



UNIVERSITY OF WARSAW

Faculty of Economic Sciences

WORKING PAPERS

No. 10/2013 (95)

PAWEŁ STRAWIŃSKI

**CONTROLLING FOR OVERLAP
IN MATCHING**

WARSAW 2013



UNIVERSITY OF WARSAW
Faculty of Economic Sciences

Controlling for overlap in matching

PAWEŁ STRAWIŃSKI

Faculty of Economic Sciences,
University of Warsaw
e-mail: pstrawinski@wne.uw.edu.pl

Abstract

The overlap problem is crucial in propensity score matching. Recently, Crump et al. (2009) showed that trimming provides a simple and robust solution to the overlap problem. In this study, we use a simulation approach to show that trimming is inferior to the caliper mechanism. We show that in most cases, both techniques provide unbiased estimates, but trimming is less efficient.

Keywords:

average treatment effect, overlap, propensity score, caliper, trimming.

JEL:

C14, C15

Working Papers contain preliminary research results.

Please consider this when citing the paper.

Please contact the authors to give comments or to obtain revised version.
Any mistakes and the views expressed herein are solely those of the authors.

Introduction

Quasi-experimental methods and estimation of average treatment effect on the treated (ATT) have gained popularity among empirical researchers. Comparisons of treated with non-treated objects usually use propensity scores, which are defined as the probability of being treated. Propensity score matching (PSM) plays such a fundamental role, since it allows for the use of one-dimensional non-parametric regression techniques, even with many confounding variables (Frölich, 2004). However, the observable characteristics of the treated group may differ substantially from those of the non-treated group. Hence, the estimation often faces a problem of lack of overlap; that is, the distributions of propensity scores in treated and non-treated groups have different supports, and the overlap may be narrow.

Limited overlap can result in poor finite sample properties of estimators for the ATT. Therefore, some studies have employed trimming methods or imposed a caliper on the propensity score to ensure the existence of greater overlap. Nevertheless, the implementation of these methods is not rooted in theory and is typically done on an ad-hoc basis or by a 'rule of thumb'. The caliper mechanism has been proposed to prevent poor matches. It imposes a limit on the maximum dissimilarity value between matched objects. Those pairs whose differences in propensity scores exceed pre-set caliper values are excluded from the analysis. Consequently, only those pairs with characteristics close to each other are used in the ATT estimation. Hence, the overlap is improved.

Recently, Crump et al. (2009) proposed a systematic approach that addresses the problem of lack of overlap. They showed that for a wide range of distributions, a simple rule of thumb to discard all units with estimated propensity scores outside the interval $[0.1, 0.9]$ provides a good approximation to the optimal rule. On the other hand, Busso et al. (2009) showed that trimming methods are not capable of correcting for the problem of lack of common support.

Both approaches that help to maintain overlap cause inefficiency of the ATT estimator, as they reduce the bias of the estimand at the cost of increased variance. Overlap is an important problem when there are small numbers of observations in the treated and non-treated samples. In this research, we compare two competing methods of controlling for overlap (i.e. the trimming method proposed by Crump et al. (2009) and the caliper method). We concentrate on the small-sample properties of both methods, address the question of efficiency, and elucidate the issue. The results of our simulation-based study show that trimming the support using the method of Crump et al. (2009) is inferior to the standard caliper mechanism. In small samples, trimming performed better than the caliper method in only three simulations, which were characterized by non-linear outcomes and χ^2 error distributions. In larger samples, when the caliper is tight (i.e. the trimming size is considerable), the caliper method evidently provides more-precise estimates.

The number of studies dealing with the finite sample properties of matching estimators is limited. Note the works by Austin (2009); Busso et al. (2009); and Frölich (2004). Frölich (2004) examined the properties of various propensity-score matching estimators and showed that one-to-one matching is outperformed by ridge matching. However, the mean squared error (MSE) of the ridge matching procedure is lower than that of one-to-one matching only if the optimal bandwidth is known. Usually, the optimal bandwidth value is not known a priori and has to be estimated. Austin (2009) also compared several matching techniques in a Monte Carlo study, concentrating mainly on one-to-one matching. All examined estimators resulted in similar numbers of matched pairs and a similar balance of variables between treated and untreated samples. Moreover, matching according to the propensity score with caliper size of ≤ 0.03 tends to result in estimates with negligible relative bias. Similarly, Busso et al. (2009), Huber et al. (2010) emphasise the role of trimming to account for common support. Controlling for common support condition effectively improves matching performance regardless of estimator used. Recently, Strawiński (2011) showed that in small samples, the bias due to inexact matching is relatively small in comparison with that due to incomplete matching. The empirical results suggest small bias even if the treatment is not constant. On the other hand, the bias

due to incomplete matching turned out to be substantial (i.e. up to 15% for a very conservative caliper size and about 10% for the most popular caliper size of 0.005).

