Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators

Hugo Bodory, Lorenzo Camponovo, Martin Huber & Michael Lechner

To cite this article: Hugo Bodory, Lorenzo Camponovo, Martin Huber & Michael Lechner (2020) The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators, Journal of Business & Economic Statistics, 38:1, 183-200, DOI: 10.1080/07350015.2018.1476247

To link to this article: https://doi.org/10.1080/07350015.2018.1476247

The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators

Hugo BODORY
Department of Economics, University of St. Gallen, St. Gallen, Switzerland (*hugo.bodory@unisg.ch*)

Lorenzo CAMPONOVO
Department of Economics, University of Surrey, Surrey, UK (*l.camponovo@surrey.ac.uk*)

Martin HUBER
Department of Economics, University of Fribourg, Fribourg, Switzerland (*martin.huber@unifr.ch*)

Michael LECHNER
Department of Economics, University of St. Gallen, St. Gallen, Switzerland (*michael.lechner@unisg.ch*)

This article 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 analyze both asymptotic approximations and bootstrap methods for computing variances and confidence intervals in our simulation designs, which are based on German register data and U.S. survey data. We vary the design w.r.t. treatment selectivity, effect heterogeneity, share of treated, and sample size. The results suggest that in general, theoretically justified bootstrap procedures (i.e., wild bootstrapping for pair matching and standard bootstrapping for "smoother" treatment effect estimators) dominate the asymptotic approximations in terms of coverage rates for both matching and weighting estimators. Most findings are robust across simulation designs and estimators.

KEY WORDS: Inference; Inverse probability weighting; Matching; Treatment effects; Variance estimation.

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; 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. While some treatment effect estimators directly control for the observed covariates, most of them are based on conditioning on the treatment propensity score instead, that is the conditional probability to receive the treatment given the covariates, 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; Dehejia and Wahba 1999) and inverse probability weighting (henceforth IPW, Horvitz and Thompson 1952; Hirano, Imbens, and Ridder 2003), which belong to the most popular methods among practitioners.

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 (2017). 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 estimation. Pingel (2015), for instance, focused on the impact of tuning parameters on the accuracy of the variance estimator of Abadie and Imbens (2016), but does not compare several classes of variance estimators. This is surprising, as the accuracy of inference appears equally important as the accuracy of point estimation.

Journal of Business & Economic Statistics
January 2020, Vol. 38, No. 1
DOI: 10.1080/07350015.2018.1476247

This article 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. 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 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 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) introduced and prove the validity of a weighted bootstrap algorithm for particular classes of pair matching estimators that, however, do not include propensity score matching. Bodory et al. (2016) generalized 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 (2012) 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 generalized methods of moments (GMM) estimation of the variance (for IPW), approximations of the variances based on the weights nontreated receive in ATET estimation as in Lechner (2002a) (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 (2016). 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 bootstrap p-values as suggested by Racine and MacKinnon (2007) to improve accuracy of inference when the number of bootstrap replications is low.

Our simulation designs make use of two different empirical datasets. The first one is based on German register data on active labor market policies as previously considered in Huber, Lechner, and Wunsch (2013) and Lechner and Wunsch (2013). The treatment selection process and the association between the outcome and the covariates on which we base our simulations are estimated from these data, instead of relying on an arbitrarily chosen model. We vary several empirically relevant design features in our simulations, namely the sample size, selection into treatment, share of treated, and effect heterogeneity. Second, we investigate two simulation designs previously considered in Busso, DiNardo, and McCrary (2014), which are based on U.S. data from the National Supported Work (NSW), see LaLonde (1986), as well as the Panel Study of Income Dynamics (PSID), and vary them w.r.t. treatment selection.

We note that the usefulness of such empirically founded Monte Carlo approaches for ranking estimators has been challenged by Advani and Sloczyński (2013), who compare the performance of various estimators in experimental data from the NSW with their performance in empirically founded Monte Carlo designs generated from the experiment. Arguably, if empirical Monte Carlo methods were informative about the true ranking, there should be a strong correlation between the ranking in the experiment and the empirical Monte Carlo methods, which is not the case in Advani and Sloczyński (2013). However, their design has itself its issues, for instance, by considering a (due to the small sample size) noisy experimental estimate to be the true effect, which could disturb the ordering of the estimators in their analysis. Yet, we acknowledge the criticism of Advani and Sloczyński (2013) that empirically founded simulation designs likely fail in perfectly matching real world data generating processes and treatment effects. We nevertheless consider this approach to be preferable to generating fully arbitrary simulation designs that are not linked to empirical distributions and associations of variables at all.

Our results suggest that inference methods based on asymptotic approximations that ignore the first step estimation of the propensity score are frequently conservative, entailing over-coverage of the true effect, though the Abadie and Imbens (2006) variance estimator for pair matching is generally a noticeable exception. GMM-based variance estimation of IPW is conservative, too, even though it accounts for the estimation of the propensity score. In general, accounting for propensity score estimation in asymptotic approximations only partially mitigates over-coverage of some inference procedures, while entailing under-coverage of others. A further finding is that the coverage rates of theoretically justified bootstrap procedures are often more accurate than those of the asymptotic approximations. For IPW and radius matching, coverage rates using the standard bootstrap for either bootstrapping t-statistics based on asymptotic variance approximations or for bootstrapping the effects come closer to nominal size than the conservative coverage rates (exclusively) based on asymptotic approximations.

For pair matching, the standard bootstrap performs well in some cases, but is prone to under-coverage in others. In contrast, the wild bootstrap is generally closer to nominal size than the standard bootstrap as well as the (conservative) asymptotic variance approximations. We therefore recommend using bootstrap procedures that are consistent for the treatment effect estimator at hand. Many findings concerning the coverage rates of inference procedures are rather stable across the different simulation features like the distribution of the outcome variable, sample size, share of treated, treatment selection, and effect heterogeneity.

The remainder of this article 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 and various bootstrap implementations. Section 4 discusses the data and simulation designs. 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 (in addition to the "stable unit treatment value assumption" (SUTVA), see, for instance, Rubin 1990). First, the so-called selection on observables or conditional independence' assumption (CIA) (see, for instance, Imbens 2004; Imbens and Wooldridge 2009) has to be satisfied:

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

where '$\perp$' stands for statistical independence. This rules out the existence of (further) confounders that jointly influence the treatment and the potential outcome under nontreatment 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 of 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 iid sample of realizations of $\{Y_i, D_i, X_i\}$ with any observation $i \in 1, \ldots, n$. $n_1 = \sum_{i=1}^{n} D_i$ 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 $\sum_{j=1}^{n} \frac{(1-D_j)\hat{p}(X_j)}{1-\hat{p}(X_j)}$ ensures that the weights add to one. It is easy to see that (7) corresponds to (6) when setting $\hat{W}_i$ in the latter to $D_i + (1 - D_i)n_0\{\frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}}{\sum_{j=1}^{n} \frac{(1-D_j)\hat{p}(X_j)}{1-\hat{p}(X_j)}}\}$. 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 nontreated units with comparable propensity scores to estimate the ATET by the average difference in the outcomes of the treated and the (appropriately weighted) nontreated 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 nontreated observation j when matched to a treated unit i. Pair (or one-to-one) matching with replacement (implying that a nontreated observation may be matched several times; see, for instance, Rubin 1973), matches to each treated observation exactly the nontreated 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. 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 nontreated 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; Huber, Lechner, and Wunsch 2013).

