{aligned} \tag{21}
$$

Note, however, that the window of M matches becomes necessarily asymmetric for observations

[14] Due to i.i.d. sampling, non-correlation across i is satisfied if (the coefficients of) the propensity scores (which ultimately determine the weighting function) are non-stochastic, which holds asymptotically. However, in finite samples, units may be correlated through the estimation of the coefficients of the propensity score.

at the upper and lower boundaries of the weights. For instance, for the largest $\tilde{W}_i$, the set $\mathcal{S}_M(i)$ includes (i) unit i itself and (ii) the $(M-1)$ nearest neighbors with a weight smaller or equal to $\tilde{W}_i$. The conditional mean and variance are then estimated by

$$\begin{aligned}\hat{\mu}(\tilde{W}_i, D_i) &= \frac{1}{M}\sum_{i\in\mathcal{S}_M(i)}^{M} Y_i, \\ \hat{\sigma}^2(\tilde{W}_i, D_i) &= \frac{1}{M}\sum_{i\in\mathcal{S}_M(i)}^{M}\Big(Y_i - \hat{\mu}(\tilde{W}_i, D_i)\Big)^2.\end{aligned}$$

We may therefore estimate the variance components (18) and (20), respectively, by

$$\hat{A} = \sum_{i=1}^{n}(1-D_i)\tilde{W}_i^2\hat{\sigma}^2(\tilde{W}_i, 0), \tag{22}$$

$$\hat{B} = \frac{n_0}{n_0-1}\sum_{i=1}^{n}(1-D_i)\left(\tilde{W}_i\hat{\mu}(\tilde{W}_i,0) - \frac{1}{n_0}\sum_{i=1}^{n}(1-D_i)\tilde{W}_i\hat{\mu}(\tilde{W}_i,0)\right)^2. \tag{23}$$

We consider variance estimation based on (i) the unconditional variance formula in (16), (ii) the decomposition based approach with $\hat{V}\left\{\hat{E}[Y_i(0)|D_i=1]\right\} = \hat{A}+\hat{B}$, and finally, based on $\hat{A}$ only. Concerning the estimation of the the conditional means and variances required in approaches (i) and (ii), we use the following sample size-dependent rule for choosing the number of nearest neighbors: $M = 2\text{round}(\kappa\sqrt{n})+1$, 'round$(\cdot)$' means that the argument is rounded to the closest integer and κ gauges the number of neighbors. In the simulations, we consider 3 choices for κ: 0.2, 0.8, 3.2.

Even though these variance estimators may be reasonable approximations, there are also several caveats. First of all, the unconditional variance estimator (i) is only valid under homoscedasticity. In contrast, estimators (ii) and (iii) allow for heteroscedasticity w.r.t. $\tilde{W}_i$. Furthermore, when using matching with bias correction, note that while the appropriate bias corrected weights enter the variance formulae, uncertainty related to the estimation of bias correction is not accounted for. Finally, any of the variance estimators omits the fact that the propensity scores entering the weights is itself an estimate rather than known, which in general affects the distribution of the ATET estimators. To tackle the latter issue, we therefore apply the variance correction of Abadie and Imbens (2012b) to (i), (ii), and (iii) to also account for propensity score estimation, see the discussion in Section 3.2.

3.4 Standard bootstrap

Inference in treatment effect estimation is frequently based on the (standard) nonparametric bootstrap (see Efron (1979) or Horowitz (2001), among others). This holds true even for applications of matching, in spite of the result of Abadie and Imbens (2008) that the nonparametric bootstrap is inconsistent for pair or one-to-many matching (with a fixed number of matches and continuous covariates) because of the non-smoothness of the estimator. Note, however, that several matching algorithms applied in practice (e.g. kernel matching or the radius matching algorithm with regression-based bias correction of Lechner, Miquel, and Wunsch (2011)) are smoother than the one considered in Abadie and Imbens (2008), so that bootstrap inference may be valid in such cases. Furthermore, bootstrapping automatically accounts for heteroscedasticity, trimming of influential observations, and uncertainty due to propensity score estimation and bias correction. Even for non-smooth estimators like pair matching, it appears interesting whether the inconsistency of the bootstrap entails practically relevant biases. For this reason we apply two nonparametric bootstrap algorithms to all of our estimators.

The first algorithm bootstraps the ATET estimator directly. To this end, one randomly draws B bootstrap samples of size n with replacement out of the initial sample and compute the ATET estimate in each draw. We denote the latter by $\hat{\theta}^b$, where b is the index of the bootstrap sample, $b \in \{1, 2, ..., B\}$. We consider two options for computing p-values and confidence intervals in our simulations. The first one is based on plugging the square root of the bootstrap variance of the ATET, $\hat{V}(\hat{\theta}^b) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^b - \frac{1}{B}\sum_{b}^{B}\hat{\theta}^b\right)^2$, into the t-statistic and evaluating the latter on its asymptotic normal distribution to obtain the p-value. Confidence intervals are standardly obtained by $\hat{\theta} + / - \sqrt{\hat{V}(\hat{\theta}^b)}c$, where c denotes the asymptotic critical value for a particular confidence level α. The second one computes the p-value directly from the quantiles of the ATET estimates $\hat{\theta}^b$ (also known as percentile method), based on how frequently zero is included in the bootstrap distribution:

$$\text{p-value} = 2\min\left(\frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\{\hat{\theta}^b \leq 0\}, \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\{\hat{\theta}^b > 0\}\right). \tag{24}$$

The lower and upper bounds of the $1-\alpha$ confidence interval are computed by the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution, respectively.

The second algorithm accounts for the fact that the bootstrap has better theoretical properties when using an asymptotically pivotal statistic such as the t-statistic. Therefore, we in a first step compute the t-statistic using the variance estimators outlined in Sections 3.1 to 3.3: $T_n = \hat{\theta}/\sqrt{\hat{V}(\hat{\theta})}$, with $\hat{V}$ denoting some variance approximation. In the second step, we randomly draw B bootstrap samples of size n with replacement. In each draw, we compute the ATET estimate, denoted by $\hat{\theta}^b$, as well as the recentered t-statistic $T_n^b = (\hat{\theta}^b - \hat{\theta})/\sqrt{\hat{V}(\hat{\theta}^b)}$. The p-value is computed by the quantile or percentile method (see for instance MacKinnon (2006), equation (5)), i.e., as the share of absolute bootstrap t-statistics that are larger than the absolute value of the t-statistic in the original sample (as the t-statistic has a symmetric distribution):

$$\text{p-value} = 1 - \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{|T_n^b| \leq |T_n|\} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\{|T_n^b| > |T_n|\}, \tag{25}$$

where $|\cdot|$ denotes the absolute value of the argument. As a second option to compute the p-value, we also consider a smoothed version of (25) as suggested by Racine and MacKinnon (2007), see their equation (4):

$$\text{p-value} = 1 - \frac{1}{B}\sum_{b=1}^{B} K(T_n^b, T_n, h). \tag{26}$$

$K(T_n^b, T_n, h) = K\left(\frac{|T_n| - |T_n^b|}{h}\right)$ denotes the Gaussian cumulative kernel function for estimating the c.d.f. of the bootstrapped T_n^b evaluated at T_n, the t-statistic in the original sample. h denotes the bandwidth which is set to the optimal value for normally distributed T_n^b, $h = 1.575B^{-4/9}\sqrt{\hat{V}(T_n^b)}$, where $\hat{V}(T_n^b)$ is the variance of the bootstrap t-statistic. Racine and MacKinnon (2007) argue that due to a more efficient use of the information in the bootstrap statistics, the smoothed version increases power and can yield quite accurate results even when B is very small.

Concerning confidence intervals, computation is based on the following formula, see MacKinnon (2006):

$$\left[\hat{\theta} - \sqrt{\hat{V}(\hat{\theta})}T_n^b(1-\alpha/2), \hat{\theta} - \sqrt{\hat{V}(\hat{\theta})}T_n^b(\alpha/2)\right], \tag{27}$$

where $T_n^b(\tau)$ denotes the τ quantile of T_n^b and $\hat{V}(\hat{\theta})$ is an analytical variance estimate. That is, in contrast to conventional confidence intervals, the quantiles of the bootstrap distribution are used instead of the asymptotic critical value c. As discussed in MacKinnon (2006), quantile (or percentile) t-statistic confidence intervals have in theory a better higher-order accuracy than conventional intervals (either based on asymptotic or bootstrap standard errors). In our simulations, the number of bootstrap draws B is set to 199 for any method.[15] In addition, smaller values of B, namely 99 and 49, are also considered, in order to analyse the relationship between

[15] As discussed in MacKinnon (2006), the accuracy of bootstrap p-values that are based on the quantile method theoretically improves when choosing B such that $(B+1)$ times the confidence level is an integer.

bootstrap performance and number of bootstrap draws.

3.5 Wild bootstrap

The definition of the wild bootstrap procedure introduced in Bodory, Camponovo, Huber, and Lechner (2016) relies on the martingale representation for matching estimators proposed in Abadie and Imbens (2012a). Unlike the standard bootstrap, we do not construct bootstrap samples $(Z_1^*, \ldots, Z_n^*)$ by randomly selecting with replacement from $(Z_1, \ldots, Z_n)$, where $Z_i = (Y_i, D_i, X_i')'$. Instead, we fix the covariates and construct the bootstrap approximation by perturbating the martingale representation for matching estimators.

Consider the matching estimator introduced in (8) with weights defined in (9). Then, as shown in Abadie and Imbens (2012a), we can write the matching estimator as $\sqrt{n}(\hat{\theta}_{\text{match}} - \theta) = T_{1n} + T_{2n} + o_p(1)$, where

$$\begin{aligned} T_{1n} &= \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n} D_i(\mu(X_i, 1) - \mu(X_i, 0) - \theta), \\ T_{2n} &= \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n} \left(D_i - (1 - D_i)K_i\right)\left(Y_i - \mu(X_i, D_i)\right). \end{aligned}$$

The wild bootstrap algorithm uses this representation to reproduce the sampling distribution of $\sqrt{n}(\hat{\theta}_{\text{match}} - \theta)$. In particular, we can approximate the sampling distribution of the first term T_{1n} using the wild bootstrap distribution of

$$T_{1n}^* = \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n} D_i \hat{\xi}_i u_i,$$

where

$$\hat{\xi}_i = Y_i - \sum_{j:D_j=0} \varpi_{i,j} Y_j,$$

and $(u_1, \ldots, u_n)$ are iid random variables with $E[u_i] = 0$ and $E[u_i^2] = 1$. Unfortunately, using similar arguments as in Abadie and Imbens (2008), we can show that this approximation does not correctly reproduce the variability of T_{1n}. However, to overcome this distortion we can introduce a correction factor in the approximation of the second term T_{2n} that compensates the different variability of the first term; see, e.g, Theorem 7 in Abadie and Imbens (2008).

The approximation of the sampling distribution of the second term T_{2n} requires some care. Indeed, besides correcting the different variability of the approximation of the first term T_{1n}, we also need to capture the variability implied by the estimation of the propensity score in the definition of K_i. To overcome this problem, we apply the following approach. First, we generate random treatments D_i^* using the estimated propensity score $\hat{p}(X_i)$. Then, we re-estimate the propensity score $\hat{p}^*(X_i)$ using these bootstrap treatments $(D_1^*, \dots, D_n^*)$. Let K_i^* denote the number of times unit i is used as a match. Then, we approximate the sampling distribution of T_{2n} by the wild bootstrap distribution of

$$T_{2n}^* = \frac{\sqrt{n}}{n_1^*} \sum_{i=1}^{n} \Omega_i^* \hat{\epsilon}_{i,D_i^*} v_i,$$

where $n_1^* = \sum_{i=1}^n D_i^*$, $\Omega_i^* = \left((1 - D_i^*)\left(K_i^*(K_i^* - 1)\right)\right)^{1/2}$, $\hat{\epsilon}_{i,D_i^*} = \left(\hat{\sigma}^2(\hat{p}(X_i, D_i^*))\right)^{1/2}$ defined in (14), and $(v_1, \dots, v_n)$ are iid random variables with $E[v_i] = 0$ and $E[v_i^2] = 1$. The scaling factor Ω_i^* corrects for the different variability of the approximation of the first term T_{1n}. This term is also used in Theorem 7 in Abadie and Imbens (2008). Finally, we approximate the sampling distribution of $\sqrt{n}(\hat{\theta}_{\text{match}} - \theta)$ with the empirical bootstrap distribution of $T_{1n}^* + T_{2n}^*$.

A similar approach has been previously adopted in Otsu and Rai (2015), who introduce and prove the consistency of a weighted bootstrap procedure. However, unlike Otsu and Rai (2015), our bootstrap method can also be applied to propensity score matching. Moreover, our approach does not require the implementation of kernel estimators. As for the standard bootstrap, B is set to 199, 99, and 49 bootstrap draws.

3.6 Summary of the inference methods

Table 1 provides a summary of which inference procedures are investigated for which point estimators in our simulation study.

4 Simulation design

4.1 Data base and sample restrictions

The idea of an Empirical Monte Carlo Study (EMCS) is to base the data generating process (DGP) at least partially on real world data rather than models that are completely artificial (and arbitrary), see for instance Huber, Lechner, and Wunsch (2013), Lechner and Wunsch (2013), Huber, Lechner, and Steinmayr (2014), Huber, Lechner, and Mellace (2014), Lechner and Strittmatter (2014), and Frölich, Huber, and Wiesenfarth (2014). Our simulations exploit the same administrative data as used in Huber, Lechner, and Wunsch (2013), which comprise a 2 % random sample of employees in Germany who are subject to social insurance from 1990 to

Table 1: Inference methods and point estimators

variance estimator (row) / ATET estimator (column)	IPW	PM	R1.5	R3	R1.5BC	R3BC
asymptotic variance using GMM (Section 3.1)	x					
percentile bootstrap of t-stat based on GMM variance (Section 3.4)	x					
smoothed percentile bootstrap of t-stat based on GMM variance (Section 3.4)	x					
asymptotic variance of Abadie and Imbens (Section 3.2)		x				
bootstrap of t-stat based on asymptotic variance (Section 3.4)		x				
smoothed bootstrap of t-stat based on asymptotic variance (Section 3.4)		x				
wild bootstrap of t-stat based on asymptotic variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on asymptotic variance (Section 3.5)		x				
asymptotic variance of Abadie and Imbens with p-score correction (Section 3.2)		x				
bootstrap of t-stat based on asymptotic variance (Section 3.4)		x				
smoothed bootstrap of t-stat based on asymptotic variance (Section 3.4)		x				
wild bootstrap of t-stat based on asymptotic variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on asymptotic variance (Section 3.5)		x				
weights-based variance: unconditional variance (Section 3.3)	x	x	x	x	x	x
bootstrap of t-stat based on weights-based variance (Section 3.4)	x	x	x	x	x	x
smoothed bootstrap of t-stat based on weights-based variance (Section 3.4)	x	x	x	x	x	x
wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
weights-based variance: decomposition ($\hat{A}+\hat{B}$) (Section 3.3)	x	x	x	x	x	x
bootstrap of t-stat based on weights-based variance (Section 3.4)	x	x	x	x	x	x
smoothed bootstrap of t-stat based on weights-based variance (Section 3.4)	x	x	x	x	x	x
wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
weights-based variance: $\hat{A}$ (Section 3.3)	x	x	x	x	x	x
bootstrap of t-stat based on weights-based variance (Section 3.4)	x	x	x	x	x	x
smoothed bootstrap of t-stat based on weights-based variance (Section 3.4)	x	x	x	x	x	x
wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
weights-based var. with p-score correction: unconditional variance (Sections 3.3, 3.2)		x	x	x	x	x
bootstrap of t-stat based on weights-based variance (Section 3.4)		x	x	x	x	x
smoothed bootstrap of t-stat based on weights-based variance (Section 3.4)		x	x	x	x	x
wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
weights-based var. with p-score correction: decomposition ($\hat{A}+\hat{B}$) (Sections 3.3, 3.2)		x	x	x	x	x
bootstrap of t-stat based on weights-based variance (Section 3.4)		x	x	x	x	x
smoothed bootstrap of t-stat based on weights-based variance (Section 3.4)		x	x	x	x	x
wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
weights-based var. with p-score correction: $\hat{A}$ (Sections 3.3, 3.2)		x	x	x	x	x
bootstrap of t-stat based on weights-based variance (Section 3.4)		x	x	x	x	x
smoothed bootstrap of t-stat based on weights-based variance (Section 3.4)		x	x	x	x	x
wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
smoothed wild bootstrap of t-stat based on weights-based variance (Section 3.5)		x				
bootstrap of ATET to plug bootstrap std. error into t-stat (Section 3.4)	x	x	x	x	x	x
bootstrap of ATET using the quantile method (Section 3.4)	x	x	x	x	x	x
wild bootstrap of ATET to plug bootstrap std. error into t-stat (Section 3.4)		x				
wild bootstrap of ATET using the quantile method (Section 3.4)		x				

Note: IPW: inverse probability weighting; PM: pair matching; R1.5, R3: radius matching with a radius size of 1.5 or 3 times the maximum difference between matches occurring in pair matching, respectively; R1.5BC, R3BC: radius matching with bias correction as considered by Lechner, Miquel, and Wunsch (2011). Any of the bootstrap procedures is based on $B = 199, 99, 49$ bootstrap replications. The number of observations M (see (21)) used for the weights-based estimation of $\hat{A}$ and $\hat{B}$ is determined by $M = 2\ \mathrm{round}(\kappa\sqrt{n}) + 1$, with $\kappa = 0.2, 0.8, 3.2$.

2006. The data set combines information from four different registers: (i) employer-provided employee records to the social insurance agency (1990-2006), (ii) unemployment insurance records

(1990-2006), (iii) the programme participation register of the Public Employment Service (PES, 2000-2006) and (4) the jobseeker register of the PES (2000-2006). This entails a rich set of individual characteristics like gender, education, nationality, marital status, number of children, labor market history (since 1990), occupation, earnings, unemployment benefit claim, participation in active labor market programs, and others. Furthermore, a range of regional characteristics was also included, e.g. information about migration and commuting, average earnings, unemployment rate, long-term unemployment, welfare dependency rates, urbanization codes, and others.

Using the same sample restrictions as in Huber, Lechner, and Wunsch (2013), we consider all individuals entering unemployment between (and including) April 2000 and December 2003 in West Germany (without West Berlin) who were aged 20-59, had not been unemployed or in any labor market program in the 12 months before unemployment, and whose previous employment was not an internship or of any other non-standard form. Those unemployed individuals who start training courses that provide job-related vocational classroom training within the first 12 months of unemployment are defined as treated (3,266 observations), while those not participating in any active labor market program in the same period (114,349) are defined as non-treated. We consider two outcome variables in our simulations: average monthly earnings over the three years after entering unemployment (semi-continuous with 50% zeros), and an indicator whether there has been some form of (unsubsidized) employment in that period (binary).

4.2 Empirical Monte Carlo Study

Based on the sample with the restrictions, henceforth referred to as 'full sample', the EMCS proceeds as follows: (i) estimation of the propensity score (the conditional training probability) in the full sample which is then considered to be the 'true' population propensity score model, (ii) sampling of non-treated observations and simulation of a treatment (based on the coefficients of the 'true' propensity score model) for which the treatment effect and its variance are estimated and (iii) repeating the second step many times to assess the performance of the estimators.