The remainder of this article is organised as follows. In section 2, we describe the matching framework, introduce the notation used for matching estimators, and describe the trimming and caliper mechanisms in detail. In section 3, we describe our Monte Carlo experiment for different distributions of the propensity score and the outcome equations. In the fourth section, we present our main results. The last section contains the summary and conclusion.

2. Matching framework

Our framework follows the seminal paper of Rosenbaum and Rubin (1983). Let us assume we have a random sample of N observations from a large population. For each observation i , let T_i be the indicator of treatment. $T_i = 1$ signifies that the treatment of interest was received, while $T_i = 0$ signifies the so-called control treatment. Let Y_{1i} be the outcome when individual i receives a treatment and Y_{0i} signify that when he or she does not. The treatment effect for an individual i can be written as:

$$Y_{1i} - Y_{0i} \quad (1)$$

The fundamental evaluation problem arises because for each individual i either Y_{0i} or Y_{1i} is observed. Therefore estimating individual treatment effects is not possible. The population averaged treatment effects are used instead. The average treatment effect on the treated (ATT) is defined as

$$ATT = E[Y_1 | P = 1] - E[Y_0 | P = 1] \quad (2)$$

where Y_1 and Y_0 are the average outcomes over the treated and untreated populations. Simply, $Y_1 = \sum Y_{1i}$ and $Y_0 = \sum Y_{0i}$.

A typical matching estimator has the form (Smith and Todd, 2005):

$$\frac{1}{N} \sum_{i=1}^N [Y_{1i} - E(Y_{0i} | T_i = 1)] \quad (3)$$

where the right term of equation (2) is an estimator of the counterfactual state. The PSM methods are usually implemented as semiparametric estimators: the propensity score is based on a parametric model, but the relationship between the outcome variable and the propensity score is nonparametric (Huber et al. 2010).

The idea of matching is to compute a similarity measure and use the algorithm to match observations from the treatment group with their closest counterparts from the control group. The aim is to construct an adequate comparison group that replaces missing data and allows us to estimate $E(Y_{0i} | P_i = 1)$ without imposing additional a priori assumptions (Blundell and Costa Diàs, 2000). Objects are matched according to the estimated value of the similarity measure. The most straightforward algorithm is to choose for each object in the treatment group an object with the most proximal value of the similarity measure p in the control group. Usually, the propensity score (i.e. the probability of receiving treatment) is chosen for that purpose. Let us define a set of control units A_i such that only one comparison unit i belongs to A_i :

$$A_i = \{j \mid j \in \{1 \dots n\} : \min \|p_i - p_j\|\} \quad (4)$$

where $p_i = Pr(P_i = 1 | X_i)$, X_i is a vector characteristic of individual i , and $\| \cdot \|$ is a metric used to measure the distance between p_i and p_j . In case of the nearest neighbour matching set, A_i can be treated as a weighting matrix. The weighting matrix $P(i; j)$ is a square matrix with zeros and ones as elements. The

value of one signifies the closest neighbour (with zeros for all remaining objects). This type of matching is called one-to-one or pair matching. Each unit from the treatment group is linked with only one element in the control group.

This estimator is not efficient, as only one non-treated observation is matched to each treated one, independent of sample size. No other control observations, even if very similar to the treated ones, are used. However, despite its inefficiency, pair matching has some advantages over other methods. First, using only the closest neighbour should reduce bias at the expense of increased variance. Second, it is fairly robust to propensity score misspecification (see, Austin 2009). Third, the nearest neighbour matching estimator has good statistical properties if p_i and p_j are defined on a common set. The role of the researcher is to decide how to treat poorly matched observations (Lee, 2005, p. 89).

