AStA Wirtsch Sozialstat Arch (2017) 11:205–231
https://doi.org/10.1007/s11943-017-0212-x

ORIGINALVERÖFFENTLICHUNG

Augmenting propensity score equations to avoid misspecification bias – Evidence from a Monte Carlo simulation

Gerhard Krug

Received: 1 February 2017 / Accepted: 21 November 2017 / Published online: 4 December 2017

Abstract Propensity score matching is a semi-parametric method of balancing covariates that estimates the causal effect of a treatment, intervention, or action on a specific outcome. Propensity scores are typically estimated using parametric models for binary outcomes, such as logistic regression. Therefore, model specification may still play an important role, even if the causal effect is estimated nonparametrically in the matched sample. Methodological research indicates that incorrect specification of the propensity score equation can lead to biased estimates. Augmenting the propensity score equation with terms that represent potential nonlinearity and nonadditivity, as proposed by Dehejia and Wahba and more recently by Imbens and Rubin, represents a means of avoiding such bias. Here, we conduct a Monte Carlo simulation and show that the misspecification bias is rather small in many situations. However, when the propensity score equation and/or the outcome equation are characterized by strong nonlinearity and nonadditivity, the misspecification bias can be severe. Augmentation is shown to reduce this bias, often substantially. The Dehejia-Wahba (2002) algorithm performs better than the Imbens-Rubin algorithm, especially when the outcome equation is strongly nonlinear and nonadditive.

Keywords Causal inference · Propensity score matching · Propensity score stratification · Misspecification bias · Monte carlo simulation

G. Krug (✉)
Institute for Employment Research (IAB), Regensburger Straße 104, 90478 Nuremberg, Germany
E-Mail: gerhard.krug@iab.de

G. Krug
Chair for empirical Economic Sociology, University of Erlangen-Nuremberg (FAU),
Findelgasse 7/9, 90402 Nuremberg, Germany

Erweiterung der Propensity Score Gleichung zur Vermeidung von Fehlspezifikationen? Eine Monte Carlo Simulation

Zusammenfassung Propensity Score Matching ist eine semi-parametrische Methode zur Drittvariablenkontrolle bei der Schätzung kausaler Effekt eines Treatments, einer Intervention oder einer Handlung auf eine bestimmte Zielvariable. Propensity-Scores werden typischerweise unter Verwendung parametrischer Modelle für binäre Ergebnisse geschätzt, etwa der logistischen Regression. Daher stellt sich trotzdem die Frage der korrekten Modellspezifikation, selbst wenn der kausale Effekt in der gematchten Stichprobe nichtparametrisch geschätzt wird. Studien zeigen, dass eine falsche Spezifikation der Propensity-Score-Gleichung zu verzerrten Schätzungen führen kann. Um solche Verzerrungen zu vermeiden, haben Dehejia und Wahba und kürzlich Imbens und Rubin Algorithmen zur Anreicherung der Propensity-Score-Gleichung mit Termen vorgeschlagen, welche eine potenzielle Nichtlinearität und Nichtadditivität in der Modellspezifikation abbilden sollen. In der vorliegenden Arbeit wird eine Monte-Carlo-Simulation durchgeführt und es zeigt sich, dass die Verzerrung aufgrund von Fehlspezifikation in vielen Situationen eher klein ist. Wenn jedoch die Propensity-Score-Gleichung und/oder die Outcome-Gleichung durch starke Nichtlinearität und Nichtadditivität gekennzeichnet sind, kann die Fehlspezifizierungs-Vorspannung schwerwiegend sein. Anreicherungsalgorithmen reduzieren solche Verzerrungen oft erheblich. Der Dehejia-Wahba Algorithmus scheint hierzu besser geeignet als der Algorithmus von Imbens-Rubin (2015), insbesondere dann, wenn auch die Ergebnisgleichung stark nichtlinear und nichtadditiv ist.

Schlüsselwörter Causal inference · Propensity score matching · Propensity score stratification · Misspecification bias · Monte carlo simulation

1 Introduction

Propensity score matching was developed by Rosenbaum and Rubin (1983a, 1983b) and closely follows Rubin's (1974) framework of potential outcomes. Matching estimators are intended to adjust a given sample of treated and untreated individuals to mimic a random assignment of individuals to a treatment and control group[1]. Therefore, when applying matching techniques, researchers are conducting a "hypothetical randomized experiment" (Rubin 1986). Nevertheless, statistics does not provide guidelines as to whether the sociological subject matter lends itself to such an interpretation; the researcher and reader must make this decision, and different research traditions may come to different conclusions (see also the discussion in Holland 1986). Matching estimators have been successfully applied to a variety of research questions, including those in sociology, economics, and other social science fields. Propensity score matching is one of the most commonly applied matching estimators in the field.

[1] For clarity, we focus on the binary case; however, matching estimators have also been generalized to non-binary treatments (cf., Imai and van Dyk 2004).

In contrast to regression methods, propensity score matching is considered a non-parametric method because it does not require the choice of a functional form. Indeed, as in randomized experiments, only a comparison of means is needed after matching on the propensity score. However, in most applications, the true propensity score is unknown and must be estimated from the data. Because the true propensity score is typically estimated using parametric estimators of treatment participation, such as probit and logit models, some scholars refer to propensity score matching as a semi-parametric method (Huber et al. 2012). As demonstrated by recent methodological studies, incorrect specification of the propensity score equation can lead to serious bias. Zhao (2008) investigated the effects of over- and under-specifying the propensity score equation on bias and found that, in either case, the causal effect "is insensitive to specification of the propensity score" (Zhao 2008, p. 313). However, Zhao's own findings do not fully support this claim; rather, they indicate that matching without replacement on an under-specified propensity score induces bias in two out of three cases. In a Monte Carlo simulation, Millimet and Tchernis (2009) find that over-specifying does not induce bias, but under-specifying the propensity score equation does. In an additional empirical application using real data, Millimet and Tchernis found that the causal effect was sensitive to the specification of the propensity score.

Because the true propensity score is unknown in most applications, there is no way for applied researchers to know whether a model is misspecified. For example, applied researchers seldom acknowledge that estimating the propensity score by logistic regression follows a different set of rules than employing logistic regression to test hypotheses. Whereas applied researchers appear to believe that the estimating the propensity score coincides with explaining the choice for or against participation in the treatment group, methodologists Dehejia and Wahba (2002: 161) note that "the role of the propensity score is only to reduce the dimensions of the conditioning; as such it has no behavioral assumptions attached to it". Thus, the usual goodness-of-fit tests do not provide meaningful information on how well the matching eliminates the influence of covariates[2] (Rubin 2004). In fact, variable selection based on goodness-of-fit tests or model-building algorithms (e. g., forward step-wise regression) often lead to inefficient estimates (Brookhart et al. 2006). Similarly, certain caveats regarding, e. g., multicollinearity do not apply, and tests for significance are only marginally informative on whether to include a covariate in the estimation (c.f., Harding 2003; Rubin 2004). However, because specific guidelines regarding model specification in the context of propensity score matching are scarce, the topic is rarely addressed in applied research. The point we wish to make refers to an even more subtle aspect of model specification than which covariates to include. Here, we draw attention to the issue of correctly specifying the propensity score equation, rather than to the consequences of including or excluding specific covariates.

[2] Rubin (2004, p. 855) distinguishes between "diagnostics for the successful prediction of probabilities and parameter estimates underlying those probabilities, possibly estimated using logistic regression" (e. g., goodness-of-fit tests), and "diagnostics for the successful design of observational studies based on estimated propensity scores, possibly estimated using logistic regression" (e. g., balancing tests). The former convey no meaningful information on the quality of the matching procedure, whereas the latter are "a critically important activity" when conducting a propensity score matching analysis.