Table 2: Descriptive statistics of the full sample

Variable	Treated		Non-treated		St.diff. in %	Probit model	
	mean	std.	mean	std.		m.eff. in %	s.e.
Some unsubsidized employment (Y)	0.63	0.48	0.56	0.5	9	-	-
av. monthly earnings (EUR) (Y)	1193	1115	1041	1152	9	-	-
Age / 10	3.67	0.84	3.56	1.11	8	7.3	0.5
Age squared / 1000	1.42	0.63	1.39	0.85	3	-9.1	0.6
20 - 25 years old	0.22	0.41	0.36	0.48	22	0.9	0.2
Women	0.57	0.5	0.46	0.5	15	-5.5	1.5
Not German	0.11	0.31	0.19	0.39	16	-0.5	0.1
Secondary degree	0.32	0.47	0.22	0.42	15	1.1	0.1
University entrance qualification	0.29	0.45	0.2	0.4	15	1	0.1
No vocational degree	0.18	0.39	0.34	0.47	26	-0.3	0.1
At least one child in household	0.42	0.49	0.28	0.45	22	-0.2	0.1
Last occupation: Non-skilled worker	0.14	0.35	0.21	0.41	13	0.4	0.2
Last occupation: Salaried worker	0.4	0.49	0.22	0.41	29	1.8	0.2
Last occupation: Part time	0.22	0.42	0.16	0.36	12	2.1	0.4
UI benefits: 0	0.33	0.47	0.44	0.5	16	-0.5	0.1
$>$ 650 EUR per month	0.26	0.44	0.22	0.41	7	0.8	0.2
Last 10 years before UE: share empl.	0.49	0.34	0.46	0.35	8	-1.4	0.2
share unemployed	0.06	0.11	0.06	0.11	1	-2.5	0.6
share in programme	0.01	0.04	0.01	0.03	9	5	1.4
share part time	0.16	0.33	0.11	0.29	10	-0.6	0.2
share out-of-the labour force (OLF)	0.28	0.4	0.37	0.44	14	-1.3	0.2
Entering UE in 2000	0.26	0.44	0.19	0.39	13	1.7	0.1
2001	0.29	0.46	0.26	0.44	5	0.9	0.1
2003	0.2	0.4	0.27	0.44	12	0	0.1
Share of pop. living in/ close to big city	0.76	0.35	0.73	0.37	6	0.4	0.1
Health restrictions	0.09	0.29	0.15	0.36	13	-0.6	0.1
Never out of labour force	0.14	0.34	0.11	0.31	6	0.6	0.1
Part time in last 10 years	0.35	0.48	0.29	0.45	9	-0.5	0.1
Never employed	0.11	0.31	0.2	0.4	17	-1.2	0.2
Duration of last employment $>$ 1 year	0.41	0.49	0.43	0.5	4	-0.6	0.1
Av. earn. last 10 yrs when empl./1000	0.59	0.41	0.52	0.4	13	-0.4	0.2
Woman × age / 10	2.13	1.95	1.65	1.94	17	2.7	0.6
× squared / 1000	0.83	0.85	0.65	0.9	15	-2.8	0.7
× no vocational degree	0.09	0.28	0.16	0.36	15	-0.9	0.1
× at least one child in household	0.32	0.47	0.17	0.37	25	1.1	0.2
× share OLF last year	0.19	0.36	0.18	0.35	3	0.8	0.2
× average earnings last 10 y. if empl.	0.26	0.34	0.19	0.3	16	-1.4	0.3
× entering UE in 2003	0.1	0.3	0.13	0.33	6	-0.6	0.1
$X_i\tilde{\beta}$	-1.7	0.4	-2.1	0.42	68	-	-
$\Phi(X_i\tilde{\beta})$	0.06	0.04	0.03	0.03	60	-	-
Number of obs., Pseudo-R2 in %	3266		114349			3.3	

Note: $\tilde{\beta}$ denotes the estimated probit coefficients in the full sample and $\Phi(X_i\tilde{\beta})$ is the c.d.f. of the standard normal distribution evaluated at $X_i\tilde{\beta}$. Pseudo-R^2 is the so-called Efron's R^2 $\left\{1-\frac{\sum_{i=1}^{n}[D_i-\Phi(X_i\tilde{\beta})]^2}{\sum_{i=1}^{n}[D_i-n^{-1}\sum_{i=1}^{n}D_i]^2}\right\}$. St.diff. (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). Mean, std., s.e. stand for mean, standard deviation, and standard error, respectively. M.eff.: Marginal effects evaluated at the mean in the probit model for treatment selection based on discrete changes for binary variables and derivatives otherwise.

Table 2 provides descriptive statistics for the treated and non-treated in the full sample, which is informative about selection into treatment relevant for step (i).[16] While the upper part presents descriptives for the two outcome variables average monthly earnings and the employment indicator, the remainder of the table focusses on the 36 confounders (among these seven interaction terms) that are included in the 'true' propensity score model used for the simulation of the placebo-treatments.[17] We also present the normalized differences between treated and non-treated as well as the marginal effects of the covariates at the means of all other covariates according to the 'true' propensity score, which point to considerable selection into treatment, as several variables are not balanced across treatment states.

After the estimation of the 'true' propensity score model in the full sample, the actually treated observations are discarded and no longer play a role in the simulations, leaving us with a 'population' of 114,349 observations. The next step is to randomly draw simulation samples of size n from the non-treated units with replacement. The sample sizes used in our simulations are 500 and 2000, in order to investigate the performance of the variance estimators both in moderate samples and in somewhat larger samples of a few 1000 observations as it frequently occurs in applied work. The extensive computational burden of some inference procedures (in particular the bootstrap) prevents us from investigating even larger samples sizes.[18] In each simulation sample, the (pseudo-)treatment is simulated among observations based on the coefficient estimates of the 'true' propensity score model in the full sample, which we denote by $\tilde{\beta}$ (note that a constant is included). To vary the strength of treatment selectivity, we consider two choices of selection into treatment based on the following equation:

$$D_i = \mathbb{I}\{\lambda X_i \tilde{\beta} + \delta + U_i > 0\}, \quad U_i \sim \mathcal{N}(0,1), \quad \lambda \in \{1, 2.5\}, \tag{28}$$

where U_i denotes a standard normally distributed random variable and λ determines selectivity (1=normal and 2.5=strong selection). As only a pseudo-treatment is assigned, the true effect on any individual is equal to zero no matter how strong selection is. Finally, δ gauges the shares of treated and non-treated and is chosen such that the expected number of treated equals 70%

[16] Note that some descriptives in Table 2 seemingly differ from those in Table 1 of Huber, Lechner, and Wunsch (2013), even though they refer to the same data. The reason is that in Huber, Lechner, and Wunsch (2013), the non-treated covariate means are incorrectly displayed in the column which claims to provide the standard deviations of the covariates of the treated, while the latter are given in the column which claims to show the non-treated covariate means. Therefore, Table 2 is correct, while the statistics in Table 1 of Huber, Lechner, and Wunsch (2013) are partially misplaced.

[17] We use almost the same covariates as Huber, Lechner, and Wunsch (2013), with the exception that drop the variable 'minor employment with earnings of no more than 400 EUR per month' and its interaction with gender, as this improves the small sample convergence of probit-based propensity score estimation.

[18] However, if a variance estimator turns out to already perform well for our n, we expect it to perform at least as well in larger samples.

or 30%, respectively.[19]

Note that in the simulation design outlined so far, effects are homogeneous as they are zero for everyone, because only a pseudo-treatment is considered. In order to investigate the performance of inference methods under heterogeneous effects, we in addition introduce models for the outcome variables. For the employment outcome, we create a uniformly distributed random variable $\epsilon_i \sim \mathcal{U}(0, 1.2)$, which is a function of the linear index of the 'true' propensity score in the full sample. To be specific,

$$\begin{aligned} f(X_i) &= \mathbb{I}\{|X_i\tilde{\beta}| \leq 3\}X_i\tilde{\beta} + \mathbb{I}\{|X_i\tilde{\beta}| > 3\}\bar{X}_i\tilde{\beta} - \min(X_i\tilde{\beta}), \\ \epsilon_i &= 1.2\frac{1.5f(X_i)/\max(f(X_i)) + W_i}{\max(1.5f(X_i)/\max(f(X_i)) + W_i)}. \end{aligned}$$

$\bar{X}_i$ denotes the vector of mean covariates in the 'population' of 114,349 observations, such that outliers with $|X_i\tilde{\beta}| > 3$ are trimmed to the average index when generating $f(X_i)$. $W_i \sim \mathcal{U}(0, 1)$ is a uniformly distributed simulated random variable. Then, among observations in the 'population' with the employment state equal to zero, the employment outcome is switched to one if $\epsilon_i > 0.7$, while among observations with employment equal to one, it is set to zero if $\epsilon_i < 0.15$. This introduces effect heterogeneity w.r.t. the index and implies that 69% of the 'population' are employed (vs. just 56% under effect homogeneity). Concerning the earnings outcome, effect heterogeneity is based on $\varepsilon_i \sim \mathcal{U}(0.994, 1.346)$ which is generated in the following way:

$$\varepsilon_i = 0.21[f(X_i)/\max(f(X_i)) + W_i] + 0.945.$$

ε_i is added to positive earnings outcomes of any individuals in the 'population' with employment equal to one under effect homogeneity. For those observations without earnings whose employment state has been switched to one to introduce effect heterogeneity, the average of all positive earnings (under effect homogeneity) multiplied by $(3\varepsilon_i - 2.4)$ is added. This entails average earnings of 1,247.29 EUR in our 'population' of 114,349 observations (vs. 1,040.96 under effect homogeneity).

Table 3 summarizes the 8 scenarios that are considered in the EMCS and gives statistics about the strength of selection implied by each.[20] Combined with two sample sizes, we therefore

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

[20] The standardized differences as well as the pseudo-R^2s are based on a re-estimated propensity score in the actually non-treated sample (114,349 obs.), the 'population' in which the pseudo-treatment is assigned. However, when reassigning observations to act as simulated treated, the pool of non-treated is changed. Together with the fact that the treatment share differs from the original share leads to different values of those statistics even in the case that mimics selection in the full sample.

run all in all 16 simulations. Similar to Huber, Lechner, and Wunsch (2013), the number of Monte Carlo replications is proportional to the sample size, consisting of 10,000 replications for the smaller and 2,500 for the the larger sample size, as the latter is computationally more expensive, but has less variability in results across simulation samples.

Table 3: Summary statistics (DGPs)

Strength of	Share of	St.diff. of	Pseudo-R^2 of	Y(1)		Y(0)		ATET		Trimming in %	
selection	treated in %	p-score in %	probit in %	mean	std	mean	std	mean	std	500 obs	2000 obs
Effect homogeneity (employment)											
Normal	70	41	8.7	0.6	0.5	0.6	0.5	0	0	5.2	0
Normal	30	42	9.1	0.6	0.5	0.6	0.5	0	0	0.1	0
Strong	70	81	33.8	0.6	0.5	0.6	0.5	0	0	28.6	8.7
Strong	30	89	34.4	0.7	0.5	0.7	0.5	0	0	5.5	0.2
Effect homogeneity (earnings)											
Normal	70	41	8.7	11.0	11.8	11.0	11.8	0	0		
Normal	30	42	9.1	11.9	12.2	11.9	12.2	0	0		
Strong	70	81	33.8	11.7	11.9	11.7	11.9	0	0		
Strong	30	89	34.4	12.9	12.7	12.9	12.7	0	0		
Effect heterogeneity (employment)											
Normal	70	42	9.1	0.8	0.4	0.6	0.5	0.2	0.4	5.3	0
Normal	30	42	8.9	0.8	0.4	0.6	0.5	0.2	0.4	0.1	0
Strong	70	81	33.9	0.8	0.4	0.6	0.5	0.2	0.4	29.3	9.1
Strong	30	89	34.2	0.8	0.4	0.7	0.5	0.1	0.4	5.6	0.2
Effect heterogeneity (earnings)											
Normal	70	42	9.1	13.6	11.2	11.1	11.8	2.5	6.6		
Normal	30	42	8.9	14.5	11.3	11.8	12.1	2.7	6.8		
Strong	70	81	33.9	14.3	11.2	11.7	12.0	2.6	6.7		
Strong	30	89	34.2	16.0	11.5	12.9	12.6	3.1	7.2		

Note: Pseudo-R^2 is the so-called Efron's R^2 $\left\{1-\frac{\sum_{i=1}^{n}[D_i-\Phi(X_i\tilde{\beta})]^2}{\sum_{i=1}^{n}[D_i-n^{-1}\sum_{i=1}^{n}D_i]^2}\right\}$. St.diff. of p-score (standardized difference of the propensity score) is defined as the difference of average propensity scores across treatment states normalized by the square root of the sum of estimated variances of the propensity scores in either state (see e.g. Imbens and Wooldridge (2009), p. 24). $Y(1)$ and $Y(0)$ denote the potential outcomes for the randomly generated treated observations under treatment and non-treatment, respectively. The (true) treatment effects on the treated (ATETs) are the differences between these potential outcomes. The means and standard deviations (std) are displayed for the potential outcomes and the corresponding ATETs. For earnings only, the values of Y(1), Y(0), and ATET are shown in hundreds. Mean and std of Y(0) can differ slightly between homogenous and heterogeneous DGPs because they are generated with different random number states (GAUSS Version 15.1.3). Trimming in % shows the share of observations dropped in the respective DGPs due to support problems (Section 2.3). Since trimming does not depend on the outcomes, the shares are presented in the employment tables only.

Table 4 presents the biases and standard deviations of the effect estimators under the different DGPs. While the upper panels refer to the various cases under homogeneity and zero effects, the lower panels refer to the case of non-zero heterogeneous effects.

We find that overall, the biases of estimators are small. Concerning their relative performance, as expected, nearest neighbor matching is the noisiest, while IPW weighting does very well, since there are no substantial issues of lack of or thin support in these DGPs. The other matching estimators are somewhat in-between these cases. Comparing the standard errors of the estimators across DGPs demonstrates that the approximation of $\sqrt{n}$-convergence usually appears to be reasonable for the case of normal selection, as the standard errors in larger samples tend to be half the size of those in the smaller samples. However, for the case of strong selection, the speed of convergence is clearly much higher which indicates that the asymptotic normal distribution may not be a good approximation of the distribution of the estimators in these cases.[21]

Table 4: Performance of ATET estimators for all DGPs

	Effect homogeneity															
	500 obs				2000 obs				500 obs				2000 obs			
Estimation	empl		earn		empl		earn		empl		earn		empl		earn	
method	bias	se	bias	se	bias	se	bias	se	bias	se	bias	se	bias	se	bias	se
	Normal selection, 30% treated								Normal selection, 70% treated							
IPW	0.2	4.7	3.6	117	-0.1	2.3	-1.0	56.5	0.4	5.3	10.7	131	0.0	2.7	1.1	68.8
PM	0.2	6.7	4.8	167	0.0	3.2	-0.3	81.6	0.2	7.8	5.0	196	-0.1	3.5	0.0	87.4
R1.5	0.2	5.6	5.0	140	0.0	2.8	-0.3	69.0	0.1	6.4	4.1	159	0.0	3.1	0.0	75.1
R3	0.3	5.5	6.4	137	0.0	2.8	-0.3	68.9	0.2	6.2	5.5	153	0.1	3.0	0.7	73.9
R1.5BC	0.0	5.6	1.3	133	-0.1	2.7	-1.6	64.3	-0.3	6.5	-4.5	151	-0.1	3.1	-0.9	71.5
R3BC	-0.1	5.5	0.7	131	-0.1	2.7	-1.9	64.2	-0.3	6.3	-5.0	147	-0.1	3.0	-1.1	70.7
	Strong selection, 30% treated								Strong selection, 70% treated							
IPW	0.2	6.2	7.7	164	0.2	3.4	-3.2	96.4	0.6	7.0	17.8	160	0.7	4.3	20.1	111
PM	0.0	9.2	-1.5	252	0.2	4.8	-4.2	142	0.3	10.5	7.5	247	0.3	7.4	7.9	189
R1.5	0.0	7.7	-0.2	208	0.2	4.0	-7.2	116	0.3	8.8	8.0	205	0.4	5.8	10.3	146
R3	0.1	7.4	2.0	199	0.2	3.9	-6.4	111	0.3	8.4	9.0	194	0.4	5.3	10.6	135
R1.5BC	-0.4	7.7	-5.4	191	0.1	4.0	-2.8	106	-0.4	8.8	-3.4	192	-0.2	5.8	-2.1	135
R3BC	-0.3	7.5	-5.2	186	0.1	3.9	-2.9	104	-0.4	8.5	-3.5	185	-0.2	5.5	-2.5	128
	Effect heterogeneity															
	Normal selection, 30% treated								Normal selection, 70% treated							
IPW	0.2	4.5	0.5	119	0.1	2.1	-2.0	54.8	0.6	5.3	13.2	130	0.0	2.5	1.0	64.4
PM	0.2	6.6	1.2	170	0.2	3.0	-2.0	79.6	0.2	7.8	4.3	196	0.0	3.4	-3.4	86.9
R1.5	0.2	5.5	1.4	143	0.2	2.6	-1.7	66.8	0.3	6.4	5.6	159	0.1	2.9	-1.2	72.4
R3	0.3	5.4	3.0	140	0.2	2.5	-1.6	66.6	0.3	6.2	6.9	153	0.1	2.9	-0.9	71.3
R1.5BC	0.0	5.5	-2.2	136	0.1	2.5	-2.1	62.2	-0.2	6.5	-1.3	151	0.0	2.9	-2.1	69.6
R3BC	0.0	5.4	-2.7	134	0.1	2.5	-2.3	62.1	-0.2	6.3	-1.9	147	0.0	2.9	-2.5	68.9
	Strong selection, 30% treated								Strong selection, 70% treated							
IPW	0.1	6.0	5.9	161	0.5	3.2	-0.3	92.0	-1.5	7.0	-25.2	160	-0.2	4.2	8.3	111
PM	-0.1	9.2	-1.9	251	0.4	4.7	-7.2	137	-1.7	10.7	-35.0	248	-0.7	7.2	-7.3	193
R1.5	0.0	7.6	-0.9	205	0.4	3.9	-6.1	112	-1.8	8.9	-35.5	204	-0.8	5.7	-5.7	152
R3	0.0	7.3	0.9	196	0.5	3.8	-5.9	108	-1.8	8.5	-34.5	193	-0.7	5.3	-4.2	140
R1.5BC	-0.4	7.6	-6.5	189	0.4	4.0	-3.1	102	-2.6	8.8	-47.3	191	-1.3	5.8	-16.4	138
R3BC	-0.4	7.4	-6.6	183	0.4	3.9	-3.3	100	-2.5	8.5	-47.2	185	-1.3	5.5	-16.6	131

Note: IPW: inverse probability weighting; PM: pair matching; R1.5, R3: radius matching with a radius size of 1.5 or 3 times the maximum difference between matches occurring in pair matching, respectively; R1.5BC, R3BC: radius matching with bias correction as considered by Lechner, Miquel, and Wunsch (2011). Sample sizes: 500 or 2000 observations (obs). Outcomes: employment (empl) and earnings (earn). The performance of the estimators is evaluated by their biases and standard errors (se).

5 Results

This section evaluates the performance of the various inference methods for the different point estimators. For the sake of brevity, we present only a limited amount of evidence in the main

[21] Table 3 (columns 11 and 12) indicates that sample size reductions due to common support issues are considerably higher in cases of strong selection. However, this should not affect the results because inference is based on samples selected after these reductions.

body of the paper which conveys the main message of our findings, as the latter seem, perhaps surprisingly, rather unambiguous. An extensive set of further results is presented in Appendix A.

Table 5 provides the rejection probabilities of the inference procedures by distinct point estimators and outcomes. The null hypothesis corresponds to classical significance tests, namely that the respective mean effect is zero. The upper panel contains the results for IPW, the intermediate one for pair matching, and the lower one for radius matching. In the case of radius matching, the rejection probabilities are averaged over the four estimators investigated (R1.5, R3, R1.5BC, R3BC), because their inference results are qualitatively very similar, see Table A.1 in Appendix A for a separate analysis of each radius matching algorithm. Furthermore, Table 5 aggregates over the different DGP features, with the exception of effect homogeneity (left panel) vs. heterogeneity (right panel). The reason is that under homogeneity, the null hypothesis is true as any effect is equal to zero, while it is violated under heterogeneity. Thus, the rejection probabilities relating to the former case reflect the size of the tests, while those under heterogeneity are informative about the power. As shown in Table A.8 in Appendix A, the coverage probabilities of the various procedures, i.e. the share of simulations in which the true value is included in the 95% confidence interval of the respective method, do not differ much for the homogeneous and heterogeneous case.