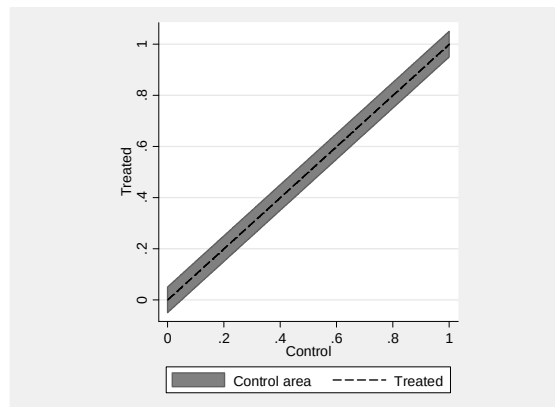
2.1. Caliper

One-to-one matching is characterised by the risk of there being poorly matched pairs (i.e. ones that are distant in terms of the chosen similarity measure). Caliper matching (Cochran and Rubin, 1973) is a variation of nearest neighbour matching that attempts to avoid poor quality matches by imposing a maximum allowable distance $\|p_i - p_j\|$. The impact of the caliper may be compared to that of the lens in a camera: when the focus is on a specific point, points at other distances are not visible. The procedure simply drops objects without closely matched counterparts in the non-treated group. In this case, the set of control units is defined as:

$$A_i = \{j \mid j \in \{1 \dots n\} : \min \|p_i - p_j\| < \delta\} \quad (5)$$

The set A_i is made of such objects j that their distance from the nearest match is less than δ . That is, a match for person i is selected only if $\|p_i - p_j\| < \delta$, where δ denotes a pre-specified level of tolerance. Treated persons for whom no matches can be found within the caliper range are excluded from the analysis; this constitutes a method of imposing a common support condition. Implementation of caliper matching may lead to smaller bias in regions in which similar controls are sparse. However, choosing an a priori reasonable tolerance level is an unresolved problem. Literature results have suggested small numbers, such as 0.005 or 0.001 (e.g. Austin 2009; Frölich 2004). Unfortunately, there is no established procedure by which to choose the size of caliper. The matching procedure without caliper seeks the closest match. After imposing a caliper, the area for reasonable matches is limited to the shaded region in Figure 1.

Figure 1. Impact of caliper



2.2. Trimming

The common support problem has been discussed by many authors (e.g. Heckman et al. 1999; Imbens 2004). Trimming is a method that helps to achieve fulfilment of the common support condition by using only those observations for which density of the propensity score exceeds a certain

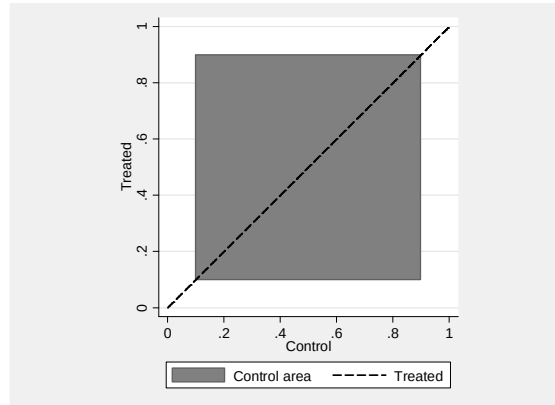
level (proposed by Heckman et al. 1998). This resolves the problem of lack of adequate comparison units; however, it may create other problems, like non-connected common support. In addition, their proposition is operationally burdensome.

Recently, Crump et al. (2009) proposed a very simple ‘rule of the thumb’ that allows for trimming of the support and provides overlap. Their approach is to remove observations with ‘extreme’ values of the propensity score in order to improve the precision of the estimator. They claim that the probability of finding a reliable counterpart for matching at the edge of the supported propensity-score range is negligible, so discarding ‘extreme’ values of the propensity score improves the overall estimation. Their ‘rule’ suggests dropping observations with estimated propensity scores lower than 0.1 and greater than 0.9. The set of control units is defined as

$$A_i = \{j \mid j \in \{1 \dots n\} : \min\|p_i - p_j\|; 0.1 < p_i, p_j < 0.9\} \quad (6)$$

However, a drawback of this procedure is the potentially increased bias of the ATT estimator, especially in situations in which distribution of the propensity score in the treated group is concentrated at the right tail.

Figure 2. Impact of trimming



Trimming has a different impact than the caliper mechanism has. It allows for unrestricted matching in the region allowed by the trimming rule. Therefore, we suspect that even with the imposition of the trimming mechanism, some poorly matched pairs will be included in the estimation process.

3. Monte Carlo study

The design of Monte Carlo simulation is partly borrowed from Frölich (2004). The experimental design is divided into two separate steps. The data are drawn in the first step, and in the second one, different matching techniques are used to estimate the ATT. During the first step, the covariates are drawn from known parametric distributions.