With this paper, we hope to contribute in several ways to the methodological literature on misspecification of the propensity score. First, whereas most methodological research focuses on the effect of misspecification on propensity score weighting (e. g. Setoguchi et al. 2008), our focus is on propensity score matching. Second, we specifically focus on investigating the performance of an algorithm proposed by Dehejia and Wahba (2002) that is intended to help applied researchers avoid misspecification of the propensity score equation. After its first implementation in the context of propensity score stratification, the algorithm has been severely criticized. In this paper, we propose a modified version in the context of propensity score matching. We argue that these modifications should eliminate the weaknesses of the algorithm while retaining its strengths; that is, it provides researchers with an easy-to-implement way of reducing the danger of bias due to misspecification. Third, we conduct a Monte Carlo simulation to test the performance of the modified Dehejia and Wahba (DW) algorithm. To our knowledge, the algorithm has not been tested in this way. In the Monte Carlo simulation, we compare two alternative proposals, a propensity score stratification approach proposed by Hong (2010) and an alternative augmentation algorithm proposed by Imbens and Rubin (2015).

The paper is structured as follows. Sect. 2 describes propensity score matching and underscores the importance of balancing tests. Sect. 3 introduces the Dehejia and Wahba (2002) algorithm and discusses its weaknesses. Sect. 4 introduces a modified version of the Dehejia–Wahba (2002) algorithm and two alternative estimators that also aim to avoid misspecification bias. Sect. 5 describes a Monte Carlo simulation performed to determine whether the modified DW algorithm avoids bias from misspecified propensity score equations. Sect. 5 presents the results, and Sect. 6 concludes the paper.

2 The propensity score tautology and the importance of balancing tests

In a randomized experiment, individuals are distributed into treatment and control groups, irrespective of their characteristics. Although randomized experiments are becoming increasingly common in many fields of research (e. g., Jackson and Cox 2013), they are often not feasible for reasons ranging from ethical problems to practical issues. In observational data, however, we must eliminate all other causally relevant factors of the outcome ex post to arrive at an unbiased comparison of outcomes from treated and untreated individuals. In other words, we must eliminate the bias introduced by confounding variables to ensure that the observed difference between the treatment and control groups reflects a causal effect. Regression estimators ideally address confounders by including them as conditioning variables and imposing functional form assumptions (e. g., linearity). Matching estimators are an alternative method of addressing confounders without imposing functional form assumptions[3].

The simplest and most intuitive matching estimators rely on exact matching. For each treatment individual, an untreated individual must be identified (and vice

[3] See Harding (2003, p. 687 f) for a short overview of the main advantages of propensity score matching.

versa). In exact matching, both individuals are characterized by identical values of the covariate vector, i. e., the vector of confounding variables, which is denoted **x** hereinafter. Given that the subsequent analysis disregards all individuals for whom no matches can be found, exact matching clearly provides researchers with subsamples in which covariates are distributed equally in the treated and untreated groups. Thereafter, the treatment and control groups are considered "balanced".

Exact matches are more difficult to obtain as more variables are controlled for and become impossible if at least one variable is continuous. Matching solely on the propensity score, rather than the confounders themselves, was developed by Rosenbaum and Rubin (1983a, 1983b) to solve this "curse of dimensionality" (Heckman et al. 1998). The propensity score collapses the information from several variables into a single scalar metric. Here, we focus on the average treatment effect on the treated (att), for which the matching estimator can be expressed as a weighted difference in means (Smith and Todd 2005a):

$$\widehat{att} = \widehat{\delta} = \frac{1}{n_1} \sum_{i \in I_1 \cap CS} t_i^1 - \frac{1}{n_1} \sum_{i \in I_1 \cap CS} \sum_{j \in I_0 \cap CS} w(i, j) t_j^0,$$

where t is a binary treatment indicator; t = 1 denotes the treatment status, and t = 0 denotes the control status. I_1 and I_0 are individuals in the treatment and control groups, respectively; CS denotes the region of common support in the propensity score distributions of both groups; n_1 is the number of individuals in the treatment group within the region of common support; and $w(i, j)$ is the weight given to observation j when it is matched to observation i. Different versions of the matching estimators can be built, depending on the choice of $w(i, j)$. Single nearest-neighbor matching (SNNM) is the most intuitive matching algorithm. In SNNM without replacement, observation j is chosen as a match for observation $_i$ when it is closest to i in terms of the absolute distance between their propensity scores $|p(\mathbf{x}_i) - p(\mathbf{x}_j)|$. SNNM then weighs the outcome of observation j, whose propensity score is closest to that of observation i, with $w(i, j) = 1$ and assigns a weight of $w(i, j) = 0$ to all other control observations, and the causal effect is then calculated. In multiple nearest-neighbor matching (MNNM), a weighted average of two or more observations j is built for each treatment individual. Often, a maximum acceptable distance (caliper) is set to avoid matches in which $p(\mathbf{x}_j)$ is extremely far from $p(\mathbf{x}_i)$, even though it is the nearest neighbor. Observations not in the CS, i. e., treatment group individuals for whom no matching partner is found and control group individuals that are not used as matching partners, are excluded from the analysis.

The propensity score is defined as the *true* probability of being in the treatment group, conditional on all confounding variables. However, the true propensity score is generally unknown and is often estimated using logistic regression (or a similar binary outcome model):

$$\widehat{p}_i\left(\mathbf{x}_i\right) = \widehat{p}_i(t_i = 1|\mathbf{x}_i) = \left(1 + \exp\left(-\left(\mathbf{x}'_i\widehat{\beta}\right)\right)\right)^{-1}$$

For the matching estimator to be unbiased, conditional independence must hold; i. e., the vector **x** must contain all variables that simultaneously influence treatment probability and the outcome of interest ("selection on observables"). Additionally, we need to know which variables to include, and if we rely on an *estimated* propensity score, the propensity score equation must be specified correctly with regard to the functional form (Ho et al. 2007: 218). Because the true propensity score equation is often unknown, it is difficult to correctly specify the functional form. However, Ho et al. note the so-called "propensity score tautology". They argue that a correctly specified score will eliminate differences in the distributions of covariates, even if the estimated propensity score cannot be proven to be correctly specified a priori. If we conduct a balancing test and find the sample to be sufficiently balanced after matching on the propensity score, we have found the correct specification. Thus, balancing tests, rather than goodness-of-fit tests, are of the utmost importance when using propensity score matching.

There are several ways to determine whether a sample obtained from propensity score matching is sufficiently balanced. For example, one can test the equality of means between the treatment and control groups with a t-test, as originally proposed by Rosenbaum and Rubin (1985). However, this approach has a major disadvantage in that t-tests depend on sample size. Matching often reduces sample size, and different procedures lead to different sample sizes. Thus, non-significant test results after matching may occur because the sample has been balanced; moreover, they may occur if matching reduced the sample size but the sample remains imbalanced. Balancing tests based on significance are also criticized because balancing is a property of the specific sample under consideration, rather than the population. If a difference in means in the sample is not significant, it can still be large and therefore lead to bias (Ho et al. 2007)[4]. Another common balancing test involves carrying out the matching procedure and then re-estimating the propensity score equation for the matched sample. If the matching procedure has successfully balanced the covariates, the pseudo-R^2 should be near zero and non-significant. Similar to t-tests, this measure depends on sample size. Large differences between the treatment and control groups can be statistically non-significant in small samples, whereas there is a high probability that even small differences will become statistically significant in large samples.

Computing the standardized difference after matching (Rosenbaum and Rubin 1985) is also a common balancing test. This measure, which is not affected by sample size, is expressed as follows:

$$sdiff = \frac{\bar{x}_{t=0} - \bar{x}_{t=1}}{\sqrt{\left(s^2_{t=0} + s^2_{t=1}\right) * 0.5}}$$

where $\bar{x}_{t=0} - \bar{x}_{t=1}$ is the difference in the mean for covariate x between the treatment and control groups and s^2 indicates the sampling variance for this covariate

[4] There is also another way to look at it, as pointed out by the anonymous reviewer. In balancing tests the desired outcome is to not reject the null hypothesis of no differences, whereas in classical hypothesis tests the desired outcome is to reject the null. In the latter case, uncertainty is making it harder to reject the undesired null, whereas in the former case, uncertainty works in favor of the researcher's goal to obtain a (seemingly) balanced sample.

in each group. The standardized difference is the difference in means in the treatment and control groups, expressed as a percentage of the "average" standard deviation over both groups for each covariate. In their original work, Rosenbaum and Rubin (1985) consider 20% to represent a large bias. Currently, however, biases below 3–5% are considered to indicate sufficient balance in matching analyses (Caliendo and Kopeinig 2008: 48).