A first observation in Table 5 is that all methods that are based on asymptotic approximations (i.e. do not rely on bootstrapping) and ignore the estimation of the propensity score are conservative. This is for instance in line with Abadie and Imbens (2012a), who show that estimating (rather than knowing) the propensity score changes the variance of the matching estimator of the ATET (albeit the direction of the change is ambiguous). Interestingly, also the GMM-based variance estimator for IPW is rather conservative, even though it accounts for the estimation of the propensity score. As any of these procedures are computationally much less expensive than the bootstrap, they may provide reasonable approximations in very large data sets. On the other hand, some procedures, namely weights-based variance estimation using term A alone and the Abadie and Imbens estimator, have excessive size when adjusted for propensity score estimation. In conclusion, weights-based estimation using A without propensity score adjustment appears to be among the best performing asymptotic methods for any point estimator in terms of size.

Table 5: Rejection probabilities

IPW	homogeneity						heterogeneity					
	binary			continuous			binary			continuous		
	as	bs		as	bs		as	bs		as	bs	
GMM	0.2	3.9		0.6	4.3		55.0	84.9		39.7	63.2	
wgt uncond var	1.0	4.3		1.6	4.6		72.6	85.1		51.5	63.3	
wgt decomp (0.2)	0.6	4.2		1.0	4.2		72.7	87.1		50.2	64.2	
wgt decomp (0.8)	0.6	4.2		1.1	4.1		73.1	87.3		51.0	64.6	
wgt decomp (3.2)	0.6	4.3		1.2	4.1		74.9	88.0		54.7	65.3	
wgt A (0.2)	3.3	3.7		3.3	3.7		88.5	87.4		64.4	63.2	
wgt A (0.8)	3.2	3.8		3.3	3.8		88.5	87.5		64.9	64.0	
wgt A (3.2)	2.9	4.0		3.8	4.0		87.8	88.2		66.9	64.7	
boot effect se		4.1			4.3			87.8			64.7	
boot effect quant		2.7			2.9			89.0			61.5	
pair matching	homogeneity						heterogeneity					
	binary			continuous			binary			continuous		
	as	bs	wbs	as	bs	wbs	as	bs	wbs	as	bs	wbs
wgt uncond var	2.6	6.1	4.6	3.7	7.4	4.9	50.2	60.2	62.5	36.9	46.9	43.9
wgt decomp (0.2)	0.4	3.1	3.5	1.0	3.2	3.8	46.8	69.7	67.6	30.1	47.3	43.5
wgt decomp (0.8)	0.3	2.9	3.5	0.9	2.9	3.6	47.2	69.9	67.7	30.6	47.2	43.7
wgt decomp (3.2)	0.3	2.6	3.5	1.4	2.1	3.8	49.9	71.0	68.2	34.1	46.3	43.9
wgt A (0.2)	3.3	2.5	3.7	4.3	2.3	3.9	69.2	69.7	67.9	46.5	45.5	43.2
wgt A (0.8)	3.2	2.4	3.7	4.3	2.2	3.8	69.2	70.1	68.1	47.1	45.7	43.5
wgt A (3.2)	2.6	2.2	3.5	5.2	1.9	3.8	68.2	70.6	68.3	49.4	45.2	43.7
Abadie Imbens	7.4	1.5	5.5	6.6	1.3	5.0	71.6	51.4	61.9	48.4	34.2	41.4
wgt uncond var ps	4.6	7.5	4.7	6.6	9.2	5.4	55.1	60.6	60.3	43.9	48.7	44.0
wgt decomp (0.2) ps	1.3	4.6	3.9	3.3	5.2	4.5	53.6	70.8	66.6	38.1	49.6	43.8
wgt decomp (0.8) ps	1.2	4.4	3.7	3.5	4.8	4.5	54.0	71.0	66.5	38.7	49.6	44.0
wgt decomp (3.2) ps	1.5	3.9	3.8	5.3	4.6	4.6	58.0	72.4	66.8	44.2	49.5	44.0
wgt A (0.2) ps	9.7	5.2	4.7	11.6	5.6	5.2	77.2	70.7	63.8	58.0	49.5	43.8
wgt A (0.8) ps	9.5	5.3	4.6	11.8	5.2	5.0	77.2	71.1	64.1	58.9	50.0	44.5
wgt A (3.2) ps	8.3	5.1	4.5	13.6	5.4	5.2	76.1	71.2	64.3	61.8	50.3	45.0
Abadie Imbens ps	14.8	2.9	6.6	15.0	3.0	6.4	82.6	58.7	64.0	61.8	41.1	44.6
boot effect se		2.0	3.6		2.3	3.8		60.5	64.6		38.0	41.3
boot effect quant		0.1	3.5		0.1	3.6		67.2	62.1		32.5	39.6
radius matching	homogeneity						heterogeneity					
	binary			continuous			binary			continuous		
	as	bs		as	bs		as	bs		as	bs	
wgt uncond var	0.9	3.3		1.6	3.7		59.2	73.4		42.0	54.3	
wgt decomp (0.2)	0.4	2.9		0.9	2.8		58.5	76.8		39.6	54.9	
wgt decomp (0.8)	0.4	2.8		0.9	2.7		59.0	77.0		40.4	55.2	
wgt decomp (3.2)	0.4	2.8		1.1	2.6		60.8	77.8		43.5	55.8	
wgt A (0.2)	3.6	2.6		3.4	2.4		78.4	76.4		54.7	53.6	
wgt A (0.8)	3.5	2.6		3.4	2.4		78.2	76.7		55.3	54.2	
wgt A (3.2)	3.1	2.5		3.9	2.4		77.6	77.6		57.3	55.0	
wgt uncond var ps	2.8	4.8		4.8	5.8		67.5	73.1		53.0	57.0	
wgt decomp (0.2) ps	2.0	4.5		3.8	5.1		67.5	76.9		50.9	57.5	
wgt decomp (0.8) ps	2.0	4.5		3.9	5.0		68.0	77.1		52.1	57.9	
wgt decomp (3.2) ps	2.1	4.6		5.1	5.5		70.0	78.1		56.5	58.9	
wgt A (0.2) ps	9.9	4.8		11.3	5.8		84.6	75.0		67.7	56.9	
wgt A (0.8) ps	9.7	4.9		11.5	5.8		84.5	75.3		68.3	57.5	
wgt A (3.2) ps	9.0	4.9		12.8	6.3		83.8	76.2		69.8	58.4	
boot effect se		3.3			3.4			74.4			53.8	
boot effect quant		1.0			1.0			77.3			50.1	

Note: 'as': the standard error is estimated by the respective method and plugged into the asymptotic approximation for confidence intervals; 'bs': using 199 (standard) bootstrap replications (without smoothing), the standard error is estimated by the respective method and plugged into the t-statistic to obtain confidence intervals based on the quantile method. Exceptions are 'boot effect se', which bootstraps the effect and plugs its standard error into the asymptotic approximation for confidence intervals, and 'boot effect quant', which obtains confidence intervals based on the quantile method on the effect (rather than the t-statistic); 'wbs': wild bootstrap (without smoothing) rather than the standard bootstrap is used for the respective method. The suffix 'ps' stands for adjustment for propensity score estimation. The results for radius matching are averages over all 4 radius matching algorithms (R1.5, R3, R1.5BC, R3BC).

In general, it appears that the bootstrap procedures dominate the asymptotic ones in that they are much closer to the nominal size under homogeneity. This holds also true for IPW estimation, where any bootstrap method performs better than its asymptotic counterpart in terms of size. For the matching estimators, the methods based on bootstrapping t-statistics which include the propensity score adjustment generally do best in terms of size, while the ones without adjustment are conservative, with the exception of the Abadie and Imbens estimator when using the wild bootstrap. Concerning pair matching, it is interesting to observe that the inconsistency of the standard bootstrap demonstrated in Abadie and Imbens (2008) seem to have little practical relevance when using the propensity score adjustment, with the exception of weights-based estimation using the unconditional variance. Furthermore, the wild bootstrap improves on several inference estimators compared to standard bootstrapping, in particular when propensity score estimation is ignored. Overall, it appears that with the exception of the unconditional variance formula, the weights-based estimators with propensity score adjustment (both based on the standard or wild bootstrap) most accurately estimate the size of pair matching. For radius matching, bootstrapping t-statistics with propensity score adjustment and weights-based estimation using A+B appears to overall dominate in terms of empirical size. It is interesting to note that for each of the point estimators, the best performing methods based on bootstrapping t-statistics outperform the approaches of directly bootstrapping the ATET (to either plug the bootstrap standard error into the asymptotic approximation or to apply the quantile method on the ATET).

The previous discussion exclusively focussed on the empirical size. Concerning the power of the inference procedures, Table 5 shows that none of the methods have severe lack-of-power issues, not even the conservative ones. Again, the bootstrap-based procedures often dominate the asymptotic approximations or are at least comparably powerful, with the exception of those asymptotic methods that are severely oversized. That is, it frequently appears that some bootstrap method is superior in terms of size and power at the same time when compared to the respective asymptotic approximation. The patterns of our results on size and power are quite similar across outcome variables, although in some cases the asymptotic procedures are less conservative for the continuous than for the binary outcome. In Tables A.2 to A.4 in Appendix A we provide the rejection probabilities of the methods (separately for each point estimator) across further simulation features like sample size, share of treated, and treatment selection.[22] It is a maybe surprising result that the size and power properties of the different inference procedures are rather stable across the different features.

[22] The corresponding coverage probabilities, i.e. the share of simulations in which the true value is included in the 95% confidence interval of the respective method, are provided in Tables A.9 to A.11 of Appendix A.

A further remark concerns the number of bootstrap replications, which we varied between 49 and 199 (even though in any of the tables mentioned so far, only the results for 199 replications are reported). On average, it seems that reliable inference is obtained already with just 49 bootstraps, no matter whether non-smoothed or smoothed statistics are considered. Even though increasing the number of replications improves the procedures, the gains appear to be rather small for any point estimator, see Tables A.6 and A.7 in Appendix A. For the non-smoothed bootstrap procedures, a larger number of replications generally slightly decreases the standard deviations of the rejection probabilities (results not reported but available on request), albeit the effect is rather minor. It is furthermore more or less non-existent for the smoothed versions, as smoothing decreases the standard deviations under a low number of bootstraps somewhat such that an increase does not entail further reductions. Smoothing has, however, virtually no effect on the coverage probabilities (i.e. on average), see Tables A.6 and A.7. It generally decreases the size under a small number of replications, but has a negligible to nonexistent impact for 199 replications.

6 Conclusion

In this paper, we investigated the finite sample properties of various inference methods for propensity score-based matching and weighting estimators of the average treatment effect on the treated. Using an 'Empirical Monte Carlo Study' (EMCS) approach based on large scale labor market data from Germany, we analysed both asymptotic approximations and several bootstrap methods for the computation of variances and confidence intervals. We found that asymptotic approximations that ignore the estimation of the propensity score tended to be conservative, while accounting for propensity score estimation led to excessive size for some procedures applicable to matching estimators. In contrast, GMM-based variance estimation of IPW was rather conservative, even though accounting for the estimation of the propensity score. In general, the bootstrap procedures dominated the asymptotic ones in terms of size. For matching, the methods based on bootstrapping t-statistics which account for propensity score estimation generally came closest to the nominal size. For pair matching, it was interesting to see that the inconsistency of the standard bootstrap bore little practical relevance for most methods accounting for propensity score estimation. Yet, a wild bootstrap algorithm applicable to propensity score matching led to a more accurate size for several inference estimators than standard bootstrapping, in particular when the estimation of the propensity score was ignored. Concerning power, none of the methods showed severe lack-of-power issues, but again, the bootstrap procedures frequently outperformed the asymptotic approximations or were at least comparably powerful.

Furthermore, the size and power properties of the inference procedures were rather stable across different simulation features. Finally, for the bootstrap procedures, we found only minor effects of the number of bootstrap replications on their average performance.

References

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

——— (2008): "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76, 1537–1557.

——— (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29, 1–11.

——— (2012a): "A Martingale representation for matching estimators," *Journal of the American Statistical Association*, 107, 833–843.

——— (2012b): "Matching on the Estimated Propensity Score," *working paper.*

Bodory, H., L. Camponovo, M. Huber, and M. Lechner (2016): "A wild bootstrap algorithm for direct and propensity score matching estimators," *mimeo.*

Busso, M., J. DiNardo, and J. McCrary (2014): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *Review of Economics and Statistics*, 96, 885–897.

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

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

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

——— (2007): "Propensity score matching without conditional independence assumption - with an application to the gender wage gap in the United Kingdom," *Econometrics Journal*, 10, 359–407.

Frölich, M., M. Huber, and M. Wiesenfarth (2014): "The finite sample performance of semi- and nonparametric estimators for treatment effects and policy evaluation," *IZA Discussion Paper No. 8756.*

Graham, B., C. Pinto, and D. Egel (2012): "Inverse probability tilting for moment condition models with missing data," *Review of Economic Studies*, 79, 1053–1079.

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

Hainmueller, J. (2012): "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20(1), 25–46.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing selection bias using experimental data," *Econometrica*, 66, 1017–1098.

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

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 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*.

HOROWITZ, J. L. (2001): “The Bootstrap,” in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Learer, pp. 3159–3228. North-Holland.

HORVITZ, D., AND D. THOMPSON (1952): “A Generalization of Sampling without Replacement from a Finite Population,” *Journal of American Statistical Association*, 47, 663–685.

HUBER, M., M. LECHNER, AND G. MELLACE (2014): “The finite sample performance of estimators for mediation analysis under sequential conditional independence,” *University of St. Gallen, Department of Economics Discussion Paper No. 2014-15.*

HUBER, M., M. LECHNER, AND A. STEINMAYR (2014): “Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behaviour,” *forthcoming in Empirical Economics.*

HUBER, M., M. LECHNER, AND C. WUNSCH (2013): “The performance of estimators based on the propensity score,” *Journal of Econometrics*, 175, 1–21.

IMAI, K., AND M. RATKOVIC (2014): “Covariate balancing propensity score,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243–263.

IMBENS, G. W. (2004): “Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review,” *The 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.

KHAN, S., AND E. TAMER (2010): “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” *Econometrica*, 78, 2021–2042.

LECHNER, M. (2002): “Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies,” *The Review of Economics and Statistics*, 84, 205–220.

——— (2009): “Sequential Causal Models for the Evaluation of Labor Market Programs,” *Journal of Business and Economic Statistics*, 27, 71–83.

LECHNER, M., R. MIQUEL, AND C. WUNSCH (2011): “Long-run Effects of Public Sector Sponsored Training in West Germany,” *Journal of the European Economic Association*, 9, 742–784.

LECHNER, M., AND A. STRITTMATTER (2014): “Practical Procedures to Deal with Common Support Problems in Matching Estimation,” *University of St. Gallen, Department of Economics Discussion Paper Discussion Paper no. 2014-10.*

LECHNER, M., AND C. WUNSCH (2013): “Sensitivity of matching-based program evaluations to the availability of control variables,” *Labour Economics*, 21, 111–121.

LUNCEFORD, J. K., AND M. DAVIDIAN (2004): “Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study,” *Statistics in Medicine*, 23, 2937–2960.

MACKINNON, J. G. (2006): “Bootstrap Methods in Econometrics,” *The Economic Record*, 82, S2–S18.

MILLIMET, D., 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.

Ñopo, H. (2008): “Matching as a Tool to Decompose Wage Gaps,” *Review of Economics and Statistics*, 90, 290–299.

Otsu, T., and Y. Rai (2015): “Bootstrap inference of matching estimators for average treatment effects,” *working paper, University of St. Gallen.*

Pingel, R. (2015): “Estimating the Variance of a Propensity Score Matching Estimator: Another Look at Right Heart Catheterization Data,” *working paper, Department of Statistics, Uppsala University.*

Racine, J. S., and J. G. MacKinnon (2007): “Inference via kernel smoothing of bootstrap values,” *Computational Statistics & Data Analysis*, 51, 5949–5957.

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.

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., A. Rotnitzky, and L. P. Zhao (1995): “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data,” *Journal of the American Statistical Association*, 90, 106–121.

Rosenbaum, P. R., and D. B. Rubin (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70(1), 41–55.

——— (1985): “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.,” *The American Statistician*, 39, 33–38.

Rothe, C., and S. Firpo (2013): “Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions,” *IZA Discussion Paper No. 7564.*

Rubin, D. B. (1973): “Matching to Remove Bias in Observational Studies,” *Biometrics*, 29, 159–183.

——— (1974): “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688–701.

——— (1979): “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies,” *Journal of the American Statistical Association*, 74, 318–328.

——— (1990): “Formal Modes of Statistical Inference For Causal Effects,” *Journal of Statistical Planning and Inference*, 25, 279–292.

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

Waernbaum, I. (2012): “Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation,” *Statistics in Medicine*, 31, 1572–1581.

Zhao, Z. (2004): “Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence,” *Review of Economics and Statistics*, 86, 91–107.

——— (2008): “Sensitivity of Propensity Score Methods to the Specifications,” *Economics Letters*, 98, 309–319.

A Appendix

Table A.1: Rejection probabilities for radius matching

	homogeneity				heterogeneity				homogeneity				heterogeneity			
	binary		continuous		binary		continuous		binary		continuous		binary		continuous	
	as	bs	as	bs	as	bs	as	bs	as	bs	as	bs	as	bs	as	bs
	radius matching R1.5								*radius matching R3*							
wgt uncond var	1.4	3.5	2.4	4.2	58.4	68.2	42.7	50.8	1.1	3.4	2.0	4.0	60.4	72.4	43.3	52.9
wgt decomp (0.2)	0.4	2.7	1.0	2.5	57.2	74.9	38.8	51.6	0.5	2.8	1.0	2.8	59.7	77.3	40.5	53.7
wgt decomp (0.8)	0.4	2.6	1.0	2.5	57.6	75.2	39.5	51.9	0.4	2.6	1.0	2.6	60.2	77.5	41.2	54.1
wgt decomp (3.2)	0.4	2.5	1.2	2.2	59.6	76.3	42.8	52.2	0.4	2.6	1.2	2.4	62.1	78.6	44.4	54.6
wgt A (0.2)	3.5	2.4	3.8	2.1	78.2	74.3	54.2	49.9	3.6	2.5	3.7	2.3	80.2	77.2	55.7	52.3
wgt A (0.8)	3.4	2.4	3.7	2.0	77.9	74.9	55.0	50.6	3.5	2.5	3.7	2.3	80.0	77.3	56.3	52.9
wgt A (3.2)	2.8	2.3	4.3	2.0	77.2	76.1	56.8	51.3	2.9	2.4	4.2	2.2	79.2	78.5	58.3	53.9
wgt uncond var ps	3.5	4.8	5.8	6.1	66.2	68.2	52.9	53.3	3.2	4.8	5.5	6.0	68.3	71.5	54.1	55.4
wgt decomp (0.2) ps	2.0	4.3	3.9	4.7	66.3	75.2	49.3	54.2	2.2	4.4	4.1	5.0	68.7	76.9	51.7	56.0
wgt decomp (0.8) ps	1.9	4.2	4.1	4.6	67.0	75.5	50.7	54.5	2.1	4.3	4.2	4.9	69.1	77.4	52.9	56.5
wgt decomp (3.2) ps	2.0	4.2	5.4	4.9	69.0	76.7	55.2	55.5	2.2	4.4	5.5	5.2	71.0	78.6	57.7	57.7
wgt A (0.2) ps	9.8	4.4	12.1	5.3	84.4	73.0	67.2	53.6	9.9	4.7	12.0	5.6	86.2	75.2	68.5	55.7
wgt A (0.8) ps	9.6	4.4	12.2	5.3	84.2	73.5	67.8	54.3	9.6	4.8	12.4	5.7	86.0	75.7	69.1	56.4
wgt A (3.2) ps	8.8	4.6	13.5	5.8	83.5	74.7	69.4	55.4	8.9	4.8	13.3	6.0	85.3	76.8	70.7	57.4
boot effect se		3.2		3.2		73.8		51.2		3.4		3.5		77.3		54.1
boot effect quant		0.8		0.8		78.6		46.1		0.9		1.0		80.9		49.9
	radius matching with bias correction R1.5BC								*radius matching with bias correction R3BC*							
wgt uncond var	0.7	3.0	1.2	3.1	58.7	75.4	41.0	55.9	0.6	3.2	1.0	3.3	59.3	77.4	41.1	57.7
wgt decomp (0.2)	0.4	3.0	0.8	2.9	58.2	76.6	39.5	56.1	0.4	3.2	0.8	3.0	58.8	78.3	39.7	57.9
wgt decomp (0.8)	0.4	2.9	0.9	2.8	58.7	76.7	40.4	56.5	0.4	3.2	0.8	3.0	59.3	78.5	40.5	58.4
wgt decomp (3.2)	0.4	3.0	1.0	2.9	60.4	77.3	43.3	57.2	0.3	3.1	1.0	3.1	61.0	79.1	43.5	59.0
wgt A (0.2)	3.8	2.6	3.2	2.5	77.4	76.2	54.5	55.1	3.6	2.9	2.9	2.8	78.0	78.1	54.6	57.0
wgt A (0.8)	3.7	2.6	3.2	2.6	77.2	76.4	55.1	55.7	3.5	2.9	2.8	2.8	77.9	78.2	54.9	57.5
wgt A (3.2)	3.4	2.6	3.7	2.6	76.7	76.9	56.9	56.4	3.1	2.8	3.3	2.8	77.4	78.8	57.0	58.3
wgt uncond var ps	2.4	4.6	4.1	5.5	67.5	75.7	52.3	58.9	2.2	5.0	3.9	5.7	68.0	77.1	52.6	60.3
wgt decomp (0.2) ps	2.0	4.4	3.7	5.3	67.2	76.9	51.0	59.1	1.9	4.8	3.5	5.5	67.9	78.5	51.4	60.6
wgt decomp (0.8) ps	2.0	4.5	3.8	5.2	67.7	77.1	52.2	59.4	2.0	4.8	3.6	5.5	68.4	78.6	52.6	61.0
wgt decomp (3.2) ps	2.2	4.7	4.8	5.8	69.7	77.6	56.4	60.5	2.1	4.9	4.7	6.2	70.3	79.2	56.8	62.1
wgt A (0.2) ps	10.3	5.0	10.7	5.9	83.7	75.0	67.6	58.5	9.8	5.2	10.3	6.3	84.2	76.7	67.7	59.9
wgt A (0.8) ps	10.0	5.1	10.8	6.1	83.5	75.2	67.9	58.9	9.6	5.2	10.5	6.2	84.1	76.9	68.3	60.2
wgt A (3.2) ps	9.4	5.0	12.4	6.5	83.0	75.8	69.5	59.7	9.1	5.4	11.9	6.8	83.6	77.4	69.7	61.1
boot effect se		3.2		3.3		72.1		54.1		3.4		3.5		74.3		56.0
boot effect quant		1.0		1.0		73.8		51.0		1.2		1.2		75.8		53.3

Note: 'as': the standard error is estimated by the respective method and plugged into the asymptotic approximation for confidence intervals; 'bs': using 199 (standard) bootstrap replications, the standard error is estimated by the respective method and plugged into the t-statistic to obtain confidence intervals based on the quantile method. Exceptions are 'boot effect se', which bootstraps the effect and plugs its standard error into the asymptotic approximation for confidence intervals, and 'boot effect quant', which obtains confidence intervals based on the quantile method on the effect (rather than the t-statistic). The suffix 'ps' stands for adjustment for propensity score estimation.