Radius matching (see, for instance, Rosenbaum and Rubin 1985; Dehejia and Wahba 1999) uses *all* nontreated observations with propensity scores within a predefined radius around that of the treated reference unit, which trades off some bias 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., nontreated 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; Abadie and Imbens 2011) to remove small and large sample bias due to mismatches. See Huber, Lechner, and Steinmayr (2015) 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.

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 (2017). 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 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) suggested using a trimming procedure first discussed in Imbens (2004), which is asymptotically unbiased if common support holds asymptotically. It is based on setting the weights of those nontreated 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. To not create an unbalanced sample by trimming the nontreated observations only, any treated with propensity scores larger than the largest value among the remaining nontreated 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 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 bootstrap p-values 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 (2016) (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 of p-values (Section 3.4) are considered. In addition, we also investigate the wild bootstrap procedure introduced in Bodory et al. (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 of p-values. 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 of p-values.

The complication in all of these methods is how to deal with the fact that the propensity score is estimated. However, there is some common sense among practitioners that ignoring estimation error of the propensity score is likely to lead to conservative inference. This is based on the theoretical insights of Hahn (1998) for the ATE as well as, for example, simulation results in Lechner (2002b). However, it must be pointed out that there is no guarantee of this to be valid for the ATET in general.

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, that is, 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, as discussed for instance in Hansen (1982). 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.

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

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

This variance formula shows that $\text{asV}\left[\sqrt{n}\hat{\theta}_{\text{IPW}}\right]$ 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:

$$\begin{aligned} &H_{\theta_{\text{IPW}}} = E[\partial h(.)/\partial\theta_{\text{IPW}}] = 1, \quad V_{hh} = E[h(.)^2] = \text{var}[n\tilde{w}_i Y_i],\\ &H_\beta(d=1) = E[\partial h(.)/\partial\beta] = 0,\\ &H_\beta(d=0) = E[\partial h(.)/\partial\beta]\\ &\quad = E\left[n\frac{\frac{X_i\phi_i}{(1-p(X_i))^2}\sum_{i=1}^{n}\frac{p(X_i)}{1-p(X_i)} - \frac{p(X_i)}{1-p(X_i)}\sum_{i=1}^{n}\frac{X_i\phi_i}{(1-p(X_i))^2}}{\left(\sum_{i=1}^{n}\frac{p(X_i)}{1-p(X_i)}\right)^2} Y_i\right],\\ &G_\beta = E[\partial g(.)/\partial\beta], \quad V_{gg} = E[g(.)g(.)'],\\ &V_{gh} = E[g(.)h(.)], \quad V_{hg} = V_{gh}'. \end{aligned}$$

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 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.

3.2 Asymptotic Variance Approximations of Abadie and Imbens

Abadie and Imbens (2006) derived 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 (nontreated) 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

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

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

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

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. For the estimation 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 article, Abadie and Imbens (2016) proposed 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 $\text{Cov}(X_i, \mu(X_i, D_i)|p(X_i))$ the covariance between X_i and $\mu(X_i, D_i)$ conditional on $p(X_i)$. Abadie and Imbens (2016) showed 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

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

$$\frac{\partial\theta}{\partial\beta} = \frac{1}{E[p(X)]}E[X\phi(X\beta)(\mu(X,1)-\mu(X,0)-\theta)].$$

$\text{cov}(X, \mu(X, D))$ (which can be shown to equal $\text{cov}(X, Y|p(X), D))$, $\mu(p(X), D)$, and $\mu(X, D)$, which enter the correction terms in (15), may be estimated by pair matching within or across treatment groups, as we do in our simulations, see Abadie and Imbens (2016) for further details. Note that the adjustment term may increase or decrease the variance estimate of the ATET. 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

According to Equation (6), all estimators considered are a difference of weighted means among the treated and controls. Therefore, Lechner (2002b) suggested approximating the variance of matching estimators based on these weights (assuming that they are nonstochastic). Thus, under iid sampling and fixed weights the variance of the estimator of the ATET is the sum of the variance of the estimator used for the treated and the estimator used for the controls. Since the potential outcome for the treated is estimated by their sample mean, the standard variance estimator for means of random variables can be applied:

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

Concerning the variance of the treated population's estimated mean potential outcome under nontreatment, Equation (6) shows that the estimated mean potential outcome under nontreatment of the treated can be expressed as a weighted sum of nontreated outcomes. The normalized nontreated weights $\tilde{W}_i$ add to one: $\hat{E}[Y_i(0)|D_i=1] = \sum_{i=1}^{n}(1-D_i)Y_i\tilde{W}_i$ (note that for conventional convenience the weights here sum up to 1, while in Equation (6) they sum up to n_0). For instance, for the IPW estimator (7) $\tilde{W}_i = \{\frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}}{\sum_{j=1}^{n}\frac{(1-D_j)\hat{p}(X_j)}{1-\hat{p}(X_j)}}\}$. One simple approximation to

the variance $V\{\hat{E}[Y_i(0)|D_i=1]\}$ is therefore the unconditional variance of $Y_i\tilde{W}_i$:

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

This result implicitly assumes homoscedasticity of non-treatment outcomes in $\tilde{W}_i$. To allow the variance of the non-treatment outcome to vary with the weights, we consider the following decomposition of the variance into the expectation of the conditional variance and the variance of the conditional expectation given the weights:

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

Note that

$$A = E\left\{\sum_{i=1}^{n}(1-D_i)\tilde{W}_i^2\sigma^2(\tilde{W}_i, D_i=0)\right\}, \tag{18}$$

$$B = V\left\{\sum_{i=1}^{n}(1-D_i)\tilde{W}_iE[Y_i|\tilde{W}_i]\right\}, \tag{19}$$

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 $\tilde{W}_iE[Y_i|\tilde{W}_i]$ is uncorrelated across i, the variance of the sum equals n_0 times the variance of its components (which are functions of W_i only):

$$V\left\{\sum_{i=1}^{n}(1-D_i)\tilde{W}_iE[Y_i|\tilde{W}_i]\right\} = n_0V\left\{\tilde{W}_i\mu(\tilde{W}_i, D_i=0),\right\}, \tag{20}$$

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 nontreated. 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 $\hat{\sigma}^2(\tilde{W}_i,0) = E[(Y_i-\hat{\mu}(\tilde{W}_i, D_i=0))^2|\tilde{W}_i, D_i=0]$. Essentially, this is a one-dimensional nonparametric estimation problem for a conditional mean and a conditional variance for which many possible estimators are available. 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 $\tilde{W}_i$:

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

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

$$\hat{\mu}(\tilde{W}_i, D_i) = \frac{1}{M}\sum_{i\in\mathcal{S}_M(i)}^{M} Y_i,$$

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

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

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

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

We consider variance estimation based on (i) the unconditional variance formula in (16), (ii) the decomposition based approach with $\hat{V}\{\hat{E}[Y_i(0)|D_i=1]\} = \hat{A}+\hat{B}$, and (iii), based on $\hat{A}$ only. Concerning the estimation of the conditional means and variances required in approaches (ii) and (iii), 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 three 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 formulas, 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 (2016) 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 nonsmoothness 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), such that the inconsistency result for the bootstrap might not apply. Furthermore, bootstrapping automatically accounts for heteroscedasticity, trimming of influential observations, and uncertainty due to propensity score estimation and bias correction. Even for nonsmooth 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, we randomly draw 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, $b \in \{1, 2, \ldots, B\}$. We consider two options for computing p-values and confidence intervals in our simulations. One is based on plugging the square root of the bootstrap variance of the ATET, $\hat{V}(\hat{\theta}^b) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^b - \frac{1}{B}\sum_b^B \hat{\theta}^b)^2$, into the t-statistic and evaluating the latter on its asymptotic normal distribution to obtain the p-value. Confidence intervals are standardly obtained by $\hat{\theta} \pm \sqrt{\hat{V}(\hat{\theta}^b)}c$, where c denotes the asymptotic critical value for a particular confidence level α. The other option is to compute 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:

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

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 bootstrap 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 3.1–3.3: $T_n = \hat{\theta}/\sqrt{\hat{V}(\hat{\theta})}$, 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)), that is, 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):

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

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):

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

$K(T_n^b, T_n, h) = K\left(\frac{|T_n|-|T_n^b|}{h}\right)$ 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) argued 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):

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

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, because 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. In addition, smaller values of B, namely 99 and 49, are also considered, to analyze the relationship between bootstrap performance and number of bootstrap draws. In this context, we note that (Andrews and Buchinsky 2000) provide a method for

choosing B to achieve a specific level of accuracy, measured as percentage deviation of a bootstrap quantity (e.g., a standard error or p-value) using a particular B from the ideal quantity under an infinite number of bootstrap replications.

3.5 Wild Bootstrap

The definition of the wild bootstrap procedure introduced in Bodory et al. (2016) relies on the martingale representation for matching estimators proposed in Abadie and Imbens (2012). 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 (2012), we can write the matching estimator as $\sqrt{n}(\hat{\theta}_{\text{match}} - \theta) = T_{1n} + T_{2n} + o_p(1)$, where

$$T_{1n} = \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n} D_i(\mu(\hat{p}(X_i), 1) - \mu(\hat{p}(X_i), 0) - \theta),$$

$$T_{2n} = \frac{\sqrt{n}}{n_1} \sum_{i=1}^{n} (D_i - (1 - D_i)K_i)\,(Y_i - \mu(\hat{p}(X_i), D_i))\,.$$

The wild bootstrap algorithm uses this representation to reproduce the sampling distribution of $\sqrt{n}(\hat{\theta}_{\text{match}} - \theta)$. In particular, we apply the following approach. First, we generate random treatments D_i^* using the estimated propensity score $\hat{p}(X_i)$. Then, we reestimate the propensity score $\hat{p}^*(X_i)$ using these bootstrap treatments $(D_1^*, \ldots, D_n^*)$. Let K_i^* denote the number of times unit i is used as a match. Furthermore, let $\hat{\mu}(p, 0)$ and $\hat{\mu}(p, 1)$ be some nonparametric estimators of $\mu(p, 0)$ and $\mu(p, 1)$, respectively. Then, we approximate the sampling distribution of $\sqrt{n}(\hat{\theta}_{\text{match}} - \theta)$ with the wild bootstrap decomposition

$$T_{1n}^* = \frac{\sqrt{n}}{n_1^*} \sum_{i=1}^{n} D_i^*(\hat{\mu}(\hat{p}^*(X_i), 1) - \hat{\mu}(\hat{p}^*(X_i), 0) - \hat{\theta}_{\text{match}})v_i,$$

$$T_{2n}^* = \frac{\sqrt{n}}{n_1^*} \sum_{i=1}^{n} \left(D_i^* - (1 - D_i^*)K_i^*\right) \hat{\epsilon}_{i,D_i^*} v_i,$$

where $n_1^* = \sum_{i=1}^{n} D_i^*$, $\hat{\epsilon}_{i,D_i^*} = (\hat{\sigma}^2(\hat{p}(X_i, D_i^*)))^{1/2}$ defined in (14), and $(v_1, \ldots, v_n)$ are iid random variables with $E[v_i] = 0$, $E[v_i^2] = 1$ and $E[v_i^4] < \infty$. As nonparametric estimators $\hat{\mu}(X_i, 0)$ and $\hat{\mu}(X_i, 1)$ we simply use the matching approach also adopted in Section 3.3. Note that a similar procedure has been previously adopted in Otsu and Rai (2015), who introduce and prove the consistency of a weighted bootstrap algorithm. However, unlike Otsu and Rai (2015), our bootstrap method can also be applied to propensity score matching. 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 the inference procedures investigated for the respective point estimators in our simulation study.

4. SIMULATION DESIGN

4.1 Empirical Monte Carlo Study Based on German Register Data

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 (2015, 2016), Lechner and Strittmatter (2017), and Frölich, Huber, and Wiesenfarth (2017). Our first set of simulations exploits the same administrative data as Huber, Lechner, and Wunsch (2013), namely a 2% random sample of employees in Germany who are subject to social insurance from 1990 to 2006. The dataset 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 (iv) 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, for example, 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 nonstandard 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).

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 nontreated 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 provides descriptive statistics for the treated and nontreated in the full sample, which is informative about selection

Table 1. Inference methods and point estimators

Variance (var) estimator (row) / ATET estimator (column)	IPW	PM	R1.5	R3	R1.5BC	R3BC
Analytical var using GMM (3.1)	x					
Standard bootstrap of t-stat with analytical var (3.4)	x					
Analytical var of Abadie and Imbens (3.2)		x				
Standard bootstrap of t-stat with analytical var (3.4)		x				
Wild bootstrap of t-stat with analytical var (3.5)		x				
Analytical var of Abadie and Imbens with p-score correction (3.2)		x				
Standard bootstrap of t-stat with analytical var (3.4)		x				
Wild bootstrap of t-stat with analytical var (3.5)		x				
Weights-based variance: uncond var (3.3)	x	x	x	x	x	x
Standard bootstrap of t-stat with weights-based var (3.4)	x	x	x	x	x	x
Wild bootstrap of t-stat with weights-based var (3.5)		x				
Weights-based var: $\hat{A} + \hat{B}$ (3.3)	x	x	x	x	x	x
Standard bootstrap of t-stat with weights-based var (3.4)	x	x	x	x	x	x
Wild bootstrap of t-stat with weights-based var (3.5)		x				
Weights-based var: $\hat{A}$ (3.3)	x	x	x	x	x	x
Standard bootstrap of t-stat with weights-based var (3.4)	x	x	x	x	x	x
Wild bootstrap of t-stat with weights-based var (3.5)		x				
Weights-based var with p-score correction: uncond var (3.3, 3.2)		x	x	x	x	x
Standard bootstrap of t-stat with weights-based var (3.4)		x	x	x	x	x
Wild bootstrap of t-stat with weights-based var (3.5)		x				
Weights-based var with p-score correction: $\hat{A} + \hat{B}$ (3.3, 3.2)		x	x	x	x	x
Standard bootstrap of t-stat with weights-based var (3.4)		x	x	x	x	x
Wild bootstrap of t-stat with weights-based var (3.5)		x				
Weights-based var with p-score correction: $\hat{A}$ (3.3, 3.2)		x	x	x	x	x
Standard bootstrap of t-stat with weights-based var (3.4)		x	x	x	x	x
wild bootstrap of t-stat with weights-based var (3.5)		x				
Standard bootstrap of ATET with bootstrap std error in t-stat (3.4)	x	x	x	x	x	x
Standard bootstrap of ATET using the quantile method (3.4)	x	x	x	x	x	x
Wild bootstrap of ATET with bootstrap std error in t-stat (3.4)		x				
Wild bootstrap of ATET using the quantile method (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 $M = 2\,\text{round}(\kappa\sqrt{n}) + 1$, with $\kappa = 0.2, 0.8, 3.2$.

into treatment relevant for step (i). 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. We use almost the same covariates as Huber, Lechner, and Wunsch (2013), with the exception of 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. 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, 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. 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 = moderate 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% or 30%, respectively. Note that the simulations are not conditional on the treatment, implying that the share of treated in each simulation sample is a random number.