3 The Dehejia and Wahba (2002) algorithm for reducing misspecification bias

To improve balance, Rosenbaum and Rubin (1984) were the first to propose refining the propensity score by including squares and interactions. In the appendix to their own paper, Dehejia and Wahba (2002) expanded on this idea and proposed a simple algorithm that can be followed if balance is not reached for a given analysis. Dehejia and Wahba (2002) used results from a randomized experiment (the benchmark causal effect) to compare the performance of different variants of the matching estimator. They were mainly concerned with "whether or not to match with replacement, how many comparison units to match to each treated unit, and (...) which matching mode to choose" (Dehejia and Wahba 2002: 153). However, Dehejia and Wahba (2002) also noted that the specification of the propensity score equation may influence the results and proposed a method of identifying the most appropriate specification.

As described by Diaz and Handa (2006: 325), the DW algorithm "essentially entails adding interaction and higher-order terms to [the] base model until tests for mean differences in covariates between control and comparison units become statistically insignificant" (see also Stuart 2010, p. 7). In detail, the DW algorithm begins by (1) stratifying the sample based on quantiles (e. g., 0–0.2, ..., 0.8–1) of a propensity score estimated with a parsimonious logistic regression (that is, a specification containing only linear terms). The balance is checked within each stratum by applying a t-test for the equality of means. If the covariates are not balanced for some strata (i. e., the t-test is statistically significant), (2) the sample should be divided into finer strata, and a new balancing test should be conducted. If these finer strata are not balanced, they recommend that (3) the logistic regression should be modified by "adding interaction terms and/or higher-order terms of the covariates" (Dehejia and Wahba 2002, p. 161), starting with the least balanced variable, until balance is achieved. Comparing the treatment effects obtained by applying their algorithm to a benchmark estimate from a randomized experiment, Dehejia and Wahba (2002) concluded that their algorithm is successful. Both Dehejia and Wahba (2002) and Rosenbaum and Rubin (1984) discussed their augmentation algorithms in the context of a propensity score *stratification* procedure, not in the context of matching on the propensity score.

The Dehejia and Wahba algorithm has several disadvantages and has thus been criticized. Smith and Todd (2005a) criticized the algorithm's lack of objective criteria for choosing and refining the initial strata. This lack of objective criteria is problematic because smaller strata include fewer cases; thus, the power of the test is lower. As a consequence, t-tests become insignificant, even if the bias is still substantial. Additionally, Smith and Todd (2005b) criticize the use of balancing tests per se because they lack formal criteria for determining when the balance is sufficient. In line with this argument, Lee (2013) demonstrated that balancing tests display size problems. For the DW algorithm, he found that the t-test for balance led to rejection in 23.8% of tested cases, instead of the conventional 5%. To alleviate these high rejection rates, Lee (2013) developed a permutation version of the traditional t-test. This updated test leads to test sizes of 3.5% for the DW algorithm; thus, it is rather conservative. Lee (2013) also considered the standardized difference, applied a 20% threshold for indicating imbalance, and found a rejection rate of nearly 100%, instead of 5%. Unfortunately, the permutation test developed by Lee (2013) is not applicable to standardized differences. However, the applied threshold is two to three times as large as the 3–5% threshold proposed by Caliendo and Kopeinig (2008: 48) and is thus too high; this fact partially explains the poor performance of the standardized difference in Lee (2013). Finally, Iacus et al. (2012: 21) used an empirical example to show that augmenting the propensity score equation can increase bias, as well as decrease it. To our knowledge, except for an unpublished manuscript by Smith and Zhang (2007), no Monte Carlo simulations have been conducted to test the performance of the DW algorithm. Furthermore, a large part of the criticism refers only to the application of the DW algorithm in the context of propensity score stratification and/or propensity score weighting, thus it is not clear if matching estimators are also subject to the same disadvantages.

4 A modified version of the Dehejia and Wahba (2002) algorithm and two recent alternatives

Instead of abandoning the idea of augmenting the propensity score equation altogether, we propose to modify the DW algorithm in a way that eliminates the above problems but keeps the basic idea of augmentation intact. We are not the first to extend the algorithm to propensity score matching. Indeed, some applied researchers already have employed some of the modifications we propose (see, for example, Diaz and Handa 2006: 325). However, we claim that we are the first to present a systematic argument for such modifications and to systematically test the performance of the modified algorithm using Monte Carlo simulations.

The modification that we propose and test begins (1) with a main-effect logit or probit specification. The specification should include all covariates that are necessary to fulfill the conditional independence assumption. (2) Propensity matching, which is performed instead of stratification, is conducted using a standard matching algorithm (e. g., nearest-neighbor matching with a caliper). By restricting the augmentation algorithm to propensity score matching, instead of propensity stratification, the subjectivity in determining the number of strata that was criticized by Smith and Todd (2005a) is eliminated. In addition, Austin (2009) found that propensity score matching outperforms propensity score stratification in terms of producing a balance between treatment and control groups. (3) After matching, we determine the standardized difference for all covariates (rather than performing t-tests), the least balanced variable is identified, and the corresponding standardized difference is recorded. By using standardized differences instead of significance tests, the sensitivity of the balancing checks to sample size is alleviated. (4) Steps 1–3 are repeated several times, and the propensity score equation is augmented each time with another interaction term and/or a higher-order term[5]. This step contrasts with that of the original algorithm, which only augments the equation if a t-test indicates imbalance. Because it is unclear what constitutes sufficient balance, we augment the equation repeatedly and select the specification that produces the best balance. This procedure should also avoid the problem that some augmented specifications reduce balance, rather than improving it (Iacus et al. 2012: 21). (5) Among all of the tested specifications, the specification that has the lowest value for the standardized bias in step three is identified. By defining best balance in terms of the maximum imbalance among all variables, the bias due to the worst balanced variable is minimized. (6) The causal effect is estimated using the specification identified in step 5.

A different route was taken by Imbens and Rubin (2015), who proposed an alternative augmentation algorithm. In contrast to the DW algorithm, they do not rely on the propensity score tautology to select the propensity score equation. Rather, their algorithm is based on step-wise logistic regression. However, the algorithm tests whether each individual covariate should be included in the propensity score equation at all, in addition to augmenting the existing set of covariates to avoid misspecification bias. Therefore, the algorithm conducts a broader specification search than the DW algorithm by including or excluding entire variables, based on the strength of their association with the treatment. As described in Imbens (2015), the algorithm starts with selecting a subset of covariates that should be included in the propensity score equation, irrespective of the strength of their association with the treatment. Additional covariates, as well as additional second-order terms, are then included in the propensity score equation, if they pass a specified threshold value. More specifically, the decision to include an additional term is based on the likelihood ratio test statistic of the null hypothesis that the coefficient from a logistic regression predicting treatment assignment is equal to zero. The threshold values are recommended based on simulation analysis, i. e., 1 for additional covariates and 2.71 for additional second-order terms Imbens (2015).

In contrast to both Dehejia and Wahba (2002) and Imbens and Rubin (2015), the estimator proposed by Hong (2010) does not use augmentation. It is based on the potentially misspecified main effects propensity score equation. Hong (2010) focused primarily on extending propensity score stratification to the case of multi-valued treatments. However, this author also argued that the "[marginal mean weighting through stratification (MMWS)] method usually provides a better approximation of nonlinear or nonadditive relationships between treatment assignment and pretreatment covariates" (Hong 2010, p. 523) and therefore is robust to incorrect specification of the functional form of the propensity score equation.

[5] Note that propensity score matching, and consequently the augmentation algorithm, do not involve the outcome variable. Therefore, algorithms like those proposed here are not considered "data mining". Balance testing can (and should) be conducted prior to hypothesis testing to maintain objectivity while searching for the specification that provides the best balance, regardless of whether this specification supports or rejects a given research hypothesis (Rubin 2001, 2007).

Hong's (2010) MMWS estimator combines weighting and propensity score stratification. First, the sample of observations is divided into a number of strata, based on the estimated and potentially misspecified propensity score. The strata are chosen such that the number of observations in each stratum is approximately equal. As a consequence of this stratification, within each stratum, the distribution of the covariates should be similar in both the treatment and control groups. Marginal mean weights (mmw) are computed based on the following equation:

$$mmw = \frac{n_s P\left(T = t\right)}{n_{t,s}}$$

where $P(T = t)$ is the probability of receiving treatment version t. In the case of binary treatments, only two versions exist, treatment and no treatment; however, the method can be extended to multiple ordered and unordered treatments. n_s is the total number of observations in stratum s, whereas $n_{t,s}$ is the number of observations subjected to treatment version t within the same stratum s. To estimate the causal effect, these stratum-specific weights are applied to all observations within the common support.