Table A.2: Rejections across simulation designs for IPW

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
GMM	0.4	0.4	0.1	0.8	0.2	0.7	0.2	0.6	20.2	74.5	50.9	43.9	58.0	36.7	55.0	39.7
wgt uncond var	1.2	1.3	0.9	1.6	1.0	1.5	1.0	1.6	41.6	82.5	69.9	54.2	73.9	50.2	72.6	51.5
wgt decomp (0.2)	0.8	0.8	0.7	1.0	0.9	0.8	0.6	1.0	40.7	82.2	69.5	53.4	73.5	49.4	72.7	50.2
wgt decomp (0.8)	0.8	0.9	0.7	1.0	0.9	0.8	0.6	1.1	40.8	83.3	70.0	54.1	73.7	50.4	73.1	51.0
wgt decomp (3.2)	0.9	0.9	0.8	1.1	1.0	0.8	0.6	1.2	43.7	85.9	72.3	57.3	76.0	53.6	74.9	54.7
wgt A (0.2)	3.1	3.6	2.8	3.9	2.8	3.9	3.3	3.3	59.8	93.1	82.0	71.0	83.5	69.5	88.5	64.4
wgt A (0.8)	3.1	3.5	2.6	3.9	2.8	3.7	3.2	3.3	59.9	93.5	82.3	71.1	83.5	69.9	88.5	64.9
wgt A (3.2)	3.0	3.7	2.8	3.9	2.8	3.8	2.9	3.8	60.4	94.3	83.0	71.7	83.8	70.9	87.8	66.9
s GMM	3.0	5.1	4.0	4.2	4.7	3.4	3.9	4.2	57.3	90.8	81.8	66.4	84.0	64.2	85.0	63.2
s wgt uncond var	3.7	5.2	4.2	4.7	5.0	3.9	4.3	4.6	58.1	90.4	81.7	66.8	84.4	64.1	85.2	63.3
s wgt decomp (0.2)	3.3	5.0	4.2	4.2	5.0	3.4	4.2	4.2	59.6	91.6	82.5	68.8	85.1	66.2	87.1	64.1
s wgt decomp (0.8)	3.3	5.0	4.2	4.1	5.0	3.3	4.2	4.1	59.8	92.2	82.9	69.1	85.1	66.8	87.3	64.6
s wgt decomp (3.2)	3.3	5.0	4.2	4.1	5.1	3.3	4.2	4.1	60.7	92.6	83.3	70.0	85.7	67.6	88.1	65.2
s wgt A (0.2)	2.9	4.6	3.7	3.8	4.5	3.0	3.7	3.8	58.7	91.9	82.2	68.5	84.7	65.9	87.4	63.3
s wgt A (0.8)	2.9	4.7	3.9	3.8	4.6	3.1	3.9	3.8	59.0	92.5	82.6	68.9	84.8	66.6	87.5	63.9
s wgt A (3.2)	3.0	4.9	4.0	3.9	4.7	3.3	4.0	3.9	60.3	92.6	83.0	69.9	85.4	67.5	88.2	64.7
boot effect se	3.2	5.2	4.1	4.3	4.4	4.0	4.1	4.3	59.3	93.2	82.6	70.0	84.6	67.9	87.8	64.7
boot effect quant	1.5	4.1	3.1	2.5	3.5	2.0	2.7	2.9	57.7	92.8	81.6	68.9	83.8	66.7	89.0	61.5

Note: Prefix ‘s’ stands for standard bootstrap. All results with prefix ‘s’ are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure.

Table A.3: Rejections across simulation designs for pair matching

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
wgt uncond var	3.3	3.0	1.6	4.7	1.6	4.7	2.6	3.7	24.9	62.1	48.3	38.8	54.2	32.9	50.2	36.9
wgt decomp (0.2)	0.8	0.6	0.6	0.8	0.7	0.7	0.4	1.0	18.2	58.8	45.4	31.5	52.3	24.6	46.8	30.1
wgt decomp (0.8)	0.7	0.5	0.5	0.7	0.7	0.5	0.3	0.9	18.0	59.8	46.1	31.7	52.8	25.0	47.2	30.6
wgt decomp (3.2)	1.1	0.6	0.8	0.9	0.8	0.9	0.3	1.4	21.2	62.8	49.8	34.3	55.5	28.6	49.9	34.1
wgt A (0.2)	3.8	3.8	3.7	3.8	3.6	3.9	3.3	4.3	37.5	78.1	65.6	50.1	68.3	47.3	69.2	46.5
wgt A (0.8)	3.7	3.8	3.7	3.8	3.6	3.9	3.2	4.3	37.6	78.7	66.0	50.3	68.5	47.8	69.2	47.1
wgt A (3.2)	3.7	4.2	3.9	3.9	3.6	4.2	2.6	5.2	38.0	79.6	66.6	51.0	68.9	48.7	68.2	49.4
Abadie Imbens	6.7	7.3	5.1	8.8	4.8	9.2	7.4	6.6	40.0	80.0	66.9	53.1	69.3	50.7	71.6	48.4
wgt uncond var ps	6.8	4.4	3.1	8.1	3.4	7.8	4.6	6.6	33.3	65.7	52.9	46.1	59.9	39.1	55.1	43.9
wgt decomp (0.2) ps	3.5	1.1	1.4	3.2	1.9	2.7	1.3	3.3	28.3	63.4	50.3	41.3	58.7	32.9	53.6	38.1
wgt decomp (0.8) ps	3.4	1.3	1.4	3.3	2.0	2.8	1.2	3.5	28.3	64.4	51.2	41.6	59.2	33.5	54.0	38.7
wgt decomp (3.2) ps	4.9	1.8	2.3	4.4	2.6	4.1	1.5	5.3	34.0	68.3	56.2	46.0	62.9	39.4	58.0	44.2
wgt A (0.2) ps	12.8	8.5	8.8	12.5	9.1	12.2	9.7	11.6	51.9	83.2	73.3	61.9	76.4	58.8	77.2	58.0
wgt A (0.8) ps	12.8	8.5	9.0	12.3	9.1	12.1	9.5	11.8	52.1	84.0	73.8	62.3	76.6	59.5	77.2	58.9
wgt A (3.2) ps	12.8	9.1	9.4	12.5	9.2	12.7	8.3	13.6	52.6	85.3	74.5	63.4	77.2	60.7	76.1	61.8
Abadie Imbens ps	18.6	11.2	10.1	19.8	11.0	18.8	14.8	15.0	58.8	85.5	75.0	69.4	79.0	65.4	82.6	61.8
s wgt uncond var	6.4	7.0	5.1	8.4	5.2	8.3	6.1	7.4	35.7	71.5	62.0	45.2	65.6	41.7	60.3	46.9
s wgt decomp (0.2)	2.4	3.9	3.1	3.3	3.4	2.9	3.1	3.2	40.0	77.1	66.0	51.0	69.3	47.8	69.7	47.3
s wgt decomp (0.8)	2.3	3.6	2.8	3.1	3.2	2.7	2.9	3.0	39.8	77.2	66.1	50.9	69.5	47.5	69.9	47.1
s wgt decomp (3.2)	1.8	2.9	2.4	2.3	2.7	2.0	2.6	2.1	40.0	77.2	66.8	50.3	70.1	47.1	70.9	46.3
s wgt A (0.2)	1.7	3.0	2.1	2.6	2.3	2.4	2.4	2.3	37.5	77.7	65.1	50.1	68.7	46.5	69.7	45.5
s wgt A (0.8)	1.6	2.9	2.0	2.5	2.4	2.1	2.4	2.2	37.8	78.0	65.4	50.4	68.9	46.9	70.1	45.7
s wgt A (3.2)	1.3	2.8	2.0	2.1	2.2	1.9	2.2	1.9	37.7	78.0	65.4	50.3	69.1	46.7	70.6	45.2
s Abadie Imbens	1.0	1.7	1.5	1.3	1.3	1.4	1.5	1.2	21.3	64.5	56.2	29.6	59.1	26.7	51.5	34.3
s wgt uncond var ps	8.2	8.4	6.7	9.9	6.8	9.9	7.5	9.1	37.3	72.2	62.9	46.6	66.6	42.8	60.7	48.7
s wgt decomp (0.2) ps	4.5	5.1	4.4	5.3	5.0	4.7	4.5	5.1	42.3	78.1	67.5	52.9	70.5	49.9	70.7	49.6
s wgt decomp (0.8) ps	4.3	4.8	4.1	5.1	4.7	4.4	4.4	4.8	42.2	78.4	67.7	52.9	70.8	49.8	71.0	49.6
s wgt decomp (3.2) ps	4.4	4.2	3.9	4.6	4.3	4.2	3.9	4.6	43.1	78.8	69.1	52.7	71.7	50.2	72.4	49.5
s wgt A (0.2) ps	5.3	5.4	4.5	6.2	5.0	5.8	5.2	5.5	40.4	79.7	68.2	51.9	70.7	49.4	70.6	49.5
s wgt A (0.8) ps	5.3	5.3	4.5	6.1	5.0	5.6	5.3	5.3	40.7	80.3	68.7	52.3	71.0	50.1	71.1	50.0
s wgt A (3.2) ps	5.1	5.4	4.6	5.9	4.9	5.6	5.0	5.4	40.7	80.7	68.7	52.6	71.2	50.1	71.1	50.2
s Abadie Imbens ps	3.1	2.7	3.0	2.9	2.7	3.2	3.0	2.9	31.2	68.6	62.1	37.7	65.1	34.7	58.8	41.1
s boot effect se	1.5	2.8	2.1	2.3	2.2	2.1	2.0	2.3	24.8	73.7	57.4	41.1	61.4	37.1	60.5	38.0
s boot effect quant	0.0	0.2	0.1	0.2	0.2	0.1	0.1	0.1	21.0	78.7	56.3	43.3	62.1	37.6	67.2	32.5
w wgt uncond var	4.3	5.2	4.0	5.5	3.9	5.6	4.6	4.9	33.2	73.3	62.3	44.2	65.1	41.4	62.6	43.9
w wgt decomp (0.2)	3.3	4.0	3.8	3.6	3.7	3.7	3.5	3.8	35.0	76.2	64.1	47.1	67.1	44.2	67.7	43.6
w wgt A (0.2)	3.5	4.1	3.9	3.7	3.7	3.9	3.7	3.9	35.0	76.2	63.9	47.3	66.9	44.3	68.0	43.2
w wgt decomp (0.8)	3.3	3.8	3.7	3.4	3.6	3.5	3.5	3.6	35.1	76.4	64.4	47.1	67.1	44.4	67.7	43.7
w wgt A (0.8)	3.4	4.0	3.7	3.7	3.7	3.7	3.7	3.7	35.1	76.6	64.2	47.5	66.9	44.8	68.2	43.5
w wgt decomp (3.2)	3.3	4.1	3.7	3.6	3.7	3.6	3.5	3.8	35.3	76.9	64.6	47.6	67.6	44.6	68.2	43.9
w wgt A (3.2)	3.2	4.1	3.7	3.6	3.6	3.7	3.5	3.8	35.1	77.0	64.3	47.8	67.2	44.9	68.3	43.8
w Abadie Imbens	4.9	5.5	4.3	6.2	4.1	6.3	5.4	5.0	32.1	71.3	62.1	41.4	64.8	38.6	62.0	41.4
w wgt uncond var ps	4.6	5.5	4.5	5.6	4.5	5.6	4.7	5.4	31.1	73.3	61.8	42.7	64.5	40.0	60.4	44.1
w wgt decomp (0.2) ps	3.7	4.6	4.3	4.1	4.2	4.2	3.9	4.5	33.4	76.8	64.4	45.9	66.6	43.7	66.6	43.7
w wgt A (0.2) ps	4.6	5.3	5.0	4.9	4.8	5.1	4.7	5.2	31.1	76.6	63.6	44.1	65.8	41.8	63.9	43.8
w wgt decomp (0.8) ps	3.7	4.5	4.3	3.9	4.1	4.1	3.7	4.5	33.5	77.1	64.7	45.9	66.7	43.9	66.6	44.0
w wgt A (0.8) ps	4.5	5.2	4.8	4.8	4.9	4.8	4.7	5.0	31.1	77.5	63.9	44.7	66.0	42.6	64.1	44.5
w wgt decomp (3.2) ps	3.8	4.6	4.3	4.1	4.2	4.1	3.8	4.6	33.5	77.4	64.9	46.1	67.2	43.7	66.9	44.1
w wgt A (3.2) ps	4.3	5.4	4.9	4.8	4.8	4.9	4.5	5.2	31.2	78.0	64.1	45.1	66.2	43.0	64.3	44.9
w Abadie Imbens ps	6.1	6.9	5.7	7.2	5.4	7.5	6.6	6.4	34.6	74.1	64.1	44.5	67.0	41.7	64.1	44.6
w boot effect se	2.9	4.5	3.2	4.2	2.9	4.5	3.6	3.8	30.4	75.5	61.4	44.5	64.0	41.9	64.6	41.3
w boot effect quant	2.7	4.3	3.2	3.9	2.7	4.3	3.5	3.6	28.0	73.7	59.8	41.9	62.3	39.4	62.1	39.6

Note: The prefixes 's' and 'w' stand for the standard and wild bootstrap, respectively. All results with prefixes 's' and 'w' are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure. The suffix 'ps' stands for adjustment for propensity score estimation.