Next, observations are divided into treated and control populations according to the rule described in equation (5). The assignment is almost random due to the size of the stochastic component; hence, this method provides substantial overlap and allows for comparisons of both the caliper and trimming methods with unrestricted ATT estimation.

The sample used in the simulation study consists of five continuous and five discrete covariates. The first three covariates are continuous (X_1 ; X_2 ; X_3) and jointly normal with variances of 2, 1, and 1, and covariances of 1, -1, and -0.5, respectively; X_4 is uniformly distributed on an interval of [-3,3], and X_5 is a $\chi^2(1)$ distribution. The variance and covariance reflect the real situations of some degree of dependency between unit characteristics. The discrete part of the sample consists of D_1 (a dummy

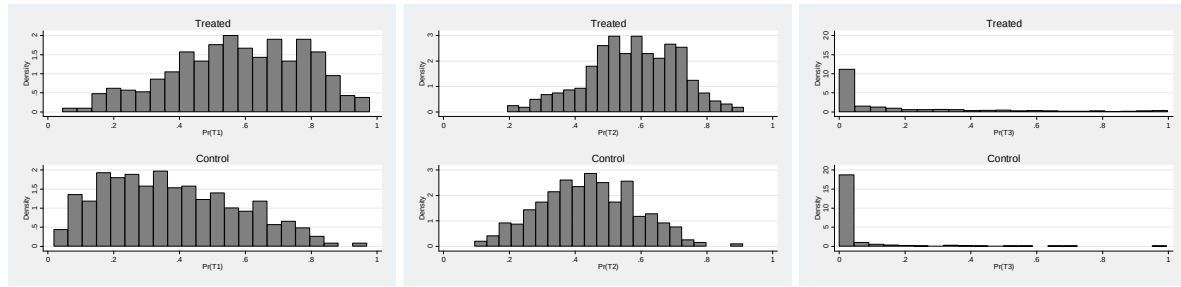
variable) and the set of variables (D_2 – D_5). The latter are created from a latent variable (LV) that is uniformly distributed on $[0,1]$. Hence, they are correlated. One might imagine that X_1 – X_5 represent *age*, *years of education*, *working experience*, any scaled variable (e.g. *opinion on quality of labour offices*), and *length of unemployment spell*, respectively. The D_1 dummy variable might be respondent *gender*, and D_2 – D_5 might represent *educational level*. This setting is similar to Rubin and Thomas (1996), who used a mix of dichotomous, ordinal, and continuous variables.

The treated ($T = 1$) and control ($T = 0$) populations are formed according to the rule

$$T = I(X_1 + 2X_2 - 2X_3 - X_4 - 0.5X_5 + D_1 + LV + \varepsilon > 0) \quad (7)$$

where $I(\cdot)$ is an indicator function. The expression in brackets is calculated and a value of T assigned for each unit i . Three different designs for the error term are considered. In the first two designs, the distribution of the error term is normal, with variances of 30 and 100, respectively. In the third design, the error term follows a $\chi^2(5)$ distribution; hence, it is asymmetric, skewed, and characterized by relatively high kurtosis. The distributions of the estimated propensity scores are summarized in Figure 3. In all designs, the probability mass in the treated group is located to the right of that of the non-treated group. This reflects the properties of the real data in which the probability of being treated is usually higher in the treated group.

Figure 3. Distributions of estimated propensity score



The distribution parameters are also used to control the treated-to-control ratio. Under a normally distributed error term, the treated-to-control ratio is equal to 1, whereas under a χ^2 distribution, it is 0.5. Three designs are considered for the outcome variable Y . They are summarised in Table 1:

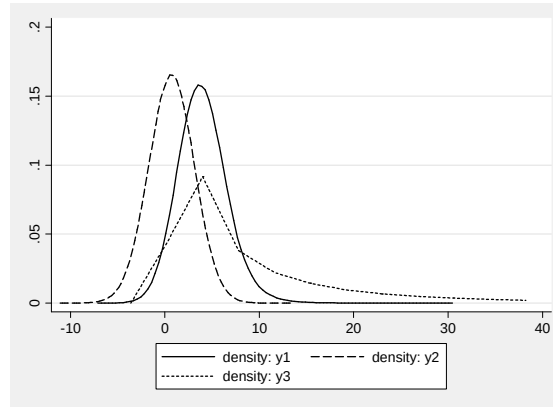
Table 1. Outcome equations for treated population