Hong (2010) supported this argument using a Monte Carlo simulation. More recently, Linden (2017a) compared Hong's (2010) version to standard propensity score stratification and found the former to be slightly more robust to misspecification than the latter; both methods outperform inverse probability weighting. Linden (2017a; see also Linden 2017b for a similar analysis) appears to both support Hong's (2010) methodological claim and to indicate that the need for augmentation is not as strong as claimed by the methodologists Dehejia and Wahba (2002) and Imbens and Rubin (2015). However, in a simulation with a more complex design, Linden et al. (2016) found that the bias in MMWS is often similar to the bias in inverse probability weighting, which can be substantial. No comparison to propensity score matching with or without augmentation has been conducted so far.

5 Monte Carlo experiment

We conduct a Monte Carlo simulation to investigate the performance of the modified DW algorithm and other estimators in the presence of misspecification of the propensity score. We use two different simulation structures that have been developed for similar purposes.

Simulation 1 follows a setup developed by Setoguchi et al. (2008) that has been modified only slightly for our purposes. The simulation involves a continuous outcome variable y and a binary treatment variable t (1 if treated, 0 if control) where p(t = 1) = ~0.5. Four covariates (x_1, x_2, x_3 and x_4) are correlated with both the treatment and the outcome, three covariates are correlated only with the treatment (x_5, x_6 and x_7), and another three covariates are correlated only with the outcome (x_8, x_9 and x_{10}). The covariates are generated such that two groups of some covariates (x_2 and x_6; x_4 and x_9) are highly correlated (0.9) with each other, but not with any of the other covariates. Other covariates are only moderately correlated (0.2) with each other (x_1 and x_5; x_3 and x_8) and also not correlated with the other covariates. The remaining correlations are set to 0. All of the covariates are generated as standard normal random variables, but they are dichotomized after introducing the correlations (x_1, x_3, x_5, x_6, x_8 and x_9). We implement a continuous outcome variable, following Lee et al. (2010), in which $y_i = 0.4t_i + \mathbf{x}'_i\alpha + \varepsilon_i$; however, the coefficient vector α is the same as in Setoguchi et al. (2008). The random error term ε is not correlated with any of the covariates, and $\varepsilon_i \sim N(0,1)$, leading to an R^2 of approximately 0.3.

The treatment indicator t (1 if treated, 0 if control) is generated from a binomial distribution with probability $p_i(t_i = 1) = (1 + \exp(-(f(\boldsymbol{x}'_i)\beta)))^{-1}$, with $(i = 1, ..., N)$, where the function $f(\boldsymbol{x}'_i)$ is specified such that the propensity score equation is characterized by increasing degrees of nonadditivity and nonlinearity from Scenarios A to G. For example, in Scenario A, the true propensity score equation contains only main effects, such that $p_i = (1 + \exp(-(\beta_0 + \beta_1 x_{\mathrm{i}1} + \beta_2 x_{\mathrm{i}2} + \beta_3 x_{\mathrm{i}3} + \beta_4 x_{\mathrm{i}4} + \beta_5 x_{\mathrm{i}5} + \beta_6 x_{\mathrm{i}6} + \beta_7 x_{\mathrm{i}7})))^{-1}$. In Scenario G, the propensity score equation features 10 two-way interaction terms and 3 quadratic terms, a situation Setoguchi et al. (2008) call moderate nonadditivity and nonlinearity (see the appendix for a detailed listing).

The simulation is conducted by generating a population of size N, where N takes the values of 200, 1000, and 2500. A logistic regression analysis is conducted to estimate the propensity score $\widehat{p_i}$ for each sample. Each Monte Carlo simulation consists of R = 1000 replicates of the process, from generating the population N to estimating the causal effect from the sample of N observations. We chose to estimate the average treatment effect on the treated, since this is most common in empirical applications.

The estimators we compare differ in several regards. They differ in terms of the specification of the propensity score equation (i. e., main effects specifications vs. augmentation algorithms), whether or not a treatment predictor is omitted and whether propensity score matching or (in one case) propensity score weighting is used. We first start with a main effects specification that does not omit a treatment predictor but is misspecified with regard to the functional form of the propensity score equation in all but Scenario A. This estimator is our benchmark because it shows us to what degree bias arises when only the functional form of the propensity score equation is misspecified. The second estimator uses a main effects specification for the propensity score equation but omits variable x_2, i. e., a variable associated both with the treatment and the outcome. The third estimator differs from the second one only in that it omits variable x_7, a variable that is associated only with the treatment and not with the outcome. By comparing the results obtained using estimators two and three with those obtained using estimator one, we can assess the size of the bias caused by misspecification. Causal effects are estimated by way of propensity score matching in all three cases, and we choose the SNNM without replacement with a caliper of 0.01. To estimate the causal effect, we use the Stata add-on “psmatch2” (Leuven and Sianesi 2014).

The fourth estimator, mean marginal weighting through stratification (MMWS), relies on weighting, rather than propensity score matching. To implement MMWS, we first use the Stata add-on “pstrata” (version 1.1.1; Linden 2016) to find the "optimal" number of strata, which corresponds to finding no significant differences in the mean propensity scores ($p = 0.05$) between the treatment and control groups within each stratum. The algorithm starts with 2 strata and tests for the equality of means. If significant differences remain, the sample is divided in 3 strata and the test is repeated, etc. The algorithm stops either when no significant differences exist or when one stratum contains no treatment or no control observations. The solution is then passed to the Stata add-on "mmws" (version 1.21; Linden 2014) to estimate the causal effect using the respective weights.

Only the fifth and sixth estimators use augmentation to find the preferred specification. Starting with the fifth estimator, the modified Dehejia and Wahba algorithm is implemented as described in Sect. 4. Dehejia and Wahba recommend including interactions and higher-order terms for unbalanced variables, although there is no guarantee that the resulting propensity score will balance the sample. Rather, analysists shouldtest for the equality of means and modify the equation if significant differences remain. Only in that case, the score is re-estimated, and a new balancing test is performed. This process is repeated until analysists find a specification that is sufficiently balanced. Here, we modify the algorithm to always augment the equation step by step, including first the squares of each variable and then all possible first-order interactions one by one. This procedure yields several different estimates. The number of these estimates depends on the number of covariates. From the various specifications we choose the version for which the highest standardized difference among the variables correlated with the treatment is the lowest of the consecutive estimates (i. e., we minimize the maximum bias from the consecutive propensity score estimates).

As implemented here, the modified DW algorithm systematically estimates several logistic regressions, where the interactions and squares are applied to all covariates, not only those that are deemed unbalanced. Because we restrict ourselves to first-order interactions and squares, the modified DW algorithm does not cover all possible specifications of the propensity score. Furthermore, once a square or an interaction term is included, it is no longer excluded from the estimation of the propensity score. Thus, the simulation results are conservative because some specifications that might reduce the balance even further are not considered. Additionally, we stress that none of the specifications coincide with the true one in either Simulation 1 or in Simulation 2 below. This procedure reflects a situation in which researchers remain agnostic about the true specification and follow what subject matter researchers would consider a "mindless" strategy; that is, the interactions and squares are included, regardless of any theoretical justification.

To implement the Imbens – Rubin (2015) algorithm, we use the Stata add-on "psestimate" (Version 1.5.3, Carril 2016). We use the default settings for the threshold values, as they are the ones recommended by Imbens and Rubin (2015). However, the Stata add-on "psestimate" unfortunately does not permit the selection of some covariates to be part of the propensity score equation, regardless of the strength of their correlation with the treatment.