Table A.4: Rejections across simulation designs for radius matching

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
	radius matching R1.5															
wgt uncond var	2.0	1.8	1.1	2.7	1.2	2.7	1.4	2.4	29.7	71.4	58.0	43.1	63.5	37.6	58.4	42.7
wgt decomp (0.2)	0.8	0.7	0.6	0.9	0.8	0.7	0.4	1.0	26.2	69.8	56.6	39.5	62.9	33.2	57.2	38.8
wgt decomp (0.8)	0.8	0.6	0.6	0.8	0.8	0.6	0.4	1.0	26.3	70.9	57.4	39.8	63.3	33.9	57.6	39.5
wgt decomp (3.2)	0.8	0.7	0.7	0.8	0.8	0.7	0.4	1.2	28.1	74.2	59.6	42.8	65.4	37.0	59.6	42.8
wgt A (0.2)	3.4	3.8	3.3	4.0	3.3	3.9	3.5	3.8	46.8	85.6	73.6	58.8	76.3	56.1	78.2	54.2
wgt A (0.8)	3.4	3.7	3.2	3.9	3.3	3.8	3.4	3.7	46.8	86.1	73.8	59.1	76.3	56.6	77.9	55.0
wgt A (3.2)	3.3	3.8	3.4	3.8	3.2	4.0	2.8	4.3	47.2	86.8	74.3	59.6	76.7	57.2	77.2	56.8
wgt uncond var ps	6.2	3.1	2.7	6.6	3.3	6.0	3.5	5.8	42.7	76.4	64.8	54.3	71.3	47.8	66.2	52.9
wgt decomp (0.2) ps	4.3	1.6	1.9	4.0	2.7	3.2	2.0	3.9	40.4	75.2	63.7	51.9	71.1	44.5	66.3	49.3
wgt decomp (0.8) ps	4.3	1.7	2.0	4.0	2.7	3.2	1.9	4.1	40.8	76.9	65.1	52.6	71.7	45.9	67.0	50.7
wgt decomp (3.2) ps	5.2	2.2	2.5	4.9	3.3	4.0	2.0	5.4	43.9	80.2	67.7	56.4	74.5	49.7	69.0	55.2
wgt A (0.2) ps	13.3	8.7	9.4	12.5	9.7	12.3	9.8	12.1	61.6	90.0	81.9	69.8	84.6	67.0	84.4	67.2
wgt A (0.8) ps	13.2	8.7	9.5	12.3	9.6	12.2	9.6	12.2	61.6	90.4	82.2	69.9	84.6	67.4	84.2	67.8
wgt A (3.2) ps	13.1	9.2	9.9	12.3	9.7	12.6	8.8	13.5	61.9	91.0	82.8	70.1	84.9	68.1	83.5	69.4
s wgt uncond var	3.4	4.2	2.9	4.7	3.5	4.1	3.4	4.2	40.5	78.8	68.7	50.5	72.5	46.8	68.4	50.9
s wgt decomp (0.2)	1.9	3.2	2.4	2.8	2.9	2.3	2.7	2.5	43.7	82.8	70.9	55.7	74.7	51.9	74.9	51.6
s wgt decomp (0.8)	1.9	3.1	2.3	2.7	3.0	2.1	2.6	2.5	43.8	83.3	71.3	55.8	74.9	52.2	75.3	51.9
s wgt decomp (3.2)	1.6	3.1	2.3	2.4	2.8	1.9	2.5	2.2	44.5	84.1	72.1	56.4	75.9	52.6	76.3	52.2
s wgt A (0.2)	1.5	3.0	2.0	2.5	2.5	2.0	2.4	2.1	41.3	83.0	69.8	54.4	74.2	50.0	74.3	49.9
s wgt A (0.8)	1.5	3.0	1.9	2.5	2.5	1.9	2.4	2.0	41.7	83.8	70.6	54.9	74.5	51.0	74.9	50.6
s wgt A (3.2)	1.3	2.9	1.9	2.3	2.5	1.7	2.3	2.0	43.1	84.4	71.4	56.1	75.2	52.3	76.1	51.4
s wgt uncond var ps	5.5	5.4	4.4	6.5	5.0	5.9	4.8	6.1	41.9	79.7	69.9	51.7	73.1	48.4	68.2	53.3
s wgt decomp (0.2) ps	4.2	4.7	4.0	5.0	4.7	4.2	4.3	4.7	45.7	83.7	72.4	57.0	75.6	53.8	75.2	54.2
s wgt decomp (0.8) ps	4.1	4.6	3.9	4.9	4.7	4.1	4.2	4.5	45.9	84.0	72.9	57.0	75.8	54.1	75.5	54.5
s wgt decomp (3.2) ps	4.1	4.9	4.0	5.0	4.8	4.2	4.1	4.9	46.8	85.4	74.2	58.0	77.1	55.1	76.7	55.5
s wgt A (0.2) ps	4.6	5.2	4.4	5.4	4.9	4.9	4.4	5.3	42.3	84.2	72.3	54.1	75.3	51.2	72.9	53.5
s wgt A (0.8) ps	4.6	5.1	4.5	5.2	4.9	4.8	4.4	5.2	42.7	84.9	73.0	54.6	75.4	52.2	73.4	54.2
s wgt A (3.2) ps	4.8	5.5	4.9	5.4	5.1	5.2	4.5	5.8	44.4	85.6	74.1	55.9	76.3	53.7	74.6	55.4
s boot effect se	2.1	4.3	2.9	3.5	3.3	3.1	3.2	3.2	39.3	85.7	70.1	54.9	74.0	50.9	73.8	51.2
s boot effect quant	0.2	1.4	0.8	0.9	1.1	0.5	0.8	0.8	37.2	87.5	69.8	54.9	74.2	50.5	78.6	46.1
	radius matching R3															
wgt uncond var	1.7	1.5	0.9	2.2	1.0	2.1	1.1	2.0	30.9	72.8	59.4	44.3	64.7	39.0	60.4	43.3
wgt decomp (0.2)	0.8	0.7	0.6	0.9	0.7	0.7	0.5	1.0	28.2	72.0	58.5	41.7	64.2	35.9	59.7	40.5
wgt decomp (0.8)	0.8	0.6	0.6	0.8	0.8	0.6	0.4	1.0	28.3	73.2	59.4	42.0	64.6	36.8	60.2	41.2
wgt decomp (3.2)	0.9	0.8	0.7	0.9	0.9	0.7	0.4	1.2	30.2	76.2	61.4	45.1	66.8	39.7	62.1	44.4
wgt A (0.2)	3.4	3.8	3.3	3.9	3.3	3.9	3.6	3.7	48.9	87.0	74.9	61.0	77.3	58.6	80.2	55.7
wgt A (0.8)	3.4	3.8	3.2	3.9	3.3	3.8	3.5	3.7	48.9	87.3	75.2	61.0	77.4	58.9	80.0	56.3
wgt A (3.2)	3.3	3.9	3.3	3.8	3.2	3.9	2.9	4.2	49.3	88.3	75.7	61.9	77.8	59.8	79.2	58.3
wgt uncond var ps	6.0	2.7	2.7	6.0	3.3	5.4	3.2	5.5	44.5	77.9	66.5	55.8	73.0	49.4	68.3	54.1
wgt decomp (0.2) ps	4.5	1.8	2.1	4.2	2.8	3.4	2.2	4.1	42.8	77.6	66.1	54.3	72.6	47.8	68.7	51.7
wgt decomp (0.8) ps	4.5	1.9	2.1	4.2	2.9	3.4	2.1	4.2	43.1	79.0	67.2	54.9	73.2	48.8	69.1	52.9
wgt decomp (3.2) ps	5.4	2.4	2.6	5.2	3.4	4.3	2.2	5.5	46.4	82.3	69.9	58.8	75.9	52.7	71.0	57.7
wgt A (0.2) ps	13.3	8.6	9.6	12.3	9.7	12.2	9.9	12.0	63.5	91.2	83.0	71.7	85.5	69.2	86.2	68.5
wgt A (0.8) ps	13.2	8.8	9.7	12.3	9.7	12.3	9.6	12.4	63.5	91.6	83.4	71.7	85.6	69.5	86.0	69.1
wgt A (3.2) ps	13.1	9.1	10.0	12.2	9.8	12.4	8.9	13.3	63.8	92.2	84.0	72.0	85.7	70.3	85.3	70.7
s wgt uncond var	3.2	4.2	2.8	4.6	3.5	3.9	3.4	4.0	43.6	81.9	70.7	54.7	74.2	51.2	72.5	53.0
s wgt decomp (0.2)	2.1	3.4	2.4	3.1	3.0	2.5	2.8	2.8	46.2	84.7	72.4	58.5	75.9	55.0	77.3	53.7
s wgt decomp (0.8)	2.0	3.3	2.4	2.9	3.1	2.2	2.7	2.6	46.4	85.2	73.0	58.6	76.1	55.5	77.5	54.1
s wgt decomp (3.2)	1.7	3.2	2.3	2.7	2.9	2.1	2.6	2.4	47.1	86.1	73.7	59.6	77.3	56.0	78.7	54.6
s wgt A (0.2)	1.7	3.1	2.0	2.8	2.6	2.2	2.5	2.3	44.1	85.2	71.7	57.7	75.6	53.7	77.1	52.2
s wgt A (0.8)	1.6	3.1	2.0	2.7	2.6	2.1	2.5	2.2	44.5	85.8	72.3	57.9	75.8	54.4	77.3	52.9
s wgt A (3.2)	1.5	3.2	2.0	2.6	2.6	2.0	2.4	2.2	46.0	86.4	73.1	59.3	76.7	55.7	78.5	53.9
s wgt uncond var ps	5.4	5.4	4.3	6.5	5.2	5.6	4.8	6.0	44.5	82.5	71.8	55.1	74.8	52.2	71.6	55.4
s wgt decomp (0.2) ps	4.3	5.0	4.0	5.3	4.7	4.6	4.4	4.9	47.7	85.1	73.7	59.1	76.6	56.2	76.9	55.9
s wgt decomp (0.8) ps	4.3	4.9	4.0	5.2	4.7	4.5	4.3	4.9	47.9	85.9	74.4	59.4	76.9	56.9	77.4	56.5
s wgt decomp (3.2) ps	4.5	5.1	4.2	5.3	4.9	4.6	4.4	5.1	49.0	87.2	75.6	60.6	78.2	58.1	78.6	57.6
s wgt A (0.2) ps	4.8	5.4	4.5	5.7	5.0	5.2	4.7	5.6	44.6	86.2	73.8	56.9	76.6	54.2	75.1	55.6
s wgt A (0.8) ps	4.9	5.5	4.7	5.7	5.1	5.3	4.8	5.6	45.0	87.0	74.6	57.4	76.8	55.2	75.6	56.3
s wgt A (3.2) ps	5.0	5.8	5.1	5.7	5.4	5.4	4.8	6.0	46.6	87.4	75.4	58.6	77.5	56.5	76.8	57.3
s boot effect se	2.3	4.5	3.0	3.8	3.4	3.4	3.4	3.5	43.3	88.2	72.4	59.0	76.0	55.4	77.3	54.1
s boot effect quant	0.3	1.6	0.9	1.1	1.2	0.8	0.9	1.0	41.4	89.5	72.3	58.5	76.2	54.6	80.9	49.9

Note: Prefix ‘s’ stands for standard bootstrap. All results with prefix ‘s’ are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure.

Table A.5: Rejections across simulation designs for radius matching with bias correction

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
	radius matching R1.5BC															
wgt uncond var	0.8	1.0	0.6	1.2	0.8	1.1	0.7	1.2	25.7	74.0	58.3	41.4	63.3	36.5	58.7	41.0
wgt decomp (0.2)	0.6	0.6	0.4	0.8	0.6	0.6	0.4	0.8	24.6	73.1	57.3	40.4	62.6	35.1	58.2	39.5
wgt A (0.2)	3.0	4.0	2.9	4.1	2.9	4.1	3.8	3.2	44.6	87.3	73.5	58.4	75.6	56.3	77.4	54.5
wgt decomp (0.8)	0.6	0.7	0.5	0.8	0.7	0.6	0.4	0.9	24.8	74.3	58.3	40.9	63.2	35.9	58.7	40.4
wgt A (0.8)	2.9	3.9	2.8	4.0	2.9	3.9	3.7	3.2	44.6	87.7	73.8	58.5	75.7	56.6	77.2	55.1
wgt decomp (3.2)	0.7	0.8	0.5	0.9	0.7	0.7	0.4	1.0	26.7	77.1	60.1	43.7	65.2	38.6	60.4	43.3
wgt A (3.2)	3.0	4.1	2.9	4.2	3.0	4.1	3.4	3.7	45.1	88.5	74.5	59.2	76.1	57.6	76.7	56.9
wgt uncond var ps	4.3	2.2	1.9	4.6	2.5	4.0	2.4	4.1	40.0	79.7	65.7	54.1	71.9	47.8	67.5	52.3
wgt decomp (0.2) ps	3.8	1.8	1.7	4.0	2.3	3.4	2.0	3.7	39.1	79.0	65.0	53.2	71.5	46.6	67.2	51.0
wgt A (0.2) ps	12.0	8.9	8.8	12.1	8.7	12.2	10.3	10.7	59.6	91.6	82.1	69.1	84.0	67.2	83.7	67.6
wgt decomp (0.8) ps	3.9	2.0	1.7	4.1	2.4	3.4	2.0	3.8	39.5	80.4	66.1	53.7	72.1	47.8	67.7	52.2
wgt A (0.8) ps	12.0	8.9	8.8	12.0	8.7	12.1	10.0	10.8	59.6	91.8	82.3	69.1	84.0	67.4	83.5	67.9
wgt decomp (3.2) ps	4.7	2.3	2.1	4.9	2.9	4.1	2.2	4.8	42.6	83.5	68.8	57.4	74.5	51.6	69.7	56.4
wgt A (3.2) ps	12.2	9.6	9.2	12.6	9.1	12.7	9.4	12.4	60.1	92.4	83.0	69.5	84.4	68.1	83.0	69.5
s wgt uncond var	2.5	3.6	2.7	3.4	3.3	2.9	3.0	3.1	46.4	85.1	73.1	58.4	76.7	54.8	75.5	56.0
s wgt decomp (0.2)	2.4	3.5	2.6	3.2	3.2	2.6	2.9	2.9	47.0	85.7	73.7	59.0	77.3	55.5	76.6	56.1
s wgt A (0.2)	2.0	3.1	2.2	2.9	2.8	2.3	2.6	2.5	45.2	86.1	72.8	58.5	76.6	54.7	76.2	55.1
s wgt decomp (0.8)	2.4	3.4	2.6	3.1	3.2	2.5	2.9	2.8	47.1	86.1	74.1	59.2	77.4	55.9	76.8	56.5
s wgt A (0.8)	2.0	3.2	2.2	3.0	2.9	2.3	2.6	2.5	45.4	86.7	73.3	58.8	76.7	55.4	76.4	55.7
s wgt decomp (3.2)	2.3	3.5	2.6	3.2	3.3	2.5	2.9	2.9	47.7	86.9	74.5	60.0	78.1	56.4	77.3	57.3
s wgt A (3.2)	2.0	3.2	2.2	2.9	2.8	2.3	2.6	2.6	46.4	87.0	73.8	59.5	77.2	56.2	76.9	56.4
s wgt uncond var ps	5.1	5.1	4.2	5.9	5.0	5.1	4.6	5.5	48.8	85.8	74.7	59.9	77.8	56.7	75.7	58.8
s wgt decomp (0.2) ps	4.9	4.7	4.0	5.6	4.9	4.8	4.4	5.3	49.4	86.5	75.3	60.6	78.4	57.5	76.8	59.1
s wgt A (0.2) ps	5.7	5.3	4.7	6.2	5.3	5.6	5.0	5.9	46.5	86.9	75.0	58.4	77.5	55.9	74.9	58.5
s wgt decomp (0.8) ps	4.9	4.7	4.0	5.7	5.0	4.7	4.5	5.1	49.6	86.8	75.6	60.8	78.5	57.9	77.0	59.4
s wgt A (0.8) ps	5.6	5.5	4.8	6.3	5.5	5.7	5.1	6.1	46.8	87.1	75.4	58.5	77.6	56.3	75.1	58.8
s wgt decomp (3.2) ps	5.2	5.1	4.2	6.1	5.3	5.1	4.6	5.7	50.3	87.7	76.4	61.6	79.4	58.6	77.6	60.4
s wgt A (3.2) ps	5.6	5.9	5.1	6.4	5.7	5.8	5.0	6.5	47.8	87.4	76.0	59.2	78.2	57.0	75.6	59.6
s boot effect se	2.1	4.4	2.9	3.6	3.3	3.2	3.2	3.3	38.7	87.5	71.6	54.6	74.6	51.7	72.1	54.1
s boot effect quant	0.3	1.7	0.9	1.1	1.2	0.8	1.0	1.0	35.7	89.0	71.5	53.2	74.4	50.4	73.8	51.0
	radius matching R3BC															
wgt uncond var	0.7	0.9	0.6	1.0	0.7	0.9	0.6	1.0	25.9	74.5	58.7	41.7	63.5	36.9	59.3	41.1
wgt decomp (0.2)	0.5	0.6	0.4	0.7	0.5	0.6	0.4	0.8	24.9	73.7	58.0	40.5	63.0	35.6	58.8	39.7
wgt decomp (0.8)	0.5	0.6	0.5	0.7	0.6	0.5	0.4	0.8	25.1	74.7	58.7	41.1	63.5	36.3	59.3	40.5
wgt decomp (3.2)	0.6	0.7	0.5	0.8	0.7	0.6	0.3	1.0	26.9	77.6	60.6	43.9	65.5	39.0	61.0	43.5
wgt A (0.2)	2.7	3.8	2.8	3.8	2.8	3.7	3.6	2.9	44.9	87.6	73.9	58.6	75.8	56.7	78.0	54.6
wgt A (0.8)	2.7	3.6	2.7	3.6	2.8	3.5	3.5	2.8	44.9	87.8	74.1	58.6	75.8	56.9	77.9	54.9
wgt A (3.2)	2.7	3.6	2.7	3.7	2.8	3.6	3.1	3.3	45.5	88.8	75.0	59.4	76.3	58.1	77.4	57.0
wgt uncond var ps	4.1	2.1	1.9	4.3	2.5	3.7	2.2	3.9	40.5	80.1	66.2	54.4	72.4	48.2	68.0	52.6
wgt decomp (0.2) ps	3.7	1.8	1.6	3.9	2.3	3.2	1.9	3.5	39.7	79.5	65.8	53.4	72.0	47.2	67.8	51.4
wgt decomp (0.8) ps	3.8	1.9	1.7	3.9	2.4	3.2	2.0	3.6	40.0	81.0	66.8	54.1	72.5	48.4	68.4	52.6
wgt decomp (3.2) ps	4.6	2.2	2.1	4.7	2.8	4.0	2.1	4.7	43.1	84.0	69.4	57.7	74.9	52.3	70.3	56.8
wgt A (0.2) ps	11.6	8.5	8.6	11.4	8.5	11.5	9.8	10.3	60.1	91.9	82.6	69.4	84.1	67.8	84.2	67.7
wgt A (0.8) ps	11.6	8.5	8.6	11.4	8.6	11.5	9.6	10.5	60.1	92.3	82.9	69.5	84.3	68.1	84.1	68.3
wgt A (3.2) ps	11.8	9.2	9.0	12.0	8.9	12.1	9.1	11.9	60.7	92.6	83.5	69.8	84.6	68.7	83.6	69.7
s wgt uncond var	2.7	3.8	2.8	3.7	3.4	3.1	3.2	3.3	48.3	86.9	74.4	60.8	77.7	57.5	77.4	57.7
s wgt decomp (0.2)	2.6	3.6	2.7	3.5	3.3	2.9	3.1	3.1	48.9	87.4	74.8	61.5	78.1	58.2	78.3	57.9
s wgt decomp (0.8)	2.6	3.6	2.7	3.5	3.4	2.8	3.1	3.0	49.1	87.8	75.2	61.7	78.3	58.6	78.5	58.4
s wgt decomp (3.2)	2.5	3.7	2.7	3.5	3.4	2.8	3.2	3.1	49.7	88.4	75.6	62.5	79.0	59.1	79.1	59.0
s wgt A (0.2)	2.3	3.4	2.3	3.3	3.0	2.6	2.9	2.7	47.3	87.8	74.0	61.1	77.6	57.6	78.1	57.0
s wgt A (0.8)	2.3	3.4	2.4	3.3	3.0	2.7	2.9	2.8	47.5	88.2	74.4	61.3	77.7	58.1	78.3	57.5
s wgt A (3.2)	2.2	3.4	2.4	3.2	3.0	2.6	2.8	2.8	48.5	88.6	75.0	62.1	78.2	59.0	78.8	58.3
s wgt uncond var ps	5.3	5.4	4.4	6.3	5.2	5.5	4.9	5.7	50.2	87.3	75.7	61.8	78.6	58.9	77.2	60.3
s wgt decomp (0.2) ps	5.2	5.0	4.2	6.0	5.0	5.2	4.8	5.4	51.1	88.1	76.4	62.8	79.2	60.0	78.5	60.7
s wgt decomp (0.8) ps	5.2	5.1	4.1	6.1	5.2	5.1	4.8	5.5	51.2	88.3	76.7	62.8	79.3	60.2	78.6	60.9
s wgt decomp (3.2) ps	5.5	5.5	4.5	6.6	5.5	5.5	4.9	6.1	51.9	89.2	77.4	63.7	80.2	60.9	79.1	62.0
s wgt A (0.2) ps	5.8	5.7	5.0	6.6	5.5	6.0	5.2	6.3	48.2	88.2	76.1	60.3	78.3	58.1	76.6	59.8
s wgt A (0.8) ps	5.8	5.6	5.0	6.5	5.6	5.9	5.2	6.2	48.4	88.7	76.4	60.6	78.5	58.6	76.9	60.2
s wgt A (3.2) ps	5.9	6.2	5.4	6.7	5.8	6.3	5.3	6.8	49.4	89.1	77.1	61.4	79.1	59.4	77.4	61.1
s boot effect se	2.3	4.6	3.0	3.9	3.6	3.4	3.4	3.5	41.2	89.1	73.1	57.2	75.8	54.5	74.3	56.0
s boot effect quant	0.4	1.9	1.0	1.3	1.3	1.0	1.2	1.2	38.5	90.6	73.2	55.9	75.8	53.4	75.8	53.3

Note: Prefix ‘s’ stands for standard bootstrap. All results with prefix ‘s’ are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure.