Table 2. Descriptive statistics of the full sample

	Treated		Nontreated			Probit model	
Variables	mean	std	mean	std	st.diff	m.eff	se
Some unsubsidized employment (Y)	0.63	0.48	0.56	0.50	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	0.50	0.46	0.50	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	0.20	0.40	15	1.0	0.1
No vocational degree	0.18	0.39	0.34	0.47	26	−0.3	0.1
At least one child in household	0.42	0.49	0.28	0.45	22	−0.2	0.1
Last occupation: Non-skilled worker	0.14	0.35	0.21	0.41	13	0.4	0.2
Last occupation: Salaried worker	0.40	0.49	0.22	0.41	29	1.8	0.2
Last occupation: Part time	0.22	0.42	0.16	0.36	12	2.1	0.4
UI benefits: 0	0.33	0.47	0.44	0.50	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	5.0	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	0.40	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	0.20	0.40	0.27	0.44	12	0.0	0.1
Pop. share 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	0.20	0.40	17	−1.2	0.2
Duration of last employment $>$ 1 year	0.41	0.49	0.43	0.50	4	−0.6	0.1
Av. earn. last 10 yrs when empl. / 1000	0.59	0.41	0.52	0.40	13	−0.4	0.2
Woman $\times$ age / 10	2.13	1.95	1.65	1.94	17	2.7	0.6
$\times$ squared / 1000	0.83	0.85	0.65	0.90	15	−2.8	0.7
$\times$ no vocational degree	0.09	0.28	0.16	0.36	15	−0.9	0.1
$\times$ at least one child in household	0.32	0.47	0.17	0.37	25	1.1	0.2
$\times$ share OLF last year	0.19	0.36	0.18	0.35	3	0.8	0.2
$\times$ average earnings last 10 y. if empl.	0.26	0.34	0.19	0.30	16	−1.4	0.3
$\times$ entering UE in 2003	0.10	0.30	0.13	0.33	6	−0.6	0.1
$X_i\tilde{\beta}$	−1.7	0.40	−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: $\tilde{\beta}$: probit coefficients. $\Phi(X_i\tilde{\beta})$: standard normal c.d.f. evaluated at $X_i\tilde{\beta}$. Pseudo-R^2 is the so-called Efron's R^2: $1-\sum_{i=1}^{n}[D_i-\Phi(X_i\tilde{\beta})]^2/\sum_{i=1}^{n}[D_i-n^{-1}\sum_{i=1}^{n}D_i]^2$. 'st.diff': standardized difference in % defined as mean difference normalized by the square root of the sum of the estimated variances of the particular variables in both subsamples (Imbens and Wooldridge 2009, p. 24). 'std': standard deviation. 'se': standard error in %. 'm.eff': marginal effect in % evaluated at the mean in the probit model for treatment selection based on discrete changes for binary variables and derivatives otherwise. Some descriptives in this Table 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 nontreated 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.

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. To investigate the performance of inference methods under heterogeneous effects, we in addition introduce models for the outcome variables with two different types of heterogeneity.

The first settings use economically motivated considerations to model heterogeneity. The DGPs change the binary employment status of unemployed (employed) individuals with a high (low) training probability. The rationale behind this is that those who are more likely to be assigned to a training program may have better employment opportunities (and vice versa). In a second step, this setting increases the earnings of those modeled as employed. We refer to these DGPs as "modeled heterogeneity."

The second type of heterogeneity exploits empirically observed differences in the outcomes of individuals with similar training probabilities but unequal treatment status. This

Table 3. Summary statistics (DGPs)

Selection	Treated	Probit (%)		Y(1)		Y(0)		ATET		Trimming	
strength	share (%)	st.diff	Pseudo-R^2	mean	std	mean	std	mean	std	500	2000
Effect homogeneity for employment											
moderate	70	41	8.7	0.6	0.5	0.6	0.5	0	0	5	0
moderate	30	42	9.1	0.6	0.5	0.6	0.5	0	0	0	0
strong	70	81	33.8	0.6	0.5	0.6	0.5	0	0	29	9
strong	30	89	34.4	0.7	0.5	0.7	0.5	0	0	6	0
Effect homogeneity for earnings											
moderate	70			11.0	11.8	11.0	11.8	0	0		
moderate	30			11.9	12.2	11.9	12.2	0	0		
strong	70			11.7	11.9	11.7	11.9	0	0		
strong	30			12.9	12.7	12.9	12.7	0	0		
Effect heterogeneity for employment (modeled heterogeneity)											
moderate	70	42	9.1	0.8	0.4	0.6	0.5	0.2	0.4	5	0
moderate	30	42	8.9	0.8	0.4	0.6	0.5	0.2	0.4	0	0
strong	70	81	33.9	0.8	0.4	0.6	0.5	0.2	0.4	29	9
strong	30	89	34.2	0.8	0.4	0.7	0.5	0.1	0.4	9	0
Effect heterogeneity for earnings (modeled heterogeneity)											
moderate	70			13.6	11.2	11.1	11.8	2.5	6.6		
moderate	30			14.5	11.3	11.8	12.1	2.7	6.8		
strong	70			14.3	11.2	11.7	12.0	2.6	6.7		
strong	30			16.0	11.5	12.9	12.6	3.1	7.2		
Effect heterogeneity for employment (empirical heterogeneity)											
moderate	70	41	8.7	0.6	0.5	0.6	0.5	0	0.7	4	0
moderate	30	42	8.9	0.6	0.5	0.6	0.5	0	0.7	2	0
strong	70	81	33.9	0.6	0.5	0.6	0.5	0	0.7	32	8
strong	30	89	34.2	0.6	0.5	0.7	0.5	0	0.7	9	0
Effect heterogeneity for earnings (empirical heterogeneity)											
moderate	70			11.2	10.7	11.0	11.8	0.2	15.6		
moderate	30			11.5	11.0	11.8	12.2	-0.3	16.1		
strong	70			11.5	11.0	11.6	12.0	-0.1	16.0		
strong	30			12.1	11.3	12.9	12.7	-0.8	16.9		

Note: 'st.diff': standardized difference defined as mean difference normalized by the square root of the sum of the estimated variances of the particular variables in both subsamples (Imbens and Wooldridge 2009, p. 24). 'std': standard deviation. Pseudo-R^2 is the so-called Efron's R^2: $1 - \sum_{i=1}^{n}[D_i - \Phi(X_i\tilde{\beta})]^2 / \sum_{i=1}^{n}[D_i - n^{-1}\sum_{i=1}^{n} D_i]^2$. For earnings, 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: share of dropped units in % due to support problems for DGPs with 500 or 2000 observations (Section 2.3). Since the statistics for both the probit model and trimming do not depend on the outcomes, they are indicated for employment only.

form of heterogeneity is generated by computing the outcome differences in the population between each individual and its nearest neighbor in the opposite training group. We term this as "empirical heterogeneity" in the remainder of this article.

Modeled heterogeneity is implemented in the following way. 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,

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

$\bar{X}_i$ 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.