Simulation 1 stops at Scenario G, a situation characterized by Setoguchi et al. (2008) as moderate nonadditivity and nonlinearity in the propensity score equation. In a second simulation (Simulation 2), we model a more extreme form of nonadditivity and nonlinearity. Furthermore, we also investigate the consequences of nonadditivity and nonlinearity in the outcome equation for the performance of the modified DW algorithm. To do this, the second simulation structure follows a setup developed by Kang and Schafer (2007) that has been slightly modified for our purposes. In contrast to Simulation 1, this setup includes fewer covariates[6] but allows for nonlinearities and nonadditivity in both the outcome and propensity score equations. In addition, the nonlinearity and nonadditivity in both equations is more complex than those in Simulation 1, explaining why we refer to it as strong nonadditivity and nonlinearity. Four covariates (z_1, z_2, z_3 and z_4) are generated following a standard normal distribution and subsequently transformed, such that

$$\begin{aligned}
x_{\mathrm{i}1} &= \exp\left(\frac{z_{\mathrm{i}1}}{2}\right)\\
x_{\mathrm{i}2} &= \frac{z_{\mathrm{i}2}}{1+\exp\left(z_{\mathrm{i}1}\right)}+10\\
x_{\mathrm{i}3} &= \left(\frac{z_{\mathrm{i}1}z_{\mathrm{i}3}}{25}+0.6\right)^3\\
x_{i4} &= (z_2+z_4+20)^2
\end{aligned}$$

This setup is used to compare four estimators. First, we analyze the performance of a propensity score matching estimator in which the analyst guesses the correct specification of the propensity score equation. Second, we analyze the performance of propensity score matching based on a misspecified main effects propensity score equation. Third, we examine whether marginal mean weighting based on the misspecified main effects propensity score equation eliminates bias associated with potential misspecification. Fourth and fifth, we investigate the performance of both the Dehejia and Wahba and the Imbens and Rubin augmentation algorithms.

We distinguish two scenarios. In Scenario 1, the outcome equation is based on z_1 to z_4 such that nonlinearity and nonadditivity is only present in the propensity score equation; $y_i = 2.1 + 0.4t_i + 2.74z_{\mathrm{i}1} + 1.37z_{\mathrm{i}2} + 1.37z_{\mathrm{i}3} + 1.37z_{\mathrm{i}4} + \varepsilon_{\mathrm{i}}$. In Scenario 2, the variables x_1 to x_4 are substituted for the z-variables in the outcome equation, thus keeping the coefficients but causing the outcome equation to also be characterized by strong nonlinearity and nonadditivity.[7]

Simulation 2 is also conducted by generating a population of size N, where N takes the values 200, 1000, and 2500, a logistic regression analysis is conducted to estimate the propensity score $\widehat{p}_i(t_i = 1)$for each sample. Moreover, the average treatment effect on the treated is estimated and averaged over the 1000 replications of the simulation. The implementations of PSM, MMMS and the two augmentation algorithms follow those in Simulation 1.

[6] However, all four covariates are correlated with both the treatment and the outcome. Given that covariates correlated to either the treatment or the outcome are not expected to contribute to eliminating bias, the simulations are more similar than it seems at first glance.

[7] Note that, in contrast to Kang and Schafer (2007), we do not perform additional regression adjustments. Instead, we merely estimate the causal effect via the (weighted) difference in mean outcomes of the treatment and control groups. Because we do not use an outcome model to estimate the causal effect, only the propensity score equation can be misspecified. However, the data-generating process can be linear and additive or non-linear and non-additive for both the treatment and the outcome.

6 Results

In this section, we present the results from the two simulations, both of which are conducted with 200, 1000 and 2500 observations. In Simulation 1, Scenarios A to G differ with regard to the extent of nonlinearity and nonadditivity in the true propensity equation, but the outcome equation is always linear and additive. We compare the effects of the different estimation strategies on the bias of the causal effect estimate, specifically the average treatment effect on the treated. We first discuss the results presented in Table 1, which are based on simulations with 1000 observations, and then compare these to the results for smaller and larger samples that are presented in Table 2 and 3, respectively.

We start with results for propensity score matching based on a main effects logistic regression that includes all seven covariates. In Scenario A, this specification coincides with the true specification, and we find only a small absolute bias. In Scenarios B–G, the true specification is characterized by increasing nonlinearity and nonadditivity, and therefore the main effects logistic regression becomes increasingly misspecified. However, contrary to expectations, the bias does not increase with the degree of misspecification. This result contrasts with that obtained for propensity score weighting performed using the same simulation setup (Setoguchi et al. 2008). Therefore, it appears that misspecification bias is less of a problem for propensity score matching than other propensity score methods, especially weighting.

Compared to misspecification bias, omitting an important covariate leads to a substantial bias of around 10%. In line with statistical theory and previous research, we find this omitted variable bias arises only when the variable is associated with both the treatment and the outcome, but not when the omitted variable is associated only with the treatment.

Even if misspecification bias is generally small, there are some differences between estimators, both between those that do and do not omit treatment predictors and between those that are based on main effects propensity score equations and those that rely on augmentation algorithms. In the special case were the main-effect logit is correctly specified, we find that matching on the propensity score leads to a small absolute bias of 1.6%. The bias is similar for marginal mean weighting (1.9%). However, the modified DW augmentation algorithm has a bias of only 1%. Thus, re-estimating the propensity score after including additional nonlinear terms further reduces the already small bias compared to the correctly specified equation. This result occurs because several attempts at propensity score matching are made per replication, and the algorithm chooses the one with the lowest bias. A similar reduction in bias is achieved by the IR augmentation algorithm. In Scenarios B to G, the misspecification bias remains small, although the degree of nonlinearity and nonadditivity is greater. For most of the tested specifications, MMWS does not reduce the bias and sometimes increases it; the absolute bias reaches 2.5% in Scenario E. In contrast, the two augmentation algorithms mostly continue to reduce bias.

Over all of the tested scenarios, the standard errors are similar for all estimators. In contrast, the mean standardized differences diverge between estimation strategies. First, even when propensity score matching is applied with misspecified main effects, the mean standardized difference is between 3 and 4% and thus well within the

Table 1 Results for Simulation 1, $N = 1000$

	Main effects PS equation												Augmented PS equation					
	PSM (full set)			*PSM (x2 omitted)*			*PSM (x7 omitted)*			*MMWS*			*mDW algorithm*			*IR algorithm*		
Scenario	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%
A	*0.016*	0.091	3.15	*0.090*	0.091	4.31	*0.017*	0.084	11.90	*0.019*	0.084	2.30	*0.009*	0.094	1.75	*0.008*	0.094	3.20
B	*0.010*	0.094	3.13	*0.081*	0.094	3.97	*0.016*	0.086	11.63	*0.018*	0.087	2.29	*0.004*	0.092	1.74	*0.004*	0.094	3.21
C	*0.013*	0.085	2.86	*0.068*	0.089	3.63	*0.013*	0.087	8.70	*0.002*	0.079	1.77	*0.010*	0.090	1.75	*0.006*	0.095	3.55
D	*0.003*	0.094	3.36	*0.098*	0.095	4.53	*0.006*	0.087	12.02	*0.020*	0.090	2.71	*0.004*	0.091	1.84	*0.005*	0.097	3.38
E	*0.009*	0.094	3.28	*0.069*	0.093	4.07	*0.007*	0.086	11.84	*0.025*	0.090	2.76	*0.002*	0.093	1.79	*0.003*	0.097	3.45
F	*0.007*	0.093	3.27	*0.133*	0.093	4.86	*0.012*	0.088	9.45	*0.005*	0.086	2.44	*0.012*	0.092	1.85	*0.005*	0.101	3.47
G	*0.001*	0.091	3.11	*0.109*	0.090	4.25	*0.010*	0.089	6.99	*0.006*	0.081	1.88	*0.002*	0.090	1.83	*0.009*	0.101	3.74

PSM Propensity Score Matching, *MMWS* Marginal Mean Weighting through Stratification, *mDW* modified Dehejia-Wahba algorithm, *IR algorithm* Imbens-Rubin algorithm; *ab%* absolute bias in percent; se: standard error, *msd%* absolute mean standardized difference over all 7 covariates, averaged over the 1000 replicates of the Monte Carlo simulation