$Y_1 = X_1 + X_2 + X_3 - X_4 + X_5 + D_1 + D_2 + D_3 + D_4 + D_5 + u$
$Y_2 = X_1 + X_2 + 0.2X_3X_4 - \sqrt{X_5} + D_1 + D_2 + D_3 + D_4 + D_5 + u$
$Y_3 = (X_1 + X_2 + X_5)^2 + D_1 + D_2 + D_3 + D_4 + D_5 + u$

The first design is a linear function of independent variables and a small, normally distributed error u with mean 0 and variance 0.01. Y_2 is moderately non-linear, and design Y_3 heavily departs from linearity. Distributions of different outcome variables are presented in Figure 4. The linear distribution mimics the case in which treatment outcomes are higher for those subjects that are more likely to take part in the experiment. The non-linear case Y_2 can be considered a slight departure from standard assumptions. The resulting distribution is platokurtic and skewed to the left. The Y_3 specification heavily departs from normality: the outcome distribution is not centred at zero and has a fat right tail.

In our numerical experiment, the outcome value does not depend on treatment status. Hence, the value of the treatment effect is zero. This value was chosen for reasons of simplicity.

Figure 4. Distributions of the outcome variable



4. Simulation results

Three different distributions of the error term $\times$ three different designs for the outcome equation results in nine experimental designs. For each pair consisting of an error term and an outcome-data-generating process, two samples of different sizes are generated. The sample of 500 observations mimics a small sample size, and the sample of 1500 observations mirrors an entire population. The data generation step was replicated 100,000 times in order to provide robust results.

The propensity score is estimated as a linear probit equation. The treatment indicator D is the dependent variable, whereas the independent variables are X_1 – X_5 and D_1 – D_5 . Therefore, Y_1 is the only correct specification for the normal error term. The probit equation used for estimation is exactly the same as that used during the data generation process. For non-normal error u , all three specifications are incorrect. Specification Y_2 mildly departs from normality: the correlation between true and estimated propensity scores is 0.7. Specification Y_3 is more misspecified, with a heavy tail and correlation of 0.2.

In each table with a results column, ‘outcome’ defines which equation is used to generate outcome values, and column ‘distribution’ informs about the distribution and parameterisation of the error term. The next column contains estimated values of ATT and its standard errors using different methods. The third and fourth columns show statistics for the standard one-to-one matching; this result is used as a benchmark. The following columns consist of the results of matching via the caliper method and trimming. We compare three distinct caliper values with three values of the trimming parameter. The values were chosen so that a comparable number of observations was discarded using both methods.

Table 2. Sample size 500, caliper 0.005 vs. trimming 0.90

outcome	distribution	1:1 Mean	1:1 SE	caliper Mean	caliper SE	trimming Mean	trimming SE
Y_1	U_1	0.036	0.450	0.003	0.373	0.009	0.411
Y_1	U_2	0.014	0.354	0.003	0.337	0.010	0.351
Y_1	U_3	0.077	0.797	0.002	0.828	0.011	0.743
Y_2	U_1	0.057	0.361	-0.002	0.294	0.006	0.313
Y_2	U_2	0.019	0.309	-0.000	0.286	0.012	0.303
Y_2	U_3	0.083	0.562	0.008	0.573	-0.019	0.495
Y_3	U_1	0.624	2.374	0.023	1.925	0.109	2.134
Y_3	U_2	0.239	1.890	0.017	1.729	0.154	1.860
Y_3	U_3	1.350	5.349	-0.118	5.208	0.100	4.799

Legend: ‘outcome’ is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; ‘1:1’ stands for one-to-one matching, ‘caliper’ for caliper matching, ‘trimming’ for one-to-one matching with trimming, ‘SE’ standard error.

Table 3. Sample size 500, caliper 0.010 vs. trimming 0.95

outcome distribution		1:1 Mean	1:1 SE	caliper Mean	caliper SE	trimming Mean	trimming SE
Y_1	U_1	0.036	0.450	0.005	0.377	0.016	0.425
Y_1	U_2	0.014	0.354	0.003	0.339	0.013	0.353
Y_1	U_3	0.077	0.797	0.004	0.732	0.015	0.728
Y_2	U_1	0.057	0.361	-0.002	0.299	0.018	0.334
Y_2	U_2	0.019	0.309	0.000	0.292	0.017	0.308
Y_2	U_3	0.083	0.562	-0.005	0.508	-0.013	0.495
Y_3	U_1	0.624	2.374	0.038	1.965	0.222	2.220
Y_3	U_2	0.239	1.890	0.025	1.767	0.214	1.883
Y_3	U_3	1.350	5.349	-0.059	4.750	0.155	4.743

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and $U_3 \chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'SE' standard error.