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
	Moderate selection, 30% treated								Moderate 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 (modeled heterogeneity)															
	Moderate selection, 30% treated								Moderate 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
	Effect heterogeneity (empirical heterogeneity)															
	Moderate selection, 30% treated								Moderate selection, 70% treated							
IPW	0.2	5.6	13.5	136	0.1	2.7	9.3	64	0.4	5.8	20.4	138	−0.1	2.8	−1.9	72
PM	0.2	7.3	12.8	182	0.2	3.5	10.0	87	0.0	8.1	11.2	204	−0.1	3.6	−3.0	90
R1.5	0.3	6.4	14.1	156	0.2	3.1	10.1	76	0.1	6.8	12.1	167	−0.1	3.2	−2.9	79
R3	0.3	6.3	15.8	154	0.2	3.1	10.0	76	0.1	6.6	13.8	161	0.0	3.1	−2.3	78
R1.5BC	0.0	6.4	9.9	151	0.1	3.0	8.7	73	−0.4	6.8	4.6	158	−0.1	3.2	−3.9	75
R3BC	0.0	6.3	9.5	150	0.1	3.0	8.4	73	−0.4	6.7	4.2	155	−0.1	3.1	−4.1	75
	Strong selection, 30% treated								Strong selection, 70% treated							
IPW	−0.2	6.9	34.9	178	0.1	3.4	−1.4	99	1.7	7.5	77.2	168	0.7	4.4	47.6	114
PM	−0.4	9.8	25.7	263	0.1	4.9	−2.5	145	1.2	11.0	63.2	254	0.3	7.3	29.6	197
R1.5	−0.4	8.3	27.0	220	0.0	4.1	−5.4	119	1.3	9.4	65.9	213	0.4	5.9	35.0	152
R3	−0.3	8.1	29.2	211	0.0	4.0	−4.6	115	1.4	8.9	67.2	202	0.4	5.5	35.2	141
R1.5BC	−0.8	8.3	21.8	206	−0.1	4.1	−1.0	109	0.7	9.4	56.0	202	−0.1	6.0	25.6	142
R3BC	−0.7	8.1	22.0	201	0.0	4.0	−1.2	107	0.7	9.1	56.2	196	−0.1	5.7	24.7	135

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).

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 scenarios that are considered in the EMCS and gives statistics about the strength of selection implied by each. The standardized differences as well as the pseudo-R^2s are based on a reestimated propensity score in the actually nontreated 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 nontreated is changed. This leads (together with the fact that the treatment share differs from the original share) to

Table 5. Coverage probabilities, German data

	Homogeneity						Modeled heterogeneity						Empirical heterogeneity					
	binary			continuous			binary			continuous			binary			continuous		
	as	bs	wbs	as	bs	wbs	as	bs	wbs	as	bs	wbs	as	bs	wbs	as	bs	wbs
IPW																		
GMM	99.9	95.4		99.7	95.1		100.0	95.6		99.8	95.7		99.9	95.3		99.5	95.2	
wgt uncond var	99.6	94.8		99.3	94.5		99.7	94.9		99.3	95.0		99.2	94.8		98.7	94.7	
wgt decomp (κ 0.2)	99.6	94.8		99.3	95.1		99.7	95.1		99.3	95.6		99.2	94.7		98.7	95.1	
wgt decomp (κ 0.8)	99.6	94.8		99.3	95.1		99.7	95.1		99.3	95.6		99.2	94.7		98.6	95.0	
wgt decomp (κ 3.2)	99.5	94.9		98.9	95.1		99.6	95.1		98.8	95.6		99.0	94.8		98.0	94.8	
wgt A (κ 0.2)	97.0	95.7		97.0	95.8		96.8	96.1		97.1	96.2		95.5	96.0		95.2	95.7	
wgt A (κ 0.8)	97.0	95.5		96.9	95.6		96.9	95.9		97.1	96.1		95.5	95.7		95.1	95.5	
wgt A (κ 3.2)	97.2	95.2		96.3	95.2		97.1	95.5		96.4	95.6		95.8	95.2		94.1	94.9	
boot effect se		96.2			96.1			96.6			96.6			96.4			95.9	
boot effect quant		96.1			96.1			96.6			96.5			96.3			96.0	
Pair matching																		
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	99.7	89.2	97.4	98.7	87.4	96.5
wgt decomp (κ 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	99.7	90.4	95.9	98.7	92.1	95.5
wgt decomp (κ 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	99.7	90.4	95.9	98.5	92.1	95.4
wgt decomp (κ 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	99.5	91.0	95.8	97.5	91.9	95.1
wgt A (κ 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	95.6	92.8	95.9	94.5	92.9	95.4
wgt A (κ 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	95.7	92.5	95.8	94.3	92.5	95.3
wgt A (κ 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	96.0	91.7	95.8	93.2	92.3	95.2
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	94.8	97.8	97.3	94.4	97.5	96.7
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	99.4	88.1	97.9	97.6	86.0	97.1
wgt decomp (κ 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	99.3	89.8	96.4	97.5	91.7	96.0
wgt decomp (κ 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	99.3	89.8	96.3	97.2	91.6	95.9
wgt decomp (κ 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	98.9	90.3	96.3	95.5	91.5	95.9
wgt A (κ 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	92.4	92.6	97.0	90.2	92.6	96.4
wgt A (κ 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	92.5	92.1	97.0	89.9	92.4	96.3
wgt A (κ 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	93.0	91.5	96.9	88.2	92.0	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	89.2	97.0	97.4	88.3	96.6	96.7
boot effect se		97.9	97.2		97.5	96.8		98.0	96.8		97.4	96.8		97.9	96.5		97.4	95.9
boot effect quant		97.9	97.2		97.5	97.0		98.0	96.9		97.6	96.9		97.8	96.7		97.3	96.3
Radius matching																		
wgt uncond var	99.7	93.9		99.2	93.9		99.7	94.1		99.2	94.0		99.3	93.8		98.8	93.5	
wgt decomp (κ 0.2)	99.7	94.1		99.3	94.6		99.7	94.3		99.2	94.7		99.4	94.0		98.8	94.1	
wgt decomp (κ 0.8)	99.7	94.0		99.2	94.6		99.7	94.3		99.2	94.6		99.3	94.0		98.7	94.0	
wgt decomp (κ 3.2)	99.6	94.0		98.7	94.3		99.6	94.3		98.7	94.3		99.2	94.0		98.0	93.8	
wgt A (κ 0.2)	96.5	95.0		96.8	95.2		96.3	95.3		96.6	95.4		95.3	95.1		95.1	94.8	
wgt A (κ 0.8)	96.5	94.8		96.6	95.1		96.4	95.0		96.5	95.1		95.4	95.0		94.9	94.6	
wgt A (κ 3.2)	96.7	94.5		96.0	94.6		96.6	94.8		95.9	94.7		95.7	94.6		94.1	94.1	
wgt uncond var ps	99.0	93.0		97.5	92.5		99.1	93.2		97.7	92.8		98.8	93.2		97.2	92.8	
wgt decomp (κ 0.2) ps	99.0	93.1		97.6	93.6		99.1	93.4		97.7	93.8		98.8	93.4		97.3	93.7	
wgt decomp (κ 0.8) ps	99.0	93.0		97.5	93.5		99.0	93.4		97.5	93.7		98.8	93.4		97.1	93.7	
wgt decomp (κ 3.2) ps	98.6	93.1		96.3	93.0		98.8	93.5		96.4	93.3		98.4	93.5		95.9	93.3	
wgt A (κ 0.2) ps	91.8	94.2		91.2	94.2		91.9	94.7		91.5	94.4		91.7	95.1		90.5	94.6	
wgt A (κ 0.8) ps	91.9	94.0		90.9	93.9		92.0	94.5		91.2	94.2		91.8	94.9		90.2	94.4	
wgt A (κ 3.2) ps	92.3	93.6		89.6	93.3		92.5	94.1		89.8	93.6		92.2	94.5		88.9	93.8	
boot effect se		96.8			96.5			97.0			96.8			96.8			96.2	
boot effect quant		96.7			96.6			97.0			96.8			96.7			96.2	