Table 2 Results for Simulation 1, $N = 200$

	Main effects PS equation												Augmented PS equation					
	PSM (full set)			*PSM (x2 omitted)*			*PSM (x7 omitted)*			*MMWS*			*mDW algorithm*			*IR algorithm*		
Scenario	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%
A	*0.010*	0.252	9.73	*0.113*	0.238	10.58	*0.001*	0.218	16.68	*0.056*	0.194	6.48	*0.008*	0.242	5.21	*0.014*	0.249	10.53
B	*0.012*	0.236	9.64	*0.107*	0.241	10.34	*0.004*	0.216	16.56	*0.063*	0.198	6.40	*0.017*	0.240	5.35	*0.010*	0.245	10.66
C	*0.024*	0.221	8.61	*0.094*	0.218	9.36	*0.013*	0.213	13.31	*0.057*	0.180	5.21	*0.017*	0.232	4.95	*0.019*	0.255	11.88
D	*0.013*	0.245	10.13	*0.107*	0.254	11.28	*0.005*	0.238	17.15	*0.055*	0.217	7.22	*0.014*	0.241	5.58	*0.009*	0.266	11.28
E	*0.026*	0.254	10.04	*0.08*	0.256	10.98	*0.005*	0.231	16.89	*0.055*	0.201	7.20	*0.016*	0.244	5.60	*0.028*	0.253	11.20
F	*0.017*	0.252	10.03	*0.126*	0.240	10.95	*0.027*	0.235	14.51	*0.040*	0.203	6.80	*0.016*	0.241	5.63	*0.038*	0.258	11.28
G	*0.015*	0.233	9.26	*0.11*	0.235	9.96	*0.007*	0.232	11.98	*0.018*	0.191	5.52	*0.024*	0.243	5.21	*0.004*	0.265	11.83

PSM Propensity Score Matching, *MMWS* Marginal Mean Weighting through Stratification, *mDW* modified Dehejia-Wahba algorithm, *IR algorithm* Imbens-Rubin algorithm; *ab%* absolute bias in percent; se: standard error, *msd%* absolute mean standardized difference over all 7 covariates, averaged over the 1000 replicates of the Monte Carlo simulation

Table 3 Results for Simulation 1, $N = 2500$

	Main effects PS equation												Augmented PS equation					
	PSM (full set)			*PSM (x2 omitted)*			*PSM (x7 omitted)*			*MMWS*			*mDW algorithm*			*IR algorithm*		
Scenario	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%
A	*0.005*	0.059	1.92	*0.101*	0.058	3.31	*0.012*	0.052	11.06	*0.010*	0.053	1.35	*0.006*	0.055	1.21	*0.003*	0.059	1.94
B	*0.004*	0.058	1.95	*0.088*	0.060	2.96	*0.012*	0.053	10.86	*0.014*	0.054	1.35	*0.002*	0.057	1.14	*0.000*	0.059	2.00
C	*0.005*	0.053	1.90	*0.078*	0.054	2.73	*0.013*	0.051	7.96	*0.003*	0.046	1.03	*0.002*	0.053	1.29	*0.002*	0.060	2.12
D	*0.002*	0.058	2.08	*0.1*	0.062	3.40	*0.015*	0.056	11.09	*0.007*	0.057	1.60	*0.006*	0.061	1.22	*0.006*	0.060	2.08
E	*0.008*	0.058	2.02	*0.074*	0.060	2.95	*0.01*	0.054	10.90	*0.020*	0.056	1.66	*0.005*	0.058	1.17	*0.002*	0.060	2.07
F	*0.010*	0.060	2.09	*0.127*	0.059	3.90	*0.01*	0.058	8.45	*0.007*	0.054	1.43	*0.013*	0.060	1.28	*0.013*	0.061	2.11
G	*0.021*	0.058	2.10	*0.122*	0.058	3.43	*0.013*	0.055	6.20	*0.006*	0.050	1.21	*0.001*	0.058	1.32	*0.006*	0.063	2.20

PSM Propensity Score Matching, *MMWS* Marginal Mean Weighting through Stratification, *mDW* modified Dehejia-Wahba algorithm, *IR algorithm* Imbens-Rubin algorithm; *ab%* absolute bias in percent; se: standard error, *msd%* absolute mean standardized difference over all 7 covariates, averaged over the 1000 replicates of the Monte Carlo simulation

range considered to indicate sufficient balance in a matching analysis (Caliendo and Kopeinig 2008: 48). Because these low values are accompanied by only small misspecification bias, there is no contradiction. Also, bias reduction using the DW algorithm is accompanied by a simultaneous reduction in the mean standardized differences, which are reduced to below 2%. However, although marginal mean weighting also reduces the mean standardized bias in the covariates compared to the misspecified propensity score matching, this change is not always accompanied by reductions in bias. Notably, in the case where a treatment predictor is omitted, the standardized difference remains low with a maximum of 5%, even in the case of omitted variable bias. On the other hand, the standardized difference is quite high, reaching almost 12%, even if omitting a variable does not induce bias.

Comparing the results for 1000 observations to those for the very small sample of 200 observations (Table 2) shows that the standard errors, as well as the mean standardized bias in the covariates, are higher, as expected. The omitted variable bias does not seem to depend strongly on sample size. Although the misspecification bias is larger in the smaller sample, it is still rather small. Also in the small samples, the bias does not change substantially, even when the degree of misspecification increases. Again, the DW and IR algorithms slightly improve upon both the correct main effects specification and the misspecified main effects logit regressions. However, when the degree of misspecification increases, the performance of both augmentation algorithms decreases. In contrast to the medium-sized sample, in smaller samples, the marginal mean weighting does not reduce the bias but instead increases it; the absolute bias often reaches 5 to 6%. Both the standard errors and mean standardized differences are generally higher in the smaller sample than in the medium-sized sample, as is to be expected. If the sample size increases to 2500 observations (Table 3), we find that the misspecification bias mostly disappears, regardless of the degree of misspecification and whether augmentation is used or not. The standard errors and mean standardized differences are very small.

In Simulation 2, where both the true propensity score and the outcome equation are allowed to contain high levels of nonlinearity and nonadditivity, the results are quite different from those of Simulation 1. Again, we start by discussing the results for the case with 1000 observations shown in Table 4. In Scenario I, the outcome equation is strictly linear and additive, whereas the true propensity score equation is not. In such cases, propensity score matching that uses the correct specification to estimate the propensity score is only slightly biased (by 1%), and this degree of bias is similar to that obtained with the correct main effects specification in Simulation 1. In contrast to Simulation 1, however, we find that the misspecified main effects estimation of the propensity score equation leads to a substantially larger absolute bias of 13%. This bias is virtually unchanged by marginal mean weighting. This result contradicts the claims of both Hong (2010) and Linden (2017a, 2017b) that stratification alleviates or even eliminates misspecification bias, but it is in line with results obtained in Linden et al. (2016). In contrast, the DW algorithm reduces the absolute bias by about half (to 7%), whereas the IR algorithm reduces the bias to 9%. It seems that, in the case of higher levels of nonlinearity and nonadditivity, misspecification of the propensity score equation actually does lead to considerable

Table 4 Results for Simulation 2, $N = 1000$

PSM Correct (nonlinear) specification			PSM Main effects specification (misspecified)			MMWS			mDW algorithm			IR algorithm		
ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%
Scenario I: Linear and additive outcome equation														
0.009	0.041	4.03	*0.131*	0.045	4.51	*0.145*	0.040	3.59	*0.071*	0.043	1.68	*0.094*	0.044	4.24
Scenario II: Non-linear and non-additive outcome equation														
0.107	0.996	4.03	*0.767*	1.053	4.51	*0.158*	0.922	3.59	*0.026*	0.984	1.68	*0.821*	1.059	4.24

PSM Propensity Score Matching, *MMWS* Marginal Mean Weighting through Stratification, *mDW* modified Dehejia-Wahba algorithm, *IR algorithm* Imbens-Rubin algorithm; *ab%* absolute bias in percent; se: standard error, *msd%* absolute mean standardized difference over all 4 covariates, averaged over the 1000 replicates of the Monte Carlo simulation

bias. Both of the augmentation algorithms, but not MMWS, are able to reduce the bias substantially, but they are not able to eliminate it altogether.