Table A.6: Coverage and rejections of bootstrap methods for IPW and pair matching

	coverage (all DGPs)						rejection (effect homogeneity)					
	no smoothing			smoothing			no smoothing			smoothing		
	49	99	199	49	99	199	49	99	199	49	99	199
	IPW											
GMM	95.8	95.6	95.4	95.8	95.6	95.4	5.0	4.1	4.1	4.1	4.1	4.1
wgt uncond var	95.3	95.0	94.8	95.3	95.0	94.8	5.4	4.6	4.5	4.4	4.5	4.4
wgt decomp (0.2)	95.4	95.2	95.1	95.4	95.2	95.1	5.2	4.3	4.2	4.2	4.2	4.2
wgt decomp (0.8)	95.3	95.3	95.1	95.3	95.3	95.1	5.2	4.2	4.2	4.1	4.2	4.2
wgt decomp (3.2)	95.2	95.1	95.2	95.2	95.1	95.2	5.2	4.3	4.2	4.2	4.2	4.1
wgt A (0.2)	96.0	96.0	95.9	96.0	96.0	95.9	4.7	3.7	3.7	3.7	3.7	3.7
wgt A (0.8)	95.8	95.8	95.8	95.8	95.8	95.8	4.8	3.9	3.8	3.8	3.9	3.8
wgt A (3.2)	95.4	95.4	95.4	95.4	95.4	95.4	5.0	4.1	4.0	4.0	4.0	4.0
boot effect se	96.3	96.3	96.4				4.6	4.3	4.2			
boot effect quant	96.3	96.3	96.3				5.1	3.6	2.8			
	pair matching											
s wgt uncond var	89.8	89.2	88.8	89.8	89.2	88.8	7.5	6.7	6.8	6.6	6.6	6.7
s wgt decomp (0.2)	91.8	91.6	91.5	91.8	91.6	91.5	3.9	3.3	3.2	3.1	3.2	3.2
s wgt decomp (0.8)	91.8	91.6	91.5	91.8	91.6	91.5	3.7	2.9	2.9	2.9	2.9	2.9
s wgt decomp (3.2)	91.7	91.6	91.5	91.7	91.6	91.5	3.1	2.4	2.4	2.4	2.4	2.4
s wgt A (0.2)	92.9	92.8	92.7	92.9	92.8	92.7	3.0	2.4	2.4	2.4	2.4	2.3
s wgt A (0.8)	92.6	92.5	92.4	92.6	92.5	92.4	3.0	2.3	2.3	2.3	2.2	2.2
s wgt A (3.2)	92.1	92.0	92.0	92.1	92.0	92.0	2.7	2.1	2.1	2.0	2.1	2.0
s Abadie Imbens	98.0	97.9	97.9	98.0	97.9	97.9	2.0	1.5	1.4	1.5	1.5	1.4
s wgt uncond var ps	89.1	87.7	87.2	89.1	87.7	87.2	9.0	8.4	8.4	8.0	8.3	8.3
s wgt decomp (0.2) ps	91.7	90.8	90.5	91.7	90.8	90.5	5.6	4.8	4.9	4.5	4.7	4.8
s wgt decomp (0.8) ps	91.7	90.8	90.4	91.7	90.8	90.4	5.4	4.5	4.6	4.5	4.4	4.5
s wgt decomp (3.2) ps	91.5	90.7	90.4	91.5	90.7	90.4	5.2	4.2	4.3	4.1	4.1	4.2
s wgt A (0.2) ps	93.3	92.2	91.6	93.3	92.2	91.6	6.3	5.4	5.4	5.3	5.3	5.4
s wgt A (0.8) ps	92.9	91.8	91.3	92.9	91.8	91.3	6.3	5.2	5.3	5.3	5.2	5.3
s wgt A (3.2) ps	92.4	91.4	90.9	92.4	91.4	90.9	6.2	5.2	5.2	5.2	5.2	5.2
s Abadie Imbens ps	96.6	96.5	96.4	96.6	96.5	96.4	3.7	3.0	3.0	3.1	3.0	2.9
s boot effect se	97.6	97.7	97.7				2.5	2.2	2.2			
s boot effect quant	97.6	97.7	97.7				0.5	0.2	0.1			
w wgt uncond var	97.7	97.7	97.6	97.7	97.7	97.6	5.6	4.8	4.7	4.7	4.7	4.8
w wgt decomp (0.2)	96.5	96.5	96.4	96.5	96.5	96.4	4.8	3.7	3.7	3.7	3.7	3.7
w wgt decomp (0.8)	96.4	96.4	96.3	96.4	96.4	96.3	4.7	3.8	3.6	3.7	3.7	3.6
w wgt decomp (3.2)	96.2	96.3	96.2	96.2	96.3	96.2	4.8	3.7	3.7	3.7	3.7	3.6
w wgt A (0.2)	96.4	96.4	96.3	96.4	96.4	96.3	4.8	3.9	3.8	3.9	3.8	3.8
w wgt A (0.8)	96.3	96.3	96.3	96.3	96.3	96.3	4.9	3.9	3.7	3.8	3.8	3.7
w wgt A (3.2)	96.2	96.2	96.1	96.2	96.2	96.1	4.8	3.8	3.6	3.7	3.7	3.7
w Abadie Imbens	97.8	97.8	97.7	97.8	97.8	97.7	6.0	5.2	5.2	5.1	5.2	5.2
w wgt uncond var ps	97.9	97.9	97.8	97.9	97.9	97.8	6.1	5.0	5.1	5.2	5.1	5.0
w wgt decomp (0.2) ps	96.7	96.6	96.5	96.7	96.6	96.5	5.3	4.1	4.2	4.3	4.2	4.2
w wgt decomp (0.8) ps	96.6	96.5	96.5	96.6	96.5	96.5	5.2	4.2	4.1	4.2	4.2	4.1
w wgt decomp (3.2) ps	96.5	96.5	96.4	96.5	96.5	96.4	5.5	4.4	4.2	4.3	4.3	4.2
w wgt A (0.2) ps	97.0	96.9	96.8	97.0	96.9	96.8	6.1	5.0	5.0	5.1	5.0	4.9
w wgt A (0.8) ps	96.9	96.9	96.7	96.9	96.9	96.7	6.0	4.8	4.8	5.0	4.9	4.9
w wgt A (3.2) ps	96.8	96.8	96.6	96.8	96.8	96.6	6.0	4.9	4.9	4.9	4.8	4.9
w Abadie Imbens ps	97.1	97.1	97.1	97.1	97.1	97.1	7.4	6.5	6.5	6.3	6.4	6.5
w boot effect se	96.8	96.9	96.9				4.3	3.9	3.7			
w boot effect quant	96.9	97.0	97.0				6.5	4.5	3.5			

Note: The prefixes 's' and 'w' for pair matching stand for the standard and wild bootstrap, respectively. The suffix 'ps' stands for adjustment for propensity score estimation.

Table A.7: Coverage and rejections of bootstrap methods for radius matching

	coverage (all DGPs)						rejection (effect homogeneity)					
	no smoothing			smoothing			no smoothing			smoothing		
	49	99	199	49	99	199	49	99	199	49	99	199
	radius matching R1.5											
wgt uncond var	94.5	93.9	93.5	94.5	93.9	93.5	4.5	3.8	3.8	3.8	3.7	3.8
wgt decomp (0.2)	95.0	94.8	94.7	95.0	94.8	94.7	3.4	2.6	2.6	2.7	2.6	2.6
wgt decomp (0.8)	94.9	94.7	94.6	94.9	94.7	94.6	3.2	2.5	2.5	2.6	2.5	2.5
wgt decomp (3.2)	94.7	94.6	94.5	94.7	94.6	94.5	3.1	2.4	2.3	2.4	2.5	2.3
wgt A (0.2)	95.7	95.7	95.7	95.7	95.7	95.7	3.0	2.3	2.2	2.3	2.2	2.2
wgt A (0.8)	95.5	95.4	95.4	95.5	95.4	95.4	2.9	2.3	2.2	2.3	2.2	2.2
wgt A (3.2)	95.0	94.9	94.9	95.0	94.9	94.9	2.8	2.1	2.1	2.2	2.1	2.1
wgt uncond var ps	94.5	92.8	92.2	94.5	92.8	92.2	6.3	5.5	5.5	5.4	5.5	5.4
wgt decomp (0.2) ps	95.0	94.1	93.7	95.0	94.1	93.7	5.2	4.5	4.5	4.3	4.3	4.4
wgt decomp (0.8) ps	95.0	94.0	93.7	95.0	94.0	93.7	5.2	4.4	4.4	4.3	4.4	4.3
wgt decomp (3.2) ps	94.4	93.8	93.5	94.4	93.8	93.5	5.4	4.5	4.6	4.4	4.4	4.5
wgt A (0.2) ps	95.8	95.2	94.9	95.8	95.2	94.9	5.9	5.0	4.9	5.0	4.9	4.9
wgt A (0.8) ps	95.4	94.8	94.5	95.4	94.8	94.5	5.9	5.0	4.9	5.1	5.0	4.8
wgt A (3.2) ps	94.9	94.2	94.0	94.9	94.2	94.0	6.2	5.3	5.2	5.3	5.3	5.1
boot effect se	96.9	96.9	96.9				3.6	3.3	3.2			
boot effect quant	96.9	96.9	97.0				2.0	1.3	0.8			
	radius matching R3											
wgt uncond var	94.7	94.2	93.9	94.7	94.2	93.9	4.4	3.7	3.7	3.7	3.7	3.7
wgt decomp (0.2)	95.0	94.8	94.6	95.0	94.8	94.6	3.5	2.8	2.8	2.8	2.7	2.7
wgt decomp (0.8)	94.9	94.7	94.6	94.9	94.7	94.6	3.4	2.7	2.6	2.7	2.7	2.7
wgt decomp (3.2)	94.6	94.6	94.5	94.6	94.6	94.5	3.3	2.6	2.5	2.6	2.6	2.5
wgt A (0.2)	95.7	95.6	95.6	95.7	95.6	95.6	3.1	2.4	2.4	2.5	2.4	2.4
wgt A (0.8)	95.5	95.4	95.3	95.5	95.4	95.3	3.1	2.4	2.4	2.5	2.4	2.3
wgt A (3.2)	95.0	94.9	94.9	95.0	94.9	94.9	3.1	2.3	2.3	2.3	2.3	2.3
wgt uncond var ps	95.0	93.3	92.7	95.0	93.3	92.7	6.3	5.4	5.4	5.3	5.4	5.4
wgt decomp (0.2) ps	95.3	94.2	93.7	95.3	94.2	93.7	5.5	4.6	4.7	4.6	4.5	4.6
wgt decomp (0.8) ps	95.1	94.2	93.7	95.1	94.2	93.7	5.5	4.5	4.6	4.6	4.5	4.6
wgt A (3.2) ps	95.1	94.4	94.0	95.1	94.4	94.0	6.3	5.4	5.4	5.4	5.4	5.4
wgt A (0.2) ps	95.8	95.2	94.9	95.8	95.2	94.9	6.1	5.1	5.1	5.2	5.1	5.1
wgt A (0.8) ps	95.5	94.8	94.6	95.5	94.8	94.6	6.1	5.2	5.2	5.2	5.2	5.2
wgt decomp (3.2) ps	94.6	93.9	93.6	94.6	93.9	93.6	5.6	4.7	4.8	4.8	4.7	4.7
boot effect se	96.6	96.6	96.7				3.9	3.6	3.4			
boot effect quant	96.7	96.7	96.6				2.3	1.4	1.0			
	radius matching with bias correction R1.5BC											
wgt uncond var	94.6	94.4	94.3	94.6	94.4	94.3	3.8	3.2	3.1	3.1	3.1	3.0
wgt decomp (0.2)	94.4	94.4	94.2	94.4	94.4	94.2	3.7	2.9	2.9	3.0	2.9	2.9
wgt decomp (0.8)	94.4	94.3	94.2	94.4	94.3	94.2	3.7	2.9	2.9	3.0	2.9	2.8
wgt decomp (3.2)	94.1	94.0	94.0	94.1	94.0	94.0	3.8	2.9	2.9	3.0	2.9	2.9
wgt A (0.2)	94.9	94.9	94.9	94.9	94.9	94.9	3.4	2.6	2.6	2.7	2.6	2.5
wgt A (0.8)	94.7	94.8	94.7	94.7	94.8	94.7	3.3	2.6	2.6	2.7	2.6	2.6
wgt A (3.2)	94.4	94.4	94.4	94.4	94.4	94.4	3.5	2.6	2.6	2.7	2.6	2.6
wgt uncond var ps	94.8	93.8	93.3	94.8	93.8	93.3	6.1	5.1	5.1	5.1	5.1	5.0
wgt decomp (0.2) ps	94.6	93.6	93.3	94.6	93.6	93.3	6.0	4.9	4.8	4.9	4.9	4.8
wgt decomp (0.8) ps	94.5	93.6	93.2	94.5	93.6	93.2	5.9	4.9	4.9	5.0	4.9	4.8
wgt decomp (3.2) ps	94.1	93.3	93.0	94.1	93.3	93.0	6.4	5.3	5.2	5.3	5.2	5.1
wgt A (0.2) ps	95.1	94.3	94.0	95.1	94.3	94.0	6.5	5.6	5.5	5.7	5.6	5.5
wgt A (0.8) ps	94.9	94.2	93.8	94.9	94.2	93.8	6.7	5.6	5.6	5.7	5.6	5.6
wgt A (3.2) ps	94.6	93.8	93.4	94.6	93.8	93.4	7.0	5.8	5.8	6.0	5.8	5.7
boot effect se	96.8	96.9	96.9				3.7	3.4	3.3			
boot effect quant	96.7	96.8	96.9				2.3	1.5	1.0			
	radius matching with bias correction R3BC											
wgt uncond var	94.5	94.3	94.2	94.5	94.3	94.2	4.0	3.3	3.3	3.4	3.3	3.2
wgt decomp (0.2)	94.3	94.2	94.1	94.3	94.2	94.1	4.0	3.1	3.1	3.2	3.1	3.1
wgt decomp (0.8)	94.2	94.2	94.1	94.2	94.2	94.1	4.0	3.2	3.1	3.2	3.1	3.1
wgt decomp (3.2)	93.9	93.9	93.9	93.9	93.9	93.9	4.0	3.1	3.1	3.2	3.1	3.1
wgt A (0.2)	94.7	94.8	94.7	94.7	94.8	94.7	3.6	2.9	2.8	3.0	2.8	2.8
wgt A (0.8)	94.6	94.6	94.6	94.6	94.6	94.6	3.7	2.9	2.9	2.9	2.9	2.8
wgt A (3.2)	94.2	94.2	94.3	94.2	94.2	94.3	3.7	2.9	2.8	3.0	2.8	2.8
wgt uncond var ps	94.8	93.6	93.2	94.8	93.6	93.2	6.3	5.3	5.4	5.3	5.3	5.3
wgt decomp (0.2) ps	94.6	93.5	93.2	94.6	93.5	93.2	6.2	5.1	5.1	5.3	5.2	5.1
wgt decomp (0.8) ps	94.5	93.5	93.1	94.5	93.5	93.1	6.1	5.1	5.1	5.1	5.1	5.1
wgt decomp (3.2) ps	94.0	93.2	92.8	94.0	93.2	92.8	6.6	5.5	5.6	5.6	5.4	5.5
wgt A (0.2) ps	95.1	94.3	93.9	95.1	94.3	93.9	6.8	5.8	5.7	5.8	5.7	5.8
wgt A (0.8) ps	94.9	94.1	93.7	94.9	94.1	93.7	7.0	5.8	5.7	5.9	5.7	5.7
wgt A (3.2) ps	94.7	93.7	93.3	94.7	93.7	93.3	7.4	6.2	6.1	6.2	6.1	6.1
boot effect se	96.6	96.6	96.7				4.0	3.6	3.5			
boot effect quant	96.5	96.7	96.7				2.5	1.7	1.2			

Note: The suffix ‘ps’ stands for adjustment for propensity score estimation.

Table A.8: Coverage probabilities

IPW	homogeneity						heterogeneity					
	binary			continuous			binary			continuous		
	as	bs		as	bs		as	bs		as	bs	
GMM	99.9	95.4	95.4	99.7	95.1	95.1	100.0	95.6	95.6	99.8	95.7	95.7
wgt uncond var	99.6	94.8	94.8	99.3	94.5	94.5	99.7	94.9	94.9	99.3	95.0	95.0
wgt decomp (0.2)	99.6	94.8	94.8	99.3	95.1	95.1	99.7	95.1	95.1	99.3	95.6	95.6
wgt decomp (0.8)	99.6	94.8	94.8	99.3	95.1	95.1	99.7	95.1	95.1	99.3	95.6	95.6
wgt decomp (3.2)	99.5	94.9	94.9	98.9	95.1	95.1	99.6	95.1	95.1	98.8	95.6	95.6
wgt A (0.2)	97.0	95.7	95.7	97.0	95.8	95.8	96.8	96.1	96.1	97.1	96.2	96.2
wgt A (0.8)	97.0	95.5	95.5	96.9	95.6	95.6	96.9	95.9	95.9	97.1	96.1	96.1
wgt A (3.2)	97.2	95.2	95.2	96.3	95.2	95.2	97.1	95.5	95.5	96.4	95.6	95.6
boot effect se		96.2			96.1			96.6			96.6	
boot effect quant		96.1			96.1			96.6			96.5	
pair matching	homogeneity						heterogeneity					
	binary			continuous			binary			continuous		
	as	bs	wbs	as	bs	wbs	as	bs	wbs	as	bs	wbs
wgt uncond var	99.8	89.2	97.9	99.0	88.3	97.4	99.7	89.3	97.7	98.7	88.3	97.4
wgt decomp (0.2)	99.8	90.6	96.7	99.1	92.3	96.4	99.8	90.7	96.4	98.8	92.4	96.1
wgt decomp (0.8)	99.8	90.7	96.6	99.0	92.0	96.3	99.7	90.8	96.3	98.7	92.3	96.1
wgt decomp (3.2)	99.7	90.8	96.6	98.2	91.9	96.1	99.6	91.1	96.2	98.0	92.2	95.9
wgt A (0.2)	96.4	92.3	96.6	95.4	92.8	96.4	96.0	92.4	96.3	95.3	93.2	96.1
wgt A (0.8)	96.5	92.0	96.5	95.2	92.7	96.3	96.0	92.0	96.2	95.1	92.9	96.1
wgt A (3.2)	96.8	91.6	96.5	94.2	92.3	96.1	96.3	91.4	96.1	94.1	92.6	95.9
Abadie Imbens	95.5	97.9	97.9	95.2	97.8	97.6	95.1	98.1	97.6	95.2	97.9	97.6
wgt uncond var ps	99.4	87.7	98.1	97.6	86.3	97.5	99.4	88.0	98.1	97.6	86.6	97.5
wgt decomp (0.2) ps	99.4	89.5	96.8	97.7	91.2	96.4	99.4	89.7	96.7	97.6	91.5	96.3
wgt decomp (0.8) ps	99.4	89.6	96.8	97.5	91.0	96.3	99.3	89.7	96.6	97.5	91.3	96.2
wgt decomp (3.2) ps	99.0	89.8	96.7	95.7	90.8	96.2	98.9	90.1	96.5	95.8	91.1	96.1
wgt A (0.2) ps	92.0	91.4	97.0	89.8	91.6	96.5	92.2	91.7	97.2	90.2	91.8	96.5
wgt A (0.8) ps	92.0	90.9	97.0	89.4	91.4	96.4	92.3	91.2	97.1	90.0	91.7	96.4
wgt A (3.2) ps	92.9	90.4	96.9	87.9	91.0	96.2	93.0	90.7	97.1	88.4	91.4	96.3
Abadie Imbens ps	88.2	96.3	97.1	87.3	96.0	96.8	88.5	96.9	97.4	88.1	96.5	96.9
boot effect se		97.9	97.2		97.5	96.8		98.0	96.8		97.4	96.8
boot effect quant		97.9	97.2		97.5	97.0		98.0	96.9		97.6	96.9
radius matching	homogeneity						heterogeneity					
	binary			continuous			binary			continuous		
	as	bs		as	bs		as	bs		as	bs	
wgt uncond var	99.7	93.9		99.2	93.9		99.7	94.1		99.2	94.0	
wgt decomp (0.2)	99.7	94.1		99.3	94.6		99.7	94.3		99.2	94.7	
wgt decomp (0.8)	99.7	94.0		99.2	94.6		99.7	94.3		99.2	94.6	
wgt decomp (3.2)	99.6	94.0		98.7	94.3		99.6	94.3		98.7	94.3	
wgt A (0.2)	96.5	95.0		96.8	95.2		96.3	95.3		96.6	95.4	
wgt A (0.8)	96.5	94.8		96.6	95.1		96.4	95.0		96.5	95.1	
wgt A (3.2)	96.7	94.5		96.0	94.6		96.6	94.8		95.9	94.7	
wgt uncond var ps	99.0	93.0		97.5	92.5		99.1	93.2		97.7	92.8	
wgt decomp (0.2) ps	99.0	93.1		97.6	93.6		99.1	93.4		97.7	93.8	
wgt decomp (0.8) ps	99.0	93.0		97.5	93.5		99.0	93.4		97.5	93.7	
wgt decomp (3.2) ps	98.6	93.1		96.3	93.0		98.8	93.5		96.4	93.3	
wgt A (0.2) ps	91.8	94.2		91.2	94.2		91.9	94.7		91.5	94.4	
wgt A (0.8) ps	91.9	94.0		90.9	93.9		92.0	94.5		91.2	94.2	
wgt A (3.2) ps	92.3	93.6		89.6	93.3		92.5	94.1		89.8	93.6	
boot effect se		96.8			96.5			97.0			96.8	
boot effect quant		96.7			96.6			97.0			96.8	

Note: 'as': the standard error is estimated by the respective method and plugged into the asymptotic approximation for confidence intervals; 'bs': using 199 (standard) bootstrap replications (without smoothing), the standard error is estimated by the respective method and plugged into the t-statistic to obtain confidence intervals based on the quantile method. Exceptions are 'boot effect se', which bootstraps the effect and plugs its standard error into the asymptotic approximation for confidence intervals, and 'boot effect quant', which obtains confidence intervals based on the quantile method on the effect (rather than the t-statistic); 'wbs': wild bootstrap (without smoothing) rather than the standard bootstrap is used for the respective method. The suffix 'ps' stands for adjustment for propensity score estimation. The results for radius matching are averages over all 4 radius matching algorithms (R1.5, R3, R1.5BC, R3BC).