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, see equation (27) of Section 3.4. Exceptions are 'boot effect se', which bootstraps the effect and plugs its standard deviation 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 rather than the standard bootstrap is used for the respective method.'Wgt' is approximation based on weights (Section 3.3). 'uncond var', 'decomp', and 'A' are based on equations (16) (unconditional variance), (17) (decomposition), and (22), respectively. κ (0.2, 0.8, 3.2) gauges the number of nearest neighbors in equation (21). 'Abadie Imbens' is the approximation of Abadie and Imbens (Section 3.2). 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 6. Coverage probabilities, NSW/PSID data

	Emp sel		Weak sel		Emp sel			Weak sel			Emp sel		Weak sel	
	as	bs	as	bs	as	bs	wbs	as	bs	wbs	as	bs	as	bs
	IPW				Pair matching						Radius matching			
GMM	96.6	96.2	99.7	95.1										
wgt uncond var	95.2	95.9	98.8	95.2	94.8	90.6	94.9	99.3	93.1	97.2	95.0	94.8	99.3	95.4
wgt decomp (κ 0.2)	96.3	96.3	98.8	95.2	97.7	90.9	94.8	99.5	92.8	97.2	97.1	95.4	99.4	95.3
wgt decomp (κ 0.8)	96.6	96.2	98.8	95.2	98.3	85.3	94.7	99.6	92.8	97.2	97.8	95.6	99.4	95.3
wgt decomp (κ 3.2)	98.5	96.2	99.0	95.1	99.9	91.3	94.6	99.8	92.3	97.1	99.3	94.8	99.6	95.4
wgt A (κ 0.2)	94.7	96.7	98.8	95.3	95.4	91.9	94.9	98.6	94.6	97.2	94.9	96.0	98.9	95.9
wgt A (κ 0.8)	95.0	96.4	98.8	95.3	96.0	88.5	94.7	98.7	94.6	97.2	95.6	95.8	98.9	95.8
wgt A (κ 3.2)	96.3	96.2	99.0	95.2	98.4	91.7	94.5	98.8	94.6	97.2	97.1	95.0	99.1	95.8
Abadie Imbens					93.2	98.1	96.1	97.2	93.9	96.9				
wgt uncond var ps					93.4	89.8	95.2	98.6	90.6	96.3	93.0	94.1	98.2	92.5
wgt decomp (κ 0.2) ps					96.9	90.2	94.9	99.0	90.5	96.3	95.7	94.9	98.5	92.5
wgt decomp (κ 0.8) ps					97.9	84.2	94.8	99.2	90.6	96.3	97.0	95.1	98.6	92.6
wgt decomp (κ 3.2) ps					99.9	91.1	94.6	99.5	90.2	96.4	99.0	94.4	99.1	93.1
wgt A (κ 0.2) ps					94.2	91.2	95.1	97.3	91.7	96.0	92.9	95.5	97.4	92.8
wgt A (κ 0.8) ps					94.8	87.6	94.9	97.5	91.8	96.1	94.0	95.3	97.5	92.9
wgt A (κ 3.2) ps					98.0	91.3	94.6	97.8	92.0	96.2	95.8	94.4	97.8	93.0
Abadie Imbens ps					91.3	97.4	95.8	94.2	89.9	95.1				
boot effect se		96.3		95.7		97.3	94.6		97.8	97.3		96.6		96.8
boot effect quant		96.3		95.6		97.3	94.8		97.8	97.4		96.5		96.7

Note: 'emp sel': empirical selection. 'weak sel': weaker selection due to dropping observations violating common support. '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, see equation (27) of Section 3.4. Exceptions are 'boot effect se', which bootstraps the effect and plugs its standard deviation 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 rather than the standard bootstrap is used for the respective method.'Wgt' is approximation based on weights (Section 3.3). 'uncond var', 'decomp', and 'A' are based on equations (16) (unconditional variance), (17) (decomposition), and (22), respectively. κ (0.2, 0.8, 3.2) gauges the number of nearest neighbors in equation (21). 'Abadie Imbens' is the approximation of Abadie and Imbens (Section 3.2). 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).

different values of those statistics even for the case mimicking selection in the full sample. Combined with two sample sizes, we run all in all 24 simulations. The Monte Carlo simulations consist of 10,000 replications for the smaller and 1000 for the larger sample size, as the latter is computationally more expensive, but has less variability in results across simulation samples.

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 nonzero 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 in-between these cases. Considering the standard deviations of the estimators across DGPs, we find that they by and large are in line with $\sqrt{n}$-convergence for the case of moderate selection: standard deviations 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 tends to be smaller in many cases, indicating that the normal distribution may not be a good approximation of the distribution of the estimators for the sample sizes considered. This is likely driven by common support issues related to the higher selection into treatment, as indicated in Table 3 by the share of units dropped due to the trimming rule outlined in Section 2.3.

Khan and Tamer (2010) showed that weak overlap may cause a slower convergence rate due to propensity scores not strictly bounded away from zero uniformly in X. Their findings suggest that conventional asymptotics may not provide a good approximation when the overlap is weak. Relating these theory-based results to our analysis, a stricter support assumption than Equation (3) would be required for $\sqrt{n}$-convergence in specifications with a strong selectivity. The slower semiparametric convergence rate can be achieved by the weaker support restriction in Equation (3), whereas the faster parametric rate requires the stronger condition on the propensity scores being strictly bounded away from zero uniformly in X.

4.2 Simulations Based on U.S. NSW/PSID Data

Our second set of simulations exploits the same design as considered in Section 4 of Busso, DiNardo, and McCrary (2014) and is based on data from the National Supported Work (NSW) Demonstration program and the Panel Study of Income Dynamics (PSID) in the U.S. The sample consists of 780 African Americans, namely 156 program participants from the experimental NSW data and 624 nonparticipants from the PSID. The available covariates include age, years of education, indicators for being married, a high school dropout, and unemployed in 1974 and 1975, earnings in thousand USD

in 1974 and 1975 and squares thereof, and interactions between the unemployment indicators in 1974 and 1975 as well as the earnings in both periods. The outcome variable shows earnings in thousand USD in 1978. Descriptive statistics of these data are presented in Table A.1 in the online Appendix.

The construction of the 10,000 simulation samples comprising 780 observations involves the following steps. First, covariates without interaction and higher order terms are simulated: indicator variables for marital status and unemployment are drawn from the empirical distribution in the original data, while age, education, and earnings are drawn from multivariate normal distributions within groups defined upon the indicator variables. The group-specific multivariate normal distributions reflect the empirical means and covariances of age, education, and earnings (in 1974 and 1975) in the original sample. Then, group-specific minima and maxima on earnings are imposed and only the integer parts of age and education are retained. Second, a threshold crossing model for the treatment is constructed based on the simulated covariates, their previously mentioned higher order and interaction terms, and an error term drawn from the normal distribution. The coefficients in the threshold-crossing model are obtained by a probit regression of the treatment in the original sample. Third, separate models for the potential outcomes under treatment and nontreatment are constructed as linear functions of the simulated covariates, interactions, higher order terms, and a normally distributed error term. The coefficients in these models are estimated among treated and nontreated observations, respectively, in the original sample and the variance of the error term in either model is set to the mean squared error of the respective regression in the original sample. Finally, the simulated outcome is equal to one of the potential outcomes as a function of the simulated treatment state. We refer to Busso, DiNardo, and McCrary (2014) for more details on the simulation design.