Turning to Scenario II, where the true outcome equation is strongly nonlinear and nonadditive, propensity score matching is biased by approximately 11%, even when it is based on a correctly specified model. This result is in line with that of Kang and Schafer (2007) for weighting and double robust estimators. If propensity score estimation is additionally based on a misspecified main effects logit, the estimate is severely biased, and the absolute bias becomes 77%. Surprisingly, MMWS is as biased in Scenario II as in Scenario I; i. e., the absolute bias is 16%. Whereas both augmentation algorithms perform similarly in Scenario I, the IR algorithm does not reduce the bias; instead, it increases it slightly (to 82%) in Scenario II. In contrast, the DW algorithm leads to a strong decrease in the bias and leads to an absolute bias of merely 2.6%. The suboptimal performance of the IR algorithm is, however, most likely attributable to the specific implementation in stata. The user written program does not allow for retaining specific variables in the propensity score equation, irrespective of their correlation with the treatment. This restriction might have lead the algorithm to eliminate variables from the propensity score equation that have only small correlation to the treatment, but if the same variables are highly correlated to the outcome, their exclusion might still be problematic.

Similar to Simulation 1, the relationship between the degree of bias and the mean standardized difference is not unambiguous. For both the unbiased and the (moderately and severely) biased estimators, the values of the mean standardized bias are similar and well within the accepted range of 2–5%. In contrast, the mean standardized bias is in fact lowest for the DW algorithm, which is the least biased estimator. We interpret this result as evidence for the point made in Ho et al. (2007), who argued that even small standardized differences should be avoided for unbiased estimation.

The observed pattern does not change when the results from Simulation 2 obtained with populations of 1000 observations are compared with those obtained using much smaller or larger samples (Tables 5 and 6). Regardless of sample size, the DW algorithm performs best among the selected estimators, even if the outcome equation is nonlinear and nonadditive.

7 Conclusions

In this paper, we tested the performance of augmentation as proposed by Dehejia and Wahba (2002) to reduce bias due to misspecification. We proposed to slightly modify the original algorithm to alleviate the problems pointed out by its critics. The original algorithm involves systematically introducing interactions and higher-order terms while checking whether balance is improved. There is, however, no guarantee that balance will improve, even if the introduced interactions are theoretically sound. Therefore, the modified Dehejia and Wahba–algorithm proposes to test several specifications and select the one that produces the best balanced sample. The step-wise cumulative augmentation of the logistic regression by introducing interactions and higher-order terms is a simple and convenient strategy to reduce bias.

Table 5 Results for Simulation 2, $N = 200$

PSM Correct (nonlinear) specification			PSM Main effects specification (misspecified)			MMWS			mDW algorithm			IR algorithm		
ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%
Scenario I: Linear and additive outcome equation														
0.007	0.111	10.69	*0.123*	0.109	9.35	*0.181*	0.087	8.09	*0.074*	0.103	4.29	*0.105*	0.112	10.10
Scenario II: Non-linear and non-additive outcome equation														
0.041	2.675	10.69	*0.543*	2.550	9.35	*0.127*	1.999	8.09	*0.058*	2.386	4.29	*0.631*	2.657	10.10

PSM Propensity Score Matching, *MMWS* Marginal Mean Weighting through Stratification, *mDW* modified Dehejia-Wahba algorithm, *IR algorithm* Imbens-Rubin algorithm; *ab%* absolute bias in percent; se: standard error, *msd%* absolute mean standardized difference over all 4 covariates, averaged over the 1000 replicates of the Monte Carlo simulation

Table 6 Results for Simulation 2, $N = 2500$

PSM Correct (nonlinear) specification			PSM Main effects specification (misspecified)			MMWS			mDW algorithm			IR algorithm		
ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%	ab%	se	msd%
Scenario I: Linear and additive outcome equation														
0.013	0.026	2.60	*0.133*	0.027	4.04	*0.137*	0.024	2.66	*0.072*	0.027	1.26	*0.089*	0.028	3.24
Scenario II: Non-linear and non-additive outcome equation														
0.040	0.641	2.60	*0.829*	0.631	4.04	*0.208*	0.559	2.66	*0.011*	0.609	1.26	*0.703*	0.684	3.24

PSM Propensity Score Matching, *MMWS* Marginal Mean Weighting through Stratification, *mDW* modified Dehejia-Wahba algorithm, *IR algorithm* Imbens-Rubin algorithm; *ab%* absolute bias in percent; se: standard error, *msd%* absolute mean standardized difference over all 4 covariates , averaged over the 1000 replicates of the Monte Carlo simulation

Based on the two preceding Monte Carlo simulations, we can draw several conclusions. From the first simulation, we learned that compared to omitting a relevant variable, misspecifying the functional form in the model that estimates the propensity score induces only small bias. Even in cases where the true equation is characterized by moderate nonlinearity and nonadditivity, misspecification is often negligible, as long as the outcome equation is linear and additive. However, the modified Dehejia and Wahba (2002) algorithm still helps to further reduce the misspecification bias. An alternative augmentation algorithm suggested by Imbens and Rubin (2015) performed similarly, whereas a variant of propensity score stratification proposed by Hong (2010) performed slightly worse, especially in small samples.

From the second simulation we learned that in case of high nonlinearity and nonadditivity, misspecification bias can be quite severe. If the functional form of the propensity score equation is misspecified, but the outcome equation is linear and additive, misspecification bias is quite substantial. If in addition, the outcome equation is also highly nonlinear and nonadditive, bias becomes severe. In both cases, however, we found the modified Dehejia and Wahba–algorithm to reduce bias markedly, whereas the Imbens and Rubin (2015) algorithm only reduced bias if the outcome equation was linear and additive.

In all, misspecification will not always induce bias, especially when the true equation is only moderately nonlinear and nonadditive. Because the true specification of the propensity score equation is unknown, it seems prudent, however, to always take measures to avoid misspecification bias. We found the modified Dehejia and Wahba–algorithm to do a good job in reducing bias, especially compared to propensity score stratification and often also compared to the recent proposal by Imbens and Rubin (2015).

Acknowledgements For helpful comments to earlier versions I am grateful to Jörg Drechsler, Susanne Rässler and her team, as well as an anonymous reviewer. Any remaining mistakes and limitations of the analysis are my own.

Appendix

Simulation 1 True propensity score models

Scenario A (a model with additivity and linearity):

$$p\left(t=1|\boldsymbol{x}\right)=\left(1+\exp\left\{-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7\right)\right\}\right)^{-1}$$

Scenario B (a model with mild non-linearity):

$$\begin{aligned}&p\left(t=1|\boldsymbol{x}\right)\\&=\left(1+\exp\left\{-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_2x_2x_2\right)\right\}\right)^{-1}\end{aligned}$$

Scenario C (a model with moderate non-linearity):

$$p(t = 1|\boldsymbol{x}) = (1 + \exp$$
$$\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_2 x_2 x_2 + \beta_4 x_4 x_4 + \beta_7 x_7 x_7)\})^{-1}$$

Scenario D (a model with mild non-additivity):

$$p(t = 1|\boldsymbol{x}) = (1 + \exp\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$
$$+\beta_1 \times 0.5 \times x_1 x_3 + \beta_2 \times 0.7 \times x_2 x_4 + \beta_4 \times 0.5 \times x_4 x_5 + \beta_5 \times 0.5 \times x_5 x_6)\})^{-1}$$

Scenario E (a model with mild non-additivity and non-linearity):

$$p(t = 1)\boldsymbol{x}) = (1 + \exp\{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$
$$+ \beta_7 x_7 + \beta_2 x_2 x_2 + \beta_1 \times 0.5 \times x_1 x_3 + \beta_2 \times 0.7 \times x_2 x_4 + \beta_4 \times 0.5 \times x_4 x_5 +$$
$$\beta_5 \times 0.5 \times x_5 x_6)\})^{-1}$$

Scenario F (a model with moderate non-additivity):

$$p\,(t = 1|\,\boldsymbol{x})$$
$$= \left(1 + \exp \left\{ - \begin{pmatrix} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 \\ +\beta_1 \times 0.5 \times x_1 x_3 + \beta_2 \times 0.7 \times x_2 x_4 + \beta_3 \times 0.5 \times x_3 x_5 \\ +\beta_4 \times 0.7 \times x_4 x_6 + \beta_5 \times 0.5 \times x_5 x_7 + \beta_1 \times 0.5 \times x_1 x_6 \\ +\beta_2 \times 0.7 \times x_2 x_3 + \beta_3 \times 0.5 \times x_3 x_4 + \beta_4 \times 0.5 \times x_4 x_5 \\ +\beta_5 \times 0.5 \times x_5 x_6 \end{pmatrix} \right\} \right)^{-1}$$