Table 4. Sample size 500, caliper 0.050 vs. trimming 0.99

outcome distribution		1:1 Mean	1:1 SE	caliper Mean	caliper SE	trimming Mean	trimming SE
Y_1	U_1	0.036	0.45	0.014	0.418	0.030	0.445
Y_1	U_2	0.014	0.354	0.006	0.348	0.014	0.354
Y_1	U_3	0.077	0.797	0.026	0.712	0.028	0.744
Y_2	U_1	0.057	0.361	0.015	0.335	0.044	0.356
Y_2	U_2	0.019	0.309	0.005	0.303	0.019	0.309
Y_2	U_3	0.083	0.562	-0.002	0.499	0.000	0.517
Y_3	U_1	0.624	2.374	0.214	2.201	0.485	2.339
Y_3	U_2	0.239	1.890	0.086	1.845	0.239	1.890
Y_3	U_3	1.350	5.349	0.362	4.762	0.354	4.915

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and $U_3 \chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'SE' standard error.

In small samples, for the normally and almost normally distributed outcome variables (Y_1 and Y_2 , respectively), the estimated ATT values do not differ significantly from zero, at usual 5% level. In the case of heavily non-linear outcomes (Y_3), departure from zero is observed, but still, the size of this effect does not differ significantly from zero. For the linear and almost-linear outcomes, the estimated standard error is the lowest for caliper matching. The standard error of matching with trimming is lower than that of one-to-one, unrestricted matching. For the non-linear outcome, the performance of caliper matching and matching with trimming is comparable.

Table 5. Sample size 500, caliper 0.005 vs. trimming 0.90

outcome distribution		1:1 RMSE	caliper RMSE	trimming RMSE
Y_1	U_1	0.332	0.213	0.272
Y_1	U_2	0.189	0.170	0.187
Y_1	U_3	1.119	1.211	0.960
Y_2	U_1	0.194	0.122	0.146
Y_2	U_2	0.129	0.112	0.125
Y_2	U_3	0.555	0.569	0.433
Y_3	U_1	12.715	6.693	9.228
Y_3	U_2	6.964	5.357	6.694
Y_3	U_3	59.502	54.195	45.096

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and $U_3 \chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'RMSE' root mean standard error.

Table 6. Sample size 500, caliper 0.010 vs. trimming 0.95

outcome	distribution	1:1 RMSE	caliper RMSE	trimming RMSE
Y_1	U_1	0.332	0.217	0.291
Y_1	U_2	0.189	0.172	0.189
Y_1	U_3	1.119	0.894	0.902
Y_2	U_1	0.194	0.126	0.163
Y_2	U_2	0.129	0.115	0.128
Y_2	U_3	0.555	0.431	0.421
Y_3	U_1	12.715	7.155	10.336
Y_3	U_2	6.964	5.668	6.897
Y_3	U_3	59.502	41.656	43.043

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'RMSE' root mean standard error.

Table 7. Sample size 500, caliper 0.050 vs. trimming 0.99

outcome	distribution	1:1 RMSE	caliper RMSE	trimming RMSE
Y_1	U_1	0.332	0.277	0.322
Y_1	U_2	0.189	0.182	0.189
Y_1	U_3	1.119	0.834	0.940
Y_2	U_1	0.194	0.160	0.187
Y_2	U_2	0.129	0.123	0.129
Y_2	U_3	0.555	0.412	0.452
Y_3	U_1	12.715	10.01	12.078
Y_3	U_2	6.964	6.446	6.963
Y_3	U_3	59.502	41.992	46.433

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'RMSE' root mean standard error.

The size of the root mean squared error (RMSE) provides synthetic information about the statistical quality of the results. For a tight caliper and heavy trimming (Table 5), the caliper-based estimates are better in terms of RMSE for all simulations with normally distributed error. With an error term having a χ^2 distribution, matching with trimming was the best option studied. In case of average caliper size and average trimming (Table 6), in all but one simulation, the caliper method performed better. An exception is that of the mildly non-linear specification and χ^2 error distribution. The loose caliper and trimming methods provide similar results, but still, the caliper method is slightly better.