The lack of common support is a well-known problem of this dataset heavily used in the research of program evaluation (which is also shown by the statistics in Table A.1). Such support problems imply that the treated and nontreated subgroups are not comparable to each other. In line with Busso, DiNardo, and McCrary (2014), we address this issue by analyzing an additional DGP based on a treatment selection model with a linear index scaled by a factor of one-fifth (to move the propensity score distribution away from the boundaries). This setting forces overlap is referred to as weak selection (compared to the empirical selection without scaling the probit coefficients).

Table A.2 of the online Appendix provides descriptive statistics on the DGPs with both empirical and weak selection. These statistics show that the weak selection model improves the overlap considerably (the standardized difference of the p-scores between treated and nontreated reduces from 210 to 38). Table A.3 provides evidence on the heterogeneous treatment effects under both selection settings. The results show that the standard deviations of ATETs based on the empirical selection are approximately twice as high as those based on the weak selection.

5. RESULTS

This section evaluates the performance of the various inference methods for the different treatment effect estimators w.r.t. their coverage rates of the true effect. For the sake of brevity, we present only a limited amount of evidence in the main body of the article which conveys the main message of our findings. An extensive set of further results is presented in online Appendix A.

Table 5 provides the coverage rates of the inference procedures by effect estimators and outcomes for the German register data, that is, the share of simulations in which the true value is included in the 95% confidence interval of the respective method. 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 coverage rates are averaged over the four estimators investigated (R1.5, R3, R1.5BC, R3BC), because their coverage rates are qualitatively very similar, see Table A.4 in online Appendix A for a separate analysis of each radius matching algorithm. Furthermore, the results in Table 5 are averages over the different DGP features, with the exception of effect homogeneity (left panel) vs. modeled heterogeneity vs. empirical heterogeneity (right panel). Tables A.9, A.10, and A.11 in online Appendix A provide the coverage rates separated by DGP features (sample size, share of treated, strength of selection, and outcome distribution). The coverage rates within the various simulation designs are frequently quite comparable to those of the average coverage rates presented in Table 5. Table 6 provides the coverage rates for the two simulation designs based on the NSW/PSID data previously analysed by Busso, DiNardo, and McCrary (2014). Here, the results are provided separately for both simulation designs with empirical selection (emp sel) vs. weak selection (weak sel), see Section 4.2 for details, and organized in three different panels for IPW (left), pair matching (center), and radius matching (right).

Considering Table 5 with the results for the German register data, we find that any inference method for IPW (upper panel) that is based on asymptotic approximations (i.e., does not rely on bootstrapping), see the columns "as," is conservative, such that coverage exceeds the nominal size of 95%. Interestingly, this is the case for both the GMM-based variance estimator ("GMM"), which accounts for first step estimation of the propensity score, and for the various weighting-based approaches outlined in Section 3.3 ("wgt"), that ignore the first step. When considering the NSW/PSID data, see Table 6, over-coverage among asymptotic approximations for IPW is generally less severe, albeit still present for most methods, in particular when considering the DGP with weak selection. By and large, the bootstrap methods (column "bs") come closer to the nominal size of IPW than the asymptotic methods, maybe with the exception of the case of the NSW/PSID data with empirical selection. This result does not only hold when bootstrapping t-statistics based on the asymptotic variance approximations. It is frequently also the case when bootstrapping the effect for either plugging its standard deviation into the asymptotic approximation for confidence intervals ("boot effect se") or for obtaining confidence intervals based on the quantiles of the bootstrapped effects ("boot effect quant").

We next investigate pair matching, starting with the asymptotic approximations ("as"). Weighting using the unconditional variance based on (16) ("wgt uncond var") or the decomposition approach based on (22) ("wgt decomp"), where κ gauges the number of nearest neighbors in (21), entails over-coverage in

Table 5. Using an estimate of part A of the decomposition only, see (22), comes generally somewhat closer to the nominal size ("wgt A"). For the scenarios in Table 6, however, the coverage rates within the weighting approaches sometimes deviate from this pattern. Notably, all weighting approaches are conservative under weak selection. Adjusting for the first step propensity estimation, see the suffix "ps," somewhat, but not fully mitigates the over-coverage of weighting based on the unconditional variance and the decomposition in cases where these methods are too conservative. In contrast, the propensity score adjustment often entails under-coverage of weighting based on term A only, albeit the picture is again less clear for the DGPs in Table 6. The asymptotic approximation of (Abadie and Imbens 2006) ignoring propensity score estimation performs decently across DGPs, while the propensity score-adjusted version of (Abadie and Imbens 2016) is generally prone to under-coverage.

As discussed in (Abadie and Imbens 2008), the standard bootstrap ("bs") is inconsistent because even in large samples, it does not reproduce the distribution of how frequently nontreated observations are used as matches for treated observations. The authors illustrate the issue by the case of a small ratio of treated to nontreated observations, implying that the probability that a non-treated unit is used as a match more than once in the data is low. In bootstrap samples, however, a particular treated observation can be sampled several times, such that the corresponding most comparable nontreated observation is matched several times, too. The bias in variance estimation derived in the example of (Abadie and Imbens 2008) does, however, not one-to-one carry over to our simulations, which are based on different DGPs. In our settings, a range of (standardly) bootstrapped statistics (including the effects) entail coverage rates not too far from nominal size, while in other cases, under-coverage is non-negligble. In line with (Abadie and Imbens 2008), the magnitude of under-coverage generally depends on the share of treated observations, see Table A.10 in online Appendix A. Contrary to the standard bootstrap, the wild bootstrap ("wbs") is consistent and has in most settings a coverage rate that is closer to nominal size. We therefore clearly recommend the wild bootstrap over the standard bootstrap in the case of pair matching.

When looking at radius matching, the results for the asymptotic approximations ('as') resemble the ones of pair matching. In Table 5, weighting based on the unconditional variance ('wgt uncond var') or on the decomposition approach ('wgt decomp') is conservative, while using part A of the decomposition ('wgt A') is closer to the nominal size. The pattern is, however, less clear in Table 6, but any weighting approach is conservative under weak selection. In Table 5, adjusting for propensity estimation ('ps') partly offsets the over-coverage of weighting based on the unconditional variance and the decomposition, but may entail under-coverage of weighting based on term A. No clear pattern arises from Table 6. Bootstrap methods ('bs') generally fare quite satisfactorily and are not too far from nominal size for any of the scenarios considered.

All in all, our results suggest that for IPW and radius matching, coverage rates using the standard bootstrap, either for bootstrapping t-statistics based on asymptotic variance approximations or for bootstrapping the effects, are more accurate than the frequently conservative asymptotic approximations. For pair matching, the standard bootstrap performs well in some cases, but is prone to under-coverage in others. We therefore recommend using the wild bootstrap, which is generally closer to nominal size than the standard bootstrap as well as the (conservative) asymptotic variance approximations.