Scenario G (a model with moderate non-additivity and non-linearity):

$$p\,(t = 1|\,\boldsymbol{x})$$
$$= \left(1 + \exp \left\{ - \begin{pmatrix} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 \\ +\beta_2 x_2 x_2 + \beta_4 x_4 x_4 + \beta_7 x_7 x_7 + \beta_1 \times 0.5 \times x_1 x_3 \\ +\beta_2 \times 0.7 \times x_2 x_4 + \beta_3 \times 0.5 \times x_3 x_5 + \beta_4 \times 0.7 \times x_4 x_6 \\ +\beta_5 \times 0.5 \times x_5 x_7 + \beta_1 \times 0.5 \times x_1 x_6 + \beta_2 \times 0.7 \times x_2 x_3 \\ +\beta_3 \times 0.5 \times x_3 x_4 + \beta_4 \times 0.5 \times x_4 x_5 + \beta_5 \times 0.5 \times x_5 x_6 \end{pmatrix} \right\} \right)^{-1}$$

Outcome model Scenario A–G:

$$\mathrm{y}(x) = (1 + \exp\{-(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_8 + \alpha_6 x_9 + \alpha_7 x_{10} + \gamma_1 \mathrm{t})\})^{-1}$$

Table 7 Coefficients of propensity score and outcome model

$\beta_0 = 0$
$\beta_1 = 0.8$
$\beta_2 = -0.25$
$\beta_3 = 0.6$
$\beta_4 = -0.4$
$\beta_5 = -0.8$
$\beta_6 = -0.5$
$\beta_7 = 0.7$
$\alpha_0 = -3.85$
$\alpha_1 = 0.3$
$\alpha_2 = -0.36$
$\alpha_3 = -0.73$
$\alpha_4 = -0.2$
$\alpha_5 = 0.71$
$\alpha_6 = -0.19$
$\alpha_7 = 0.26$
$\gamma_1 = -0.4$

References

Austin PC (2009) Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biom J 51:171–184. https://doi.org/10.1002/bimj.200810488

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T (2006) Variable selection for propensity score models. Am J Epidemiol 163:1149–1156. https://doi.org/10.1093/aje/kwj149

Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. J Econ Surv 22:31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Carril A (2016) PSESTIMATE: stata module to estimate the propensity score proposed by Imbens and Rubin. Statistical software components S458179. Boston College, Department of Economics (https://ideas.repec.org/c/boc/bocode/s458179.html, accessed: August 2017)

Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 84:151–161. https://doi.org/10.1162/003465302317331982

Diaz JJ, Handa S (2006) An assessment of propensity score matching as a nonexperimental impact estimator: evidence from Mexico's PROGRESA program. J Hum Resour 41:319–345. https://doi.org/10.3368/jhr.XLI.2.319

Harding DJ (2003) Counterfactual models of neighborhood effects: the effect of neighborhood poverty on dropping out and teenage pregnancy. Am J Sociol 109:676–719. https://doi.org/10.1086/379217

Heckman JJ, Ichimura H, Todd P (1998) Matching as an econometric evaluation estimator. Rev Econ Stud 65:261–294. https://doi.org/10.1111/1467-937X.00044

Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 15:199–236. https://doi.org/10.1093/pan/mpl013

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81:945–960. https://doi.org/10.2307/2289064

Hong G (2010) Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. J Educ Behav Stat 35:499–531. https://doi.org/10.3102/1076998609359785

Huber M, Lechner M, Steinmayr A (2012) Radius matching on the propensity score with bias adjustment: finite sample behaviour, tuning parameters and software implementation. Empir Econ 49:1–31. https://doi.org/10.1007/s00181-014-0847-1

Iacus SM, King G, Porro G (2012) Causal inference without balance checking: coarsened exact matching. Polit Anal 20:1–24. https://doi.org/10.1093/pan/mpr013

Imai K, van Dyk DA (2004) Causal inference with general treatment regimes. Generalizing the propensity score. J Am Stat Assoc 99:854–866. https://doi.org/10.1198/016214504000001187

Imbens G (2015) Matching methods in practice. J Hum Resour 50:373–419. https://doi.org/10.3368/jhr.50.2.373

Imbens G, Rubin D (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, New York

Jackson M, Cox DR (2013) The principles of experimental design and their application in sociology. Annu Rev Sociol 39:27–49. https://doi.org/10.1146/annurev-soc-071811-145443

Kang JDY, Schafer JL (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci 22:523–539. https://doi.org/10.1214/07-STS227

Lee WS (2013) Propensity score matching and variations on the balancing test. Empir Econ 44:47–80. https://doi.org/10.1007/s00181-011-0481-0

Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. Stat Med 29:337–346. https://doi.org/10.1002/sim.3782

Leuven E, Sianesi B (2014) PSMATCH2: stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical software components S432001. Boston College, Boston (http://ideas.repec.org/c/boc/bocode/s432001.html, accessed March 2014)

Linden A (2014) MMWS: stata module to perform marginal mean weighting through stratification. Statistical software components S457886. Boston College, Boston (https://ideas.repec.org/c/boc/bocode/s457886.html, accessed August 2017)

Linden A (2016) PSTRATA: stata module for optimal propensity score stratification. Statistical software components S458232. Boston College, Boston (https://ideas.repec.org/c/boc/bocode/s458232.html, accessed August 2017)

Linden A (2017a) A comparison of approaches for stratifying on the propensity score to reduce bias. J Eval Clin Pract 23:690–696. https://doi.org/10.1111/jep.12701

Linden A (2017b) Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting. J Eval Clin Pract 23:697–702. https://doi.org/10.1111/jep.12714

Linden A, Uysal SD, Ryan A, Adams JL (2016) Estimating causal effects for multivalued treatments: a comparison of approaches. Stat Med 35:534–552. https://doi.org/10.1002/sim.6768

Millimet DL, Tchernis R (2009) On the specification of propensity scores, with applications to the analysis of trade policies. J Bus Econ Stat 27:397–415. https://doi.org/10.1198/jbes.2009.06045

Rosenbaum PR, Rubin DB (1983a) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J R Stat Soc Series B Stat Methodol 45:212–218

Rosenbaum PR, Rubin DB (1983b) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 79:516–524. https://doi.org/10.2307/2288398

Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39:33–38. https://doi.org/10.1080/00031305.1985.10479383

Rubin DB (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. J Educ Psychol 66:688–701. https://doi.org/10.1037/h0037350

Rubin DB (1986) Which ifs have causal answers. Comment to Holland 1986. J Am Stat Assoc 81:961–962. https://doi.org/10.1080/01621459.1986.10478355

Rubin DB (2001) Using propensity score to help design observational studies: application to tabacco litigation. Health Serv Outcomes Res Methodol 2:169–188. https://doi.org/10.1023/A:1020363010465

Rubin DB (2004) On principles for modeling propensity scores in medical research. Pharmacoepidemiol Drug Saf 13:855–857. https://doi.org/10.1002/pds.968

Rubin DB (2007) The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 26:20–36. https://doi.org/10.1002/sim.2739

Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF (2008) Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf 17:546–555. https://doi.org/10.1002/pds.1555

Smith JA, Todd PE (2005a) Does matching overcome LaLondes's critique of nonexperimental estimators? J Econom 125:305–353. https://doi.org/10.1016/j.jeconom.2004.04.011

Smith JA, Todd PE (2005b) Rejoinder. J Econom 125:365–375

Smith JA, Zhang Y (2007) The variety of balancing tests. Association for Public Administration and Management, Washington D.C. (http://scholar.googleusercontent.com/scholar?q=cache:2dqWg7FZrSEJ:scholar.google.com, accessed: January 2009)

Stuart EA (2010) Matching methods for causal inference: a review and a look forward. Stat Sci 25:1–21. https://doi.org/10.1214/09-STS313

Zhao Z (2008) Sensitivity of propensity score methods to the specifications. Econ Lett 98:309–319. https://doi.org/10.1016/j.econlet.2007.05.010