Table 8. Sample size 1500, caliper 0.005 vs. trimming 0.90

outcome	distribution	1:1 Mean	1:1 SE	caliper Mean	caliper SE	trimming Mean	trimming SE
Y_1	U_1	0.016	0.265	0.001	0.221	0.002	0.237
Y_1	U_2	0.005	0.203	0.001	0.196	0.003	0.202
Y_1	U_3	0.032	0.469	0.006	0.399	0.004	0.425
Y_2	U_1	0.028	0.213	0.000	0.176	0.002	0.180
Y_2	U_2	0.007	0.177	-0.001	0.169	0.004	0.174
Y_2	U_3	0.035	0.328	-0.005	0.277	-0.011	0.281
Y_3	U_1	0.313	1.443	0.014	1.193	0.031	1.274
Y_3	U_2	0.097	1.112	0.006	1.055	0.054	1.099
Y_3	U_3	0.616	3.268	0.033	2.755	0.028	2.859

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'SE' standard error.

Table 9. Sample size 1500, caliper 0.010 vs. trimming 0.95

outcome	distribution	1:1 Mean	1:1 SE	caliper Mean	caliper SE	trimming Mean	trimming SE
Y_1	U_1	0.016	0.265	0.001	0.231	0.004	0.247
Y_1	U_2	0.005	0.203	0.001	0.198	0.005	0.203
Y_1	U_3	0.032	0.469	0.007	0.397	0.006	0.418
Y_2	U_1	0.028	0.213	0.001	0.184	0.006	0.194
Y_2	U_2	0.007	0.177	0.000	0.172	0.006	0.177
Y_2	U_3	0.035	0.328	-0.005	0.277	-0.009	0.283
Y_3	U_1	0.313	1.443	0.021	1.250	0.070	1.335
Y_3	U_2	0.097	1.112	0.008	1.073	0.087	1.109
Y_3	U_3	0.616	3.268	0.091	2.767	0.040	2.836

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'SE' standard error.

Table 10. Sample size 1500, caliper 0.050 vs. trimming 0.99

outcome	distribution	1:1 Mean	1:1 SE	caliper Mean	caliper SE	trimming Mean	trimming SE
Y_1	U_1	0.016	0.265	0.010	0.257	0.012	0.262
Y_1	U_2	0.005	0.203	0.003	0.202	0.005	0.203
Y_1	U_3	0.032	0.469	0.024	0.454	0.010	0.433
Y_2	U_1	0.028	0.213	0.016	0.206	0.020	0.21
Y_2	U_2	0.007	0.177	0.003	0.176	0.007	0.177
Y_2	U_3	0.035	0.328	0.021	0.318	-0.004	0.299
Y_3	U_1	0.313	1.443	0.186	1.395	0.225	1.422
Y_3	U_2	0.097	1.112	0.045	1.101	0.096	1.112
Y_3	U_3	0.616	3.268	0.461	3.167	0.113	2.975

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'SE' standard error.

In larger samples of 1500 observations, the estimated treatment effect size is very close to zero for the normal and close-to-normal distributions of the outcome variable. In the case of the heavily non-linear outcome distribution, both methods of controlling for overlap provide better results than one-to-one matching without restriction. When the caliper size is set tightly and the trimming size is relatively large, evidently, matching via the caliper method provides more-precise estimates. Only in case of low trimming and a wide caliper, both methods give estimates of similar magnitudes. In those simulations, a relatively low share of total observations is discarded. When a larger number of observations are omitted, the caliper estimates have better statistical properties than the ones produced by the trimming method.

Table 11. Sample size 1500, caliper 0.005 vs. trimming 0.90

outcome	distribution	1:1 RMSE	caliper RMSE	trimming RMSE
Y_1	U_1	0.114	0.075	0.090
Y_1	U_2	0.062	0.057	0.061
Y_1	U_3	0.378	0.254	0.307
Y_2	U_1	0.065	0.043	0.047
Y_2	U_2	0.041	0.038	0.040
Y_2	U_3	0.183	0.123	0.136
Y_3	U_1	4.595	2.576	3.158
Y_3	U_2	2.261	1.944	2.197
Y_3	U_3	21.246	12.654	14.890

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'RMSE' root mean standard error.