Any of the bootstrap results reported in this section are based on 199 replications. In Tables A.6, A.7, and A.8 in online Appendix A we vary the number of replications between 49 and 199 and provide (i) the coverage rates when pooling all simulations as well as (ii) the rejection rates at the 5% significance level for the German data with effect homogeneity, where the null of no effect holds. The results on the rejection rates are given separately without and with smoothing the bootstrap-based p-values, see (26). By and large, it seems that reliable inference is obtained already with just 49 bootstraps, no matter whether non-smoothed or smoothed statistics are considered and gains from increasing the number of bootstraps appear to be rather small. For the nonsmoothed 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 difference is rather minor. The difference is close to nonexistent for the smoothed versions, as smoothing decreases the standard deviations under a low number of bootstraps somewhat such that an increase in the number of replications does not entail further reductions. We therefore see in Tables A.6, A.7, and A.8 that smoothing has some effect on the rejection frequencies when the number of bootstrap replications is low, but no longer when it is increased.

6. CONCLUSION

In this article, we investigated the finite sample properties of various inference methods for three classes of propensity score-based estimators of the average treatment effect on the treated: inverse probability weighting (IPW), pair matching, and radius matching. Using empirically motivated simulation designs based on German register data as well as on U.S. NSW/PSID data previously considered by (Busso, DiNardo, and McCrary 2014), we analysed both asymptotic approximations and bootstrap methods for the computation of variances, confidence intervals, and p-values. We found that asymptotic approximations that ignored the first step estimation of the propensity score frequently tended to be conservative, entailing over-coverage of the true effect, albeit the (Abadie and Imbens 2006) variance estimator for pair matching generally was a noticeable exception. GMM-based variance estimation of IPW was mostly conservative, too, even though accounting for propensity score estimation. In general, we found that accounting for propensity score estimation in asymptotic approximations only partially mitigated over-coverage of some inference methods, while it entailed even under-coverage of others.

Appropriately implemented bootstrap procedures (accounting for the different consistency requirements of pair matching and "smoother" treatment effect estimators) frequently outperformed the asymptotic approximations. For IPW and radius matching, coverage rates using the standard bootstrap for either bootstrapping t-statistics based on asymptotic variance approximations or for bootstrapping the effects came closer to nominal size than the conservative coverage rates (exclusively) based on asymptotic approximations. For pair matching, the standard

bootstrap performed well in some cases, but was prone to under-coverage in others. In contrast, the wild bootstrap came generally closer to nominal size than the standard bootstrap as well as the (conservative) asymptotic variance approximations. We therefore recommend only using bootstrap procedures that are theoretically justified for the treatment effect estimator at hand. Finally, we found only minor effects of the number of bootstrap replications on the performance of bootstrap-based inference procedures in terms of coverage and rejection rates.

[Received June 2016. Revised April 2018.]

REFERENCES

Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267. [184,187,188,199]

——— (2008), "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, 76, 1537–1557. [184,190,199]

——— (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29, 1–11. [184,186]

——— (2012), "A Martingale Representation for Matching Estimators," *Journal of the American Statistical Association*, 107, 833–843. [184,191]

Abadie, A., and Imbens, G. W. (2016), "Matching on the Estimated Propensity Score," *Econometrica*, 84, 781–807. [183,184,187,188,190,199]

Advani, A., and Sloczyński, T. (2013), "Mostly Harmless Simulations? On the Internal Validity of Empirical Monte Carlo Studies," IZA Discussion Paper No. 7874. [184]

Andrews, D. W., and Buchinsky, M. (2000), "A Three-Step Method for Choosing the Number of Bootstrap Repetitions," *Econometrica*, 68, 23–51. [190]

Bodory, H., Camponovo, L., Huber, M., and Lechner, M. (2016), "A Wild Bootstrap Algorithm for Direct and Propensity Score Matching Estimators," SEPS Discussion paper. [184,187,191]

Busso, M., DiNardo, J., and McCrary, J. (2014), "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *Review of Economics and Statistics*, 96, 885–897. [183,184,185,197,198,199]

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programmes," *Journal of American Statistical Association*, 94, 1053–1062. [183,186]

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26. [190]

Frölich, M. (2004), "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," *The Review of Economics and Statistics*, 86, 77–90. [183,186]

Frölich, M., Huber, M., and Wiesenfarth, M. (2017), "The Finite Sample Performance of Semi- and Non-Parametric Estimators for Treatment Effects and Policy Evaluation," *Computational Statistics & Data Analysis*, 115, 91–102. [183,191]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [185,187]

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054. [187]

Heckman, J. J., Ichimura, H., and Todd, P. (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294. [183]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [183,185]

Horowitz, J. L. (2001), "The Bootstrap," in *Handbook of Econometrics*, eds. J. J. Heckman, and E. Learer, North-Holland, pp. 3159–3228. [190]

Horvitz, D., and Thompson, D. (1952), "A Generalization of Sampling Without Replacement From a Finite Population," *Journal of American Statistical Association*, 47, 663–685. [183,185]

Huber, M., Lechner, M., and Steinmayr, A. (2015), "Radius Matching on the Propensity Score With Bias Adjustment: Tuning Parameters and Finite Sample Behaviour," *Empirical Economics*, 49, 1–31. [186,191]

Huber, M., Lechner, M., and Wunsch, C. (2013), "The Performance of Estimators Based on the Propensity Score," *Journal of Econometrics*, 175, 1–21. [183,184,186]

Imbens, G. W. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29. [183,185,186]

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [183,185,186]

Khan, S., and Tamer, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042. [185,197]

LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, 76, 604–620. [184]

Lechner, M. (2002a), "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. [184]

——— (2002b), "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society*, Series A, 165, 59–82. [187,188]

Lechner, M., Miquel, R., and Wunsch, C. (2011), "Long-Run Effects of Public Sector Sponsored Training in West Germany," *Journal of the European Economic Association*, 9, 742–784. [184,186,190]

Lechner, M., and Strittmatter, A. (2017), "Practical Procedures to Deal With Common Support Problems in Matching Estimation," *Econometric Reviews* (Forthcoming). [186,191]

Lechner, M., and Wunsch, C. (2013), "Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables," *Labour Economics*, 21, 111–121. [184,191]

Lunceford, J. K., and Davidian, M. (2004), "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study," *Statistics in Medicine*, 23, 2937–2960. [183]

MacKinnon, J. G. (2006), "Bootstrap Methods in Econometrics," *The Economic Record*, 82, S2–S18. [190]

Millimet, D., and Tchernis, R. (2009), "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies," *Journal of Business & Economic Statistics*, 27, 297–315. [186]

Newey, W. K. (1984), "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14, 201–206. [187]

Otsu, T., and Rai, Y. (2015), "Bootstrap Inference of Matching Estimators for Average Treatment Effects," Working paper, University of St. Gallen. [184,191]

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. [183]

Racine, J. S., and MacKinnon, J. G. (2007), "Inference via Kernel Smoothing of Bootstrap P Values," *Computational Statistics & Data Analysis*, 51, 5949–5957. [184,186,190]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [185]

——— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38. [183,186]

Rubin, D. B. (1973), "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183. [186]

——— (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [185]

——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328. [186]

——— (1990), "Formal Modes of Statistical Inference For Causal Effects," *Journal of Statistical Planning and Inference*, 25, 279–292. [185]

Smith, J., and Todd, P. (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125, 305–353. [185]

Waernbaum, I. (2012), "Model Misspecification and Robustness in Causal Inference: Comparing Matching With Doubly Robust Estimation," *Statistics in Medicine*, 31, 1572–1581. [186]

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. [183]

——— (2008), "Sensitivity of Propensity Score Methods to the Specifications," *Economics Letters*, 98, 309–319. [186]