Table A.9: Coverage across simulation designs for IPW

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
GMM	99.8	99.9	99.9	99.8	99.9	99.8	99.9	99.7	99.9	99.9	99.9	99.8	99.9	99.9	100.0	99.8
wgt uncond var	99.5	99.4	99.4	99.4	99.3	99.6	99.6	99.3	99.4	99.5	99.4	99.5	99.3	99.6	99.7	99.3
wgt decomp (0.2)	99.5	99.5	99.5	99.4	99.3	99.6	99.6	99.3	99.4	99.5	99.4	99.5	99.3	99.6	99.7	99.3
wgt decomp (0.8)	99.4	99.4	99.4	99.4	99.3	99.6	99.6	99.3	99.4	99.5	99.4	99.5	99.3	99.6	99.7	99.2
wgt decomp (3.2)	99.2	99.2	99.2	99.1	99.0	99.4	99.5	98.9	99.1	99.3	99.2	99.2	99.2	99.2	99.6	98.8
wgt A (0.2)	97.3	96.7	97.3	96.7	97.5	96.5	97.0	97.0	96.9	97.0	97.1	96.9	97.4	96.5	96.8	97.1
wgt A (0.8)	97.3	96.7	97.3	96.7	97.5	96.5	97.0	96.9	96.9	97.0	97.1	96.9	97.4	96.5	96.9	97.1
wgt A (3.2)	97.1	96.4	97.2	96.3	97.3	96.2	97.2	96.3	96.7	96.8	96.9	96.6	97.3	96.2	97.1	96.4
s GMM	96.2	94.3	95.5	95.0	94.9	95.6	95.4	95.1	96.2	95.1	95.7	95.6	95.4	95.9	95.6	95.7
s wgt uncond var	95.2	94.0	95.0	94.3	94.3	95.0	94.8	94.5	95.3	94.7	95.0	94.9	94.8	95.2	94.9	95.0
s wgt decomp (0.2)	95.6	94.3	95.0	94.8	94.5	95.4	94.8	95.1	95.5	95.2	95.2	95.5	94.9	95.8	95.1	95.6
s wgt decomp (0.8)	95.5	94.3	95.0	94.9	94.5	95.4	94.8	95.1	95.5	95.1	95.2	95.5	94.9	95.7	95.1	95.6
s wgt decomp (3.2)	95.6	94.4	95.0	95.0	94.5	95.5	94.9	95.1	95.4	95.2	95.1	95.5	94.8	95.8	95.1	95.6
s wgt A (0.2)	96.4	95.1	95.7	95.8	95.3	96.2	95.7	95.8	96.3	96.0	95.9	96.4	95.7	96.6	96.1	96.2
s wgt A (0.8)	96.3	94.8	95.5	95.6	95.1	96.0	95.5	95.6	96.2	95.7	95.7	96.2	95.6	96.4	95.9	96.1
s wgt A (3.2)	95.9	94.5	95.3	95.1	94.8	95.6	95.2	95.2	95.7	95.4	95.4	95.7	95.2	95.9	95.5	95.6
boot effect se	97.1	95.2	96.1	96.2	95.9	96.4	96.2	96.1	97.1	96.1	96.4	96.8	96.4	96.7	96.6	96.6
boot effect quant	97.1	95.2	96.1	96.2	95.8	96.4	96.1	96.1	97.0	96.1	96.3	96.8	96.4	96.7	96.6	96.5

Note: Prefix ‘s’ stands for standard bootstrap. All results with prefix ‘s’ are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure.

Table A.10: Coverage across simulation designs for pair matching

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
wgt uncond var	99.3	99.4	99.5	99.2	99.5	99.3	99.8	99.0	99.2	99.2	99.4	99.1	99.4	99.1	99.7	98.7
wgt decomp (0.2)	99.3	99.5	99.5	99.3	99.5	99.3	99.8	99.1	99.2	99.3	99.4	99.1	99.4	99.1	99.8	98.8
wgt decomp (0.8)	99.3	99.4	99.5	99.3	99.5	99.3	99.8	99.0	99.2	99.2	99.4	99.1	99.4	99.1	99.7	98.7
wgt decomp (3.2)	98.8	99.1	99.1	98.8	99.1	98.7	99.7	98.2	98.7	99.0	99.0	98.7	99.0	98.6	99.6	98.0
wgt A (0.2)	96.1	95.8	96.3	95.6	96.5	95.3	96.4	95.4	95.7	95.6	96.0	95.2	96.2	95.1	96.0	95.3
wgt A (0.8)	96.0	95.6	96.2	95.5	96.5	95.2	96.5	95.2	95.6	95.5	96.0	95.2	96.2	95.0	96.0	95.1
wgt A (3.2)	95.7	95.2	95.9	95.1	96.3	94.6	96.8	94.2	95.4	95.1	95.6	94.8	95.9	94.5	96.3	94.1
Abadie Imbens	95.7	95.1	95.9	94.8	96.1	94.6	95.5	95.2	95.1	95.2	95.7	94.6	95.9	94.4	95.1	95.2
wgt uncond var ps	98.1	98.9	99.0	98.1	98.8	98.3	99.4	97.6	98.1	98.9	99.0	98.0	98.7	98.2	99.4	97.6
wgt decomp (0.2) ps	98.1	99.1	99.0	98.1	98.7	98.4	99.4	97.7	98.0	99.0	99.0	98.0	98.7	98.3	99.4	97.6
wgt decomp (0.8) ps	98.0	98.8	98.9	98.0	98.6	98.2	99.4	97.5	98.0	98.8	98.9	97.9	98.6	98.2	99.3	97.5
wgt decomp (3.2) ps	96.4	98.2	97.9	96.8	97.9	96.7	99.0	95.7	96.3	98.4	98.1	96.7	98.0	96.8	98.9	95.8
wgt A (0.2) ps	89.9	91.9	92.1	89.6	92.2	89.6	92.0	89.8	89.5	92.9	92.4	90.0	92.3	90.1	92.2	90.2
wgt A (0.8) ps	89.8	91.7	92.0	89.5	92.1	89.4	92.0	89.4	89.5	92.8	92.3	90.0	92.3	89.9	92.3	90.0
wgt A (3.2) ps	89.5	91.3	91.6	89.2	91.9	88.9	92.9	87.9	89.0	92.3	91.8	89.5	91.9	89.4	93.0	88.4
Abadie Imbens ps	85.1	90.4	91.1	84.3	90.2	85.2	88.2	87.3	84.9	91.7	91.4	85.2	90.7	85.9	88.5	88.1
s wgt uncond var	89.2	88.3	90.7	86.9	90.4	87.2	89.2	88.3	88.9	88.8	90.6	87.0	90.4	87.2	89.3	88.3
s wgt decomp (0.2)	92.6	90.3	92.0	90.9	91.6	91.3	90.6	92.3	92.3	90.8	92.1	90.9	91.7	91.3	90.7	92.4
s wgt decomp (0.8)	92.7	90.0	91.9	90.8	91.6	91.1	90.7	92.0	92.4	90.8	92.2	91.0	91.9	91.3	90.8	92.3
s wgt decomp (3.2)	92.3	90.4	91.7	91.0	91.7	91.1	90.8	91.9	92.0	91.3	92.1	91.2	92.1	91.2	91.1	92.2
s wgt A (0.2)	93.8	91.3	93.4	91.7	93.1	92.0	92.3	92.8	93.4	92.2	93.6	91.9	93.3	92.3	92.4	93.2
s wgt A (0.8)	93.6	91.0	93.1	91.6	93.1	91.6	92.0	92.7	93.2	91.7	93.2	91.7	93.3	91.6	92.0	92.9
s wgt A (3.2)	93.3	90.6	92.7	91.1	92.7	91.1	91.6	92.3	92.7	91.3	92.8	91.2	92.8	91.2	91.4	92.6
s Abadie Imbens	98.5	97.2	96.8	98.9	96.7	99.0	97.9	97.8	98.4	97.5	97.0	98.9	96.9	99.1	98.1	97.9
s wgt uncond var ps	87.2	86.8	88.9	85.1	88.6	85.4	87.7	86.3	87.2	87.5	89.3	85.4	89.0	85.7	88.0	86.6
s wgt decomp (0.2) ps	91.8	88.9	90.6	90.1	90.3	90.5	89.5	91.2	91.5	89.7	91.0	90.2	90.5	90.7	89.7	91.5
s wgt decomp (0.8) ps	91.9	88.8	90.6	90.0	90.3	90.4	89.6	91.0	91.6	89.4	91.0	90.0	90.6	90.4	89.7	91.3
s wgt decomp (3.2) ps	91.4	89.1	90.2	90.4	90.2	90.3	89.8	90.8	91.1	90.1	90.7	90.5	90.6	90.6	90.1	91.1
s wgt A (0.2) ps	93.6	89.4	91.3	91.8	91.4	91.6	91.4	91.6	93.5	90.0	91.8	91.7	91.7	91.8	91.7	91.8
s wgt A (0.8) ps	93.3	89.0	90.9	91.4	91.2	91.1	90.9	91.4	93.2	89.6	91.4	91.5	91.6	91.2	91.2	91.7
s wgt A (3.2) ps	92.9	88.5	90.3	91.0	90.8	90.6	90.4	91.0	92.7	89.4	91.1	91.0	91.2	90.8	90.7	91.4
s Abadie Imbens ps	96.6	95.6	94.4	97.8	94.5	97.7	96.3	96.0	96.9	96.6	95.4	98.0	95.3	98.2	96.9	96.5
s boot effect se	98.4	97.0	97.9	97.5	97.8	97.5	97.9	97.5	98.3	97.2	98.1	97.3	97.9	97.5	98.0	97.4
s boot effect quant	98.4	97.1	98.0	97.5	97.8	97.6	97.9	97.5	98.2	97.3	98.1	97.4	97.9	97.6	98.0	97.6
w wgt uncond var	98.1	97.2	97.2	98.1	97.2	98.1	97.9	97.4	97.9	97.2	97.1	97.9	97.0	98.0	97.7	97.4
w wgt decomp (0.2)	96.8	96.3	96.6	96.5	96.6	96.5	96.7	96.4	96.5	96.0	96.3	96.2	96.3	96.2	96.4	96.1
w wgt decomp (0.8)	96.8	96.1	96.5	96.5	96.6	96.4	96.6	96.3	96.5	95.9	96.2	96.2	96.3	96.1	96.3	96.1
w wgt decomp (3.2)	96.6	96.1	96.4	96.3	96.4	96.2	96.6	96.1	96.2	95.9	96.1	96.0	96.2	96.0	96.2	95.9
w wgt A (0.2)	96.7	96.3	96.6	96.4	96.6	96.4	96.6	96.4	96.3	96.1	96.3	96.1	96.4	96.0	96.3	96.1
w wgt A (0.8)	96.6	96.2	96.5	96.3	96.6	96.2	96.5	96.3	96.3	96.0	96.2	96.0	96.4	95.9	96.2	96.1
w wgt A (3.2)	96.5	96.0	96.4	96.1	96.5	96.1	96.5	96.1	96.2	95.8	96.2	95.9	96.2	95.8	96.1	95.9
w Abadie Imbens	97.8	97.7	97.3	98.3	97.2	98.3	97.9	97.6	97.6	97.6	97.1	98.1	97.0	98.2	97.6	97.6
w wgt uncond var ps	98.6	97.0	97.2	98.5	97.0	98.6	98.1	97.5	98.5	97.1	97.3	98.3	97.1	98.5	98.1	97.5
w wgt decomp (0.2) ps	97.4	95.8	96.2	97.0	96.4	96.8	96.8	96.4	97.2	95.8	96.2	96.8	96.3	96.6	96.7	96.3
w wgt decomp (0.8) ps	97.4	95.6	96.1	96.9	96.4	96.7	96.8	96.3	97.2	95.6	96.1	96.7	96.3	96.5	96.6	96.2
w wgt decomp (3.2) ps	97.3	95.6	96.0	96.9	96.2	96.7	96.7	96.2	97.1	95.5	96.0	96.6	96.2	96.5	96.5	96.1
w wgt A (0.2) ps	98.0	95.5	96.2	97.3	96.2	97.3	97.0	96.5	97.9	95.7	96.3	97.3	96.4	97.2	97.2	96.5
w wgt A (0.8) ps	98.0	95.4	96.1	97.2	96.2	97.1	97.0	96.4	97.9	95.6	96.3	97.2	96.4	97.0	97.1	96.4
w wgt A (3.2) ps	97.9	95.2	96.0	97.1	96.1	97.1	96.9	96.2	97.8	95.5	96.2	97.2	96.4	97.0	97.0	96.3
w Abadie Imbens ps	97.4	96.5	96.1	97.9	96.2	97.8	97.1	96.8	97.4	96.8	96.4	97.9	96.4	97.9	97.4	96.9
w boot effect se	97.5	96.4	97.2	96.8	97.3	96.6	97.2	96.8	97.3	96.3	97.1	96.5	97.2	96.4	96.8	96.8
w boot effect quant	97.7	96.5	97.2	97.0	97.3	96.9	97.2	97.0	97.5	96.4	97.1	96.7	97.2	96.6	96.9	96.9

Note: The prefixes ‘s’ and ‘w’ stand for the standard and wild bootstrap, respectively. All results with prefixes ‘s’ and ‘w’ are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure. The suffix ‘ps’ stands for adjustment for propensity score estimation.

Table A.11: Coverage across simulation designs for radius matching

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
	radius matching R1.5															
wgt uncond var	99.4	99.4	99.5	99.4	99.4	99.4	99.7	99.1	99.4	99.5	99.5	99.4	99.4	99.5	99.7	99.2
wgt decomp (0.2)	99.4	99.4	99.5	99.4	99.4	99.5	99.8	99.1	99.4	99.6	99.5	99.4	99.4	99.5	99.7	99.2
wgt decomp (0.8)	99.4	99.4	99.4	99.4	99.4	99.4	99.8	99.1	99.4	99.5	99.5	99.3	99.4	99.4	99.7	99.1
wgt decomp (3.2)	99.1	99.2	99.2	99.0	99.1	99.1	99.6	98.6	99.0	99.2	99.3	99.0	99.2	99.0	99.6	98.6
wgt A (0.2)	96.7	96.2	96.8	96.0	96.9	96.0	96.5	96.3	96.2	96.3	96.5	96.0	96.9	95.7	96.5	96.0
wgt A (0.8)	96.7	96.1	96.7	96.0	96.8	95.9	96.6	96.1	96.2	96.3	96.5	96.0	96.8	95.6	96.5	95.9
wgt A (3.2)	96.5	95.8	96.5	95.8	96.7	95.6	96.9	95.5	96.0	96.0	96.2	95.8	96.7	95.3	96.7	95.3
wgt uncond var ps	97.7	98.7	98.6	97.8	98.2	98.2	99.1	97.3	97.7	98.9	98.8	97.8	98.5	98.1	99.2	97.4
wgt decomp (0.2) ps	97.7	98.8	98.6	97.9	98.2	98.3	99.1	97.4	97.7	99.0	98.9	97.8	98.5	98.2	99.1	97.5
wgt decomp (0.8) ps	97.7	98.6	98.5	97.8	98.2	98.2	99.1	97.3	97.6	98.8	98.7	97.7	98.4	98.1	99.1	97.3
wgt decomp (3.2) ps	96.7	98.0	98.0	96.8	97.6	97.2	98.7	96.0	96.6	98.2	98.1	96.7	97.7	97.1	98.9	96.0
wgt A (0.2) ps	90.2	92.2	91.9	90.5	91.8	90.6	91.9	90.5	90.0	93.0	92.3	90.7	92.4	90.6	92.2	90.8
wgt A (0.8) ps	90.2	92.0	91.8	90.4	91.8	90.4	92.1	90.1	89.9	92.8	92.1	90.7	92.3	90.5	92.3	90.4
wgt A (3.2) ps	89.8	91.3	91.2	90.0	91.5	89.6	92.5	88.6	89.7	92.2	91.6	90.3	91.9	90.0	92.8	89.0
s wgt uncond var	94.0	92.8	94.5	92.3	93.8	93.0	93.6	93.2	93.9	93.5	94.5	92.8	94.3	93.1	93.9	93.4
s wgt decomp (0.2)	95.3	93.9	95.0	94.2	94.2	95.0	94.2	95.0	95.1	94.3	95.1	94.3	94.7	94.7	94.4	95.1
s wgt decomp (0.8)	95.3	93.7	94.9	94.2	94.3	94.7	94.1	94.9	95.1	94.4	95.2	94.3	94.8	94.7	94.5	95.0
s wgt decomp (3.2)	95.2	93.6	94.9	94.0	94.2	94.7	94.2	94.6	94.9	94.2	95.0	94.2	94.7	94.4	94.6	94.6
s wgt A (0.2)	96.4	94.6	96.0	95.0	95.3	95.7	95.4	95.6	96.2	95.4	96.2	95.4	95.8	95.8	95.8	95.8
s wgt A (0.8)	96.3	94.4	95.9	94.8	95.3	95.4	95.2	95.4	96.1	95.0	96.0	95.0	95.7	95.4	95.5	95.5
s wgt A (3.2)	95.6	94.1	95.5	94.2	94.9	94.8	94.8	94.9	95.4	94.6	95.5	94.5	95.3	94.7	95.0	95.0
s wgt uncond var ps	92.7	91.4	92.9	91.2	92.4	91.7	92.5	91.6	92.8	91.9	93.3	91.4	93.1	91.6	92.7	92.0
s wgt decomp (0.2) ps	94.6	92.4	93.7	93.3	93.0	94.0	93.0	94.0	94.5	93.3	94.1	93.6	93.7	94.0	93.5	94.2
s wgt decomp (0.8) ps	94.6	92.5	93.7	93.3	93.0	94.0	93.0	94.1	94.5	93.2	94.0	93.6	93.8	93.9	93.5	94.1
s wgt decomp (3.2) ps	94.5	92.0	93.4	93.2	92.8	93.7	93.2	93.4	94.5	92.9	93.8	93.5	93.7	93.6	93.7	93.6
s wgt A (0.2) ps	96.4	92.9	94.4	94.9	93.9	95.4	94.7	94.6	96.5	93.7	94.9	95.2	94.5	95.6	95.3	94.8
s wgt A (0.8) ps	96.2	92.5	94.1	94.6	93.6	95.1	94.3	94.3	96.3	93.2	94.6	94.9	94.3	95.1	95.0	94.5
s wgt A (3.2) ps	95.5	92.0	93.5	94.0	93.2	94.3	93.9	93.6	95.5	92.8	93.9	94.4	94.0	94.3	94.5	93.8
s boot effect se	97.9	95.6	97.1	96.5	96.7	96.8	96.9	96.6	97.8	96.3	97.3	96.8	97.2	96.9	97.3	96.9
s boot effect quant	97.9	95.7	97.0	96.5	96.7	96.8	96.8	96.8	97.8	96.4	97.3	96.9	97.3	97.0	97.2	97.1
	radius matching R3															
wgt uncond var	99.5	99.4	99.5	99.5	99.4	99.5	99.8	99.1	99.4	99.6	99.5	99.4	99.4	99.5	99.7	99.2
wgt decomp (0.2)	99.5	99.5	99.5	99.5	99.4	99.6	99.8	99.2	99.4	99.5	99.5	99.4	99.4	99.5	99.7	99.2
wgt decomp (0.8)	99.4	99.4	99.4	99.4	99.4	99.5	99.8	99.1	99.4	99.5	99.5	99.4	99.4	99.5	99.7	99.2
wgt decomp (3.2)	99.1	99.2	99.2	99.1	99.1	99.2	99.7	98.7	99.0	99.3	99.3	99.1	99.2	99.1	99.6	98.7
wgt A (0.2)	96.8	96.4	96.9	96.2	97.0	96.2	96.6	96.5	96.4	96.5	96.5	96.3	97.0	95.9	96.5	96.3
wgt A (0.8)	96.8	96.3	96.9	96.2	97.0	96.1	96.7	96.4	96.4	96.4	96.5	96.3	96.9	95.8	96.6	96.2
wgt A (3.2)	96.6	96.0	96.6	96.0	96.8	95.7	96.9	95.6	96.1	96.2	96.3	96.0	96.8	95.5	96.7	95.6
wgt uncond var ps	97.7	98.7	98.6	97.9	98.1	98.3	99.1	97.4	97.6	99.1	98.8	97.9	98.4	98.3	99.2	97.6
wgt decomp (0.2) ps	97.7	98.8	98.6	98.0	98.2	98.4	99.1	97.5	97.7	99.1	98.8	98.0	98.4	98.3	99.2	97.6
wgt decomp (0.8) ps	97.7	98.7	98.5	97.9	98.1	98.3	99.0	97.3	97.6	98.9	98.7	97.9	98.3	98.2	99.1	97.4
wgt decomp (3.2) ps	96.7	98.0	97.9	96.8	97.5	97.3	98.7	96.0	96.6	98.4	98.2	96.8	97.7	97.3	98.9	96.1
wgt A (0.2) ps	90.3	92.3	91.8	90.8	91.9	90.7	92.0	90.5	90.1	93.0	92.1	91.0	92.4	90.7	92.3	90.9
wgt A (0.8) ps	90.3	92.0	91.6	90.7	91.8	90.5	92.1	90.2	90.1	92.8	92.1	90.9	92.3	90.6	92.4	90.6
wgt A (3.2) ps	90.0	91.4	91.1	90.3	91.5	89.9	92.5	88.9	89.8	92.1	91.5	90.4	92.0	90.0	92.8	89.1
s wgt uncond var	94.4	93.3	94.6	93.0	94.0	93.7	94.0	93.7	94.2	93.7	94.7	93.2	94.4	93.6	94.2	93.8
s wgt decomp (0.2)	95.3	93.8	95.0	94.1	94.2	94.9	94.3	94.9	95.0	94.3	95.1	94.2	94.6	94.7	94.5	94.8
s wgt decomp (0.8)	95.3	93.7	95.0	94.1	94.2	94.8	94.2	94.8	95.0	94.4	95.1	94.4	94.8	94.7	94.6	94.9
s wgt decomp (3.2)	95.2	93.7	95.0	93.9	94.2	94.7	94.3	94.6	95.0	94.3	95.0	94.3	94.7	94.6	94.7	94.7
s wgt A (0.2)	96.3	94.5	95.9	94.9	95.2	95.6	95.4	95.5	96.2	95.2	96.1	95.3	95.8	95.7	95.7	95.8
s wgt A (0.8)	96.2	94.3	95.8	94.7	95.2	95.3	95.2	95.3	96.1	94.8	95.9	95.0	95.6	95.3	95.4	95.5
s wgt A (3.2)	95.6	94.0	95.5	94.2	94.8	94.9	94.8	94.9	95.4	94.7	95.5	94.6	95.3	94.7	95.0	95.1
s wgt uncond var ps	93.4	91.8	93.2	92.0	92.6	92.6	93.0	92.1	93.4	92.3	93.5	92.2	93.2	92.5	93.2	92.4
s wgt decomp (0.2) ps	94.7	92.5	93.7	93.5	93.0	94.2	93.3	93.9	94.6	93.2	94.1	93.7	93.8	94.0	93.7	94.0
s wgt decomp (0.8) ps	94.7	92.5	93.7	93.4	93.0	94.1	93.2	93.9	94.6	93.2	94.1	93.7	93.7	94.0	93.7	94.0
s wgt decomp (3.2) ps	94.7	92.1	93.4	93.4	92.9	93.9	93.3	93.5	94.6	93.1	93.9	93.8	93.7	94.0	93.9	93.7
s wgt A (0.2) ps	96.4	93.0	94.5	94.9	93.9	95.5	94.7	94.7	96.4	93.6	94.9	95.2	94.5	95.6	95.2	94.9
s wgt A (0.8) ps	96.3	92.6	94.1	94.7	93.7	95.1	94.5	94.3	96.3	93.3	94.5	95.0	94.4	95.2	95.0	94.6
s wgt A (3.2) ps	95.6	91.9	93.5	94.0	93.2	94.3	93.9	93.7	95.6	92.9	94.0	94.5	94.0	94.5	94.5	94.0
s boot effect se	97.7	95.3	96.9	96.1	96.5	96.5	96.6	96.4	97.6	96.0	97.1	96.5	97.1	96.5	96.9	96.7
s boot effect quant	97.7	95.2	96.9	96.1	96.5	96.5	96.4	96.5	97.6	96.1	97.0	96.6	97.1	96.5	96.9	96.8

Note: Prefix ‘s’ stands for standard bootstrap. All results with prefix ‘s’ are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure.

Table A.12: Coverage across simulation designs for radius matching with bias correction

	homogeneity								heterogeneity							
	sample		% treated		strong sel		outcome		sample		% treated		strong sel		outcome	
	500	2000	30	70	no	yes	bin	cont	500	2000	30	70	no	yes	bin	cont
	radius matching R1.5BC															
wgt uncond var	99.5	99.4	99.6	99.3	99.5	99.4	99.6	99.3	99.5	99.3	99.5	99.2	99.5	99.3	99.6	99.2
wgt decomp (0.2)	99.5	99.5	99.7	99.3	99.5	99.4	99.6	99.3	99.5	99.3	99.6	99.2	99.5	99.3	99.6	99.2
wgt decomp (0.8)	99.5	99.3	99.6	99.2	99.5	99.3	99.6	99.2	99.4	99.2	99.5	99.2	99.5	99.2	99.6	99.1
wgt decomp (3.2)	99.2	99.0	99.4	98.8	99.3	98.9	99.5	98.7	99.2	99.1	99.4	98.8	99.3	98.9	99.5	98.7
wgt A (0.2)	97.1	96.1	97.2	96.0	97.2	96.0	96.2	97.0	96.7	96.2	96.8	96.1	97.2	95.7	96.0	96.9
wgt A (0.8)	97.0	96.0	97.2	95.9	97.2	95.9	96.2	96.8	96.7	96.1	96.8	96.1	97.1	95.7	96.1	96.8
wgt A (3.2)	96.9	95.8	97.0	95.7	97.1	95.5	96.4	96.3	96.5	96.1	96.7	95.9	97.1	95.5	96.3	96.3
wgt uncond var ps	98.0	98.5	98.8	97.7	98.4	98.1	98.9	97.6	97.9	98.7	98.9	97.7	98.5	98.1	98.9	97.7
wgt decomp (0.2) ps	98.0	98.6	98.8	97.7	98.4	98.1	98.9	97.7	97.9	98.7	98.9	97.7	98.6	98.1	98.9	97.7
wgt decomp (0.8) ps	97.9	98.4	98.7	97.6	98.4	98.0	98.8	97.5	97.9	98.6	98.8	97.7	98.5	98.0	98.9	97.6
wgt decomp (3.2) ps	97.1	97.9	98.3	96.7	97.8	97.2	98.5	96.5	97.0	98.1	98.4	96.7	98.0	97.1	98.6	96.6
wgt A (0.2) ps	91.0	92.0	92.4	90.7	92.6	90.5	91.4	91.6	90.9	92.3	92.4	90.8	92.9	90.3	91.3	91.8
wgt A (0.8) ps	91.0	91.8	92.2	90.6	92.5	90.3	91.5	91.3	90.9	92.2	92.4	90.7	92.8	90.3	91.5	91.6
wgt A (3.2) ps	90.7	91.3	91.8	90.2	92.3	89.7	91.9	90.1	90.5	91.8	92.0	90.3	92.5	89.8	92.0	90.3
s wgt uncond var	94.6	93.9	95.1	93.5	94.2	94.3	94.1	94.4	94.5	94.1	94.9	93.7	94.4	94.1	94.1	94.4
s wgt decomp (0.2)	94.4	94.0	95.0	93.4	94.1	94.3	94.0	94.4	94.2	94.3	95.0	93.6	94.4	94.1	94.0	94.5
s wgt decomp (0.8)	94.4	93.9	94.9	93.4	94.1	94.2	93.9	94.4	94.2	94.2	94.9	93.5	94.3	94.1	94.1	94.4
s wgt decomp (3.2)	94.2	93.7	94.9	93.0	93.9	94.0	93.8	94.1	94.0	94.1	94.8	93.3	94.2	93.9	94.1	94.0
s wgt A (0.2)	95.1	94.6	95.8	93.9	94.9	94.8	94.7	95.0	94.9	95.0	95.8	94.1	95.1	94.8	94.9	95.0
s wgt A (0.8)	95.0	94.3	95.6	93.7	94.8	94.6	94.5	94.8	94.9	94.6	95.6	93.9	95.0	94.5	94.8	94.7
s wgt A (3.2)	94.6	94.1	95.4	93.3	94.6	94.1	94.3	94.4	94.5	94.4	95.3	93.6	94.8	94.1	94.6	94.3
s wgt uncond var ps	93.8	92.7	93.8	92.8	92.9	93.6	93.2	93.3	93.8	93.0	93.9	92.9	93.3	93.5	93.3	93.5
s wgt decomp (0.2) ps	93.6	92.8	93.7	92.6	93.0	93.4	93.0	93.3	93.6	93.1	93.9	92.8	93.3	93.4	93.1	93.5
s wgt decomp (0.8) ps	93.5	92.6	93.5	92.6	92.9	93.3	93.0	93.2	93.5	93.0	93.8	92.7	93.2	93.3	93.1	93.4
s wgt decomp (3.2) ps	93.3	92.4	93.4	92.3	92.6	93.1	93.0	92.7	93.3	92.9	93.7	92.5	93.0	93.2	93.2	93.0
s wgt A (0.2) ps	94.8	92.7	93.9	93.6	93.4	94.1	93.7	93.8	94.9	93.4	94.5	93.9	94.0	94.4	94.3	94.1
s wgt A (0.8) ps	94.7	92.5	93.7	93.5	93.2	93.9	93.6	93.5	94.9	93.2	94.3	93.8	93.8	94.3	94.1	93.9
s wgt A (3.2) ps	94.4	92.1	93.3	93.1	93.0	93.5	93.4	93.1	94.4	92.7	93.8	93.3	93.4	93.7	93.7	93.4
s boot effect se	97.8	95.7	97.3	96.3	96.6	97.0	97.0	96.6	97.8	96.2	97.2	96.7	97.1	96.9	97.1	96.9
s boot effect quant	97.9	95.8	97.3	96.4	96.6	97.1	97.0	96.7	97.8	96.3	97.2	96.8	97.1	96.9	97.1	96.9
	radius matching R3BC															
wgt uncond var	99.6	99.5	99.6	99.5	99.5	99.6	99.7	99.4	99.5	99.5	99.6	99.4	99.5	99.5	99.7	99.3
wgt decomp (0.2)	99.6	99.5	99.7	99.5	99.5	99.6	99.7	99.4	99.5	99.5	99.6	99.4	99.5	99.5	99.7	99.3
wgt decomp (0.8)	99.6	99.4	99.6	99.4	99.5	99.5	99.7	99.4	99.5	99.4	99.6	99.4	99.5	99.5	99.7	99.3
wgt decomp (3.2)	99.3	99.1	99.4	99.0	99.3	99.1	99.6	98.9	99.3	99.2	99.4	99.1	99.3	99.2	99.6	98.9
wgt A (0.2)	97.3	96.5	97.4	96.5	97.3	96.5	96.5	97.3	97.0	96.5	97.0	96.5	97.3	96.2	96.3	97.2
wgt A (0.8)	97.3	96.5	97.3	96.5	97.3	96.5	96.6	97.2	97.0	96.5	97.0	96.5	97.3	96.2	96.3	97.1
wgt A (3.2)	97.1	96.3	97.2	96.2	97.2	96.2	96.7	96.7	96.8	96.3	96.8	96.3	97.3	95.9	96.5	96.6
wgt uncond var ps	98.2	98.6	98.8	98.0	98.5	98.3	99.0	97.8	98.1	98.9	99.0	98.0	98.6	98.4	99.1	97.9
wgt decomp (0.2) ps	98.2	98.6	98.9	98.0	98.5	98.3	99.0	97.8	98.1	98.9	99.0	98.0	98.6	98.4	99.1	98.0
wgt decomp (0.8) ps	98.1	98.5	98.8	97.9	98.4	98.2	98.9	97.7	98.1	98.8	98.9	98.0	98.6	98.3	99.1	97.8
wgt decomp (3.2) ps	97.3	98.1	98.4	97.1	97.9	97.5	98.6	96.8	97.3	98.4	98.5	97.1	98.1	97.5	98.8	96.9
wgt A (0.2) ps	91.6	92.4	92.7	91.3	92.7	91.2	91.8	92.1	91.4	92.9	92.8	91.6	93.1	91.2	91.9	92.4
wgt A (0.8) ps	91.5	92.2	92.5	91.2	92.7	91.1	91.9	91.9	91.4	92.7	92.6	91.5	93.0	91.1	92.0	92.1
wgt A (3.2) ps	91.3	91.7	92.1	90.9	92.5	90.4	92.3	90.7	91.0	92.3	92.3	91.0	92.7	90.6	92.5	90.9
s wgt uncond var	94.6	93.7	94.9	93.3	94.0	94.2	94.0	94.2	94.4	94.1	95.0	93.5	94.4	94.1	94.2	94.3
s wgt decomp (0.2)	94.3	93.7	94.8	93.2	94.1	94.0	93.8	94.2	94.2	94.2	94.9	93.5	94.3	94.0	94.1	94.3
s wgt decomp (0.8)	94.3	93.7	94.8	93.2	94.0	94.0	93.8	94.2	94.1	94.1	94.8	93.5	94.3	94.0	94.1	94.2
s wgt decomp (3.2)	94.1	93.6	94.8	92.9	93.9	93.8	93.7	93.9	93.9	94.0	94.8	93.2	94.1	93.8	94.1	93.9
s wgt A (0.2)	95.0	94.3	95.7	93.7	94.8	94.6	94.5	94.8	94.8	94.8	95.6	94.0	95.1	94.5	94.7	94.9
s wgt A (0.8)	95.0	94.1	95.5	93.6	94.7	94.4	94.3	94.7	94.8	94.5	95.4	93.9	94.9	94.3	94.6	94.7
s wgt A (3.2)	94.6	93.9	95.2	93.2	94.5	93.9	94.2	94.2	94.3	94.4	95.2	93.5	94.7	94.0	94.5	94.3
s wgt uncond var ps	93.8	92.4	93.7	92.5	92.8	93.4	93.1	93.1	93.8	92.8	93.9	92.7	93.2	93.4	93.4	93.2
s wgt decomp (0.2) ps	93.6	92.5	93.6	92.5	92.8	93.3	93.0	93.1	93.6	93.0	93.9	92.6	93.2	93.3	93.3	93.2
s wgt decomp (0.8) ps	93.5	92.3	93.5	92.4	92.7	93.1	92.9	92.9	93.5	93.0	93.9	92.6	93.1	93.4	93.2	93.3
s wgt decomp (3.2) ps	93.2	92.0	93.2	92.1	92.5	92.8	92.8	92.5	93.3	92.7	93.7	92.4	92.9	93.1	93.2	92.8
s wgt A (0.2) ps	94.8	92.7	93.8	93.7	93.3	94.2	93.8	93.7	94.9	93.1	94.3	93.8	93.7	94.3	94.1	93.9
s wgt A (0.8) ps	94.8	92.3	93.6	93.5	93.2	93.9	93.6	93.5	94.8	92.9	94.1	93.7	93.6	94.1	94.0	93.7
s wgt A (3.2) ps	94.4	91.9	93.2	93.1	92.9	93.4	93.3	93.0	94.4	92.5	93.7	93.2	93.2	93.7	93.7	93.2
s boot effect se	97.7	95.5	97.1	96.1	96.5	96.7	96.7	96.4	97.6	95.9	97.0	96.5	97.0	96.5	96.9	96.6
s boot effect quant	97.7	95.4	97.0	96.1	96.4	96.7	96.7	96.4	97.6	95.9	97.0	96.5	97.0	96.5	96.9	96.6

Note: Prefix ‘s’ stands for standard bootstrap. All results with prefix ‘s’ are based on both smoothed and nonsmoothed versions of the respective bootstrap procedure.

Working Paper 466 january 2016

Authors

Hugo BODORY
University of St. Gallen, Department of Economics, Swiss Institute for Empirical Economic Research, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland. Phone: +41 71 224 2767; Email: hugo.bodory@unisg.ch;
Website: http://www.sew.unisg.ch/de/ueber_uns/team?person=3157ccab-2014-4ad8-b038-6fea03292d05

Lorenzo CAMPONOVO
University of St. Gallen, Department of Economics, Maths and Statistics, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland. Phone: +41 71 224 2432; Email: lorenzo.camponovo@unisg.ch;
Website: http://www.mathstat.unisg.ch/en/people?person=ce049465-b49b-4ea4-a930-463b167fd532

Martin HUBER
University of Fribourg, Faculty of Economics and Social Sciences, Chair of Applied Econometrics - Evaluation of Public Policies, Bd. de Pérolles 90, 1700 Fribourg, Switzerland. Phone: +41 26 300 8274; Email: martin.huber@unifr.ch;
Website: http://www.unifr.ch/appecon/en/team/martin-huber

Michael LECHNER
University of St. Gallen, Department of Economics, Swiss Institute for Empirical Economic Research, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland. Phone: +41 71 224 23 20; Email: michael.lechner@unisg.ch; Website: www.michael-lechner.eu

Abstract

This paper investigates the finite sample properties of a range of inference methods for propensity score-based matching and weighting estimators frequently applied to evaluate the average treatment effect on the treated. We analyse both asymptotic approximations and bootstrap methods for computing variances and confidence intervals in our simulation design, which is based on large scale labor market data from Germany and varies w.r.t. treatment selectivity, effect heterogeneity, the share of treated, and the sample size. The results suggest that in general, the bootstrap procedures dominate the asymptotic ones in terms of size and power for both matching and weighting estimators. Furthermore, the results are qualitatively quite robust across the various simulation features.

Citation proposal

Bodory Hugo, Camponovo Lorenzo, Huber Martin, Lechner Michael. 2016. «The finite sample performance of inference methods for propensity score matching and weighting estimators». Working Papers SES 466, Faculty of Economics and Social Sciences, University of Fribourg (Switzerland)

Jel Classification

C21

Keywords

Inference, variance estimation, treatment effects, matching, inverse probability weighting

Working Papers SES collection

Last published

460 Grossmann V., Osikominu A., Osterfeld M.: Are Sociocultural Factors Important for Studying a Science University Major?; 2015
461 Bücker J.J.L.E., Furrer O., Lin Yanyan.: Measuring Cultural Intelligence: A New Test of the CQ Scale; 2015
462 Denisova-Schmidt E., Huber M., Prytula Y.: An Experimental Evaluation of an Anti-Corruption Intervention among Ukrainian University Students; 2015
463 Böhm S.: Regional Economic Integration and Factor Mobility in Unified Germany; 2015
464 Grossmann V., Strulik H.: Optimal Social Insurance and Health Inequality; 2015
465 Baguet M., Dumas C.: Birth weight and long-term outcomes in a developing country; 2015

Catalogue and download links

http://www.unifr.ch/ses/wp
http://doc.rero.ch/collection/WORKING_PAPERS_SES

Publisher

Université de Fribourg, Suisse, Faculté des sciences économiques et sociales
Universität Freiburg, Schweiz, Wirtschafts- und sozialwissenschaftliche Fakultät
University of Fribourg, Switzerland, Faculty of Economics and Social Sciences

Bd de Pérolles 90, CH-1700 Fribourg
Tél.: +41 (0) 26 300 82 00
decanat-ses@unifr.ch www.unifr.ch/ses