Table 12. Sample size 1500, caliper 0.010 vs. trimming 0.95

outcome	distribution	1:1 RMSE	caliper RMSE	trimming RMSE
Y_1	U_1	0.114	0.082	0.098
Y_1	U_2	0.062	0.059	0.062
Y_1	U_3	0.378	0.251	0.292
Y_2	U_1	0.065	0.047	0.054
Y_2	U_2	0.041	0.039	0.041
Y_2	U_3	0.183	0.123	0.134
Y_3	U_1	4.595	2.963	3.637
Y_3	U_2	2.261	2.036	2.249
Y_3	U_3	21.246	13.046	14.510

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'RMSE' root mean standard error.

Table 13. Sample size 1500, caliper 0.050 vs. trimming 0.99

outcome	distribution	1:1 RMSE	caliper RMSE	trimming RMSE
Y_1	U_1	0.114	0.105	0.110
Y_1	U_2	0.062	0.061	0.062
Y_1	U_3	0.378	0.348	0.313
Y_2	U_1	0.065	0.060	0.063
Y_2	U_2	0.041	0.041	0.041
Y_2	U_3	0.183	0.168	0.148
Y_3	U_1	4.595	4.110	4.377
Y_3	U_2	2.261	2.190	2.260
Y_3	U_3	21.246	19.324	16.368

Legend: 'outcome' is as defined in table 1, distribution U_1 is $N(0,30)$, U_2 is $N(0,100)$ and U_3 $\chi^2(5)$; '1:1' stands for one-to-one matching, 'caliper' for caliper matching, 'trimming' for one-to-one matching with trimming, 'RMSE' root mean standard error.

The results of RMSE analysis of the caliper and trimming methods in large samples indicate that trimming provided more precise estimates than the caliper method in only 3 out of 27 simulations. This particular result includes all specifications of outcome variable (Y) and χ^2 error distribution.

5. Summary and conclusions

The problem of overlap is an important issue in PSM. In the presented simulation study, we used an experimental design similar those presented in the literature (Frölich 2004; Rubin and Thomas 1996) and have analysed the small- and large-sample properties of two competing methods for controlling for overlap. We pay more attention to the small-sample properties of both methods, because in large samples, the overlap problem is largely irrelevant. We address the question of efficiency and shed some light on this particular issue. Our simulation-based study showed two important results. First, trimming the support using the method from Crump et al. (2009) is inferior to the standard caliper mechanism. In small samples, trimming outperformed the caliper method in only four simulations with non-linear outcomes and χ^2 error distributions. This implies that when the variable of interest and the outcome follow rather standard continuous or discrete distributions, the caliper method is an efficient tool to control for overlap. On the other hand, trimming drops observations with very low and very high probabilities of treatment. This causes larger bias and greater variance. In larger samples, where the problem of overlap is less severe, the caliper mechanism outperforms trimming. Especially, when the caliper is configured as tight and the trimming size considerably large, the caliper method evidently provides more-precise estimates. In only three simulations was the RMSE of the trimming method lower than that of the caliper method when a χ^2 -distributed error term was used.

Second, we conclude that misspecification of the propensity score is not a significant problem for estimation of the propensity score with linear specification. Often, RMSE for the Y_2 outcome turned out to be lower than that for Y_1 . More-severe bias is caused by a non-standard distribution of the outcome variable: the RMSEs for the Y_3 specification were considerably higher than those for the two other equations. Further, a non-standard error distribution leads to increased bias. The aforementioned problems are more severe in smaller than larger samples.

References

1. P. Austin *Some methods of Propensity Score Matching Had Superior Performance to Others: Result of an Empirical Investigation and Monte Carlo Simulations*, Biometrical Journal 5 (2009), pp. 171-184.
2. R.Blundell and M.Costa-Diàs *Evaluation Methods for Non-Experimental Data*, Fiscal Studies, 21 (2000), pp. 427-468.
3. M.Busso, J.DiNardo and J.McCrary *New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators*, IZA DP, 3998 (2009).
4. W.Cochrane and D.Rubin *Controlling Bias in Observational Studies. A Review*, Sankhya, 35 (1973), pp. 417-466.
5. R.Crump, J.Hotz, G.Imbens and O.Mitnik *Dealing with limited overlap in estimation of average treatment effects*, Biometrika, 96 (2009), pp. 187-199.
6. M.Frölich *Finite sample properties of propensity-score matching and weighting estimators*, Review of Economics and Statistics, 86 (2004), pp. 77-90.
7. J.J. Heckman, H. Ichimura, J. Smith and P. Todd *Characterizing selection bias using experimental data*, Econometrica, 66 (1998), pp. 1017-1098.
8. J.J. Heckman, H. Ichimura and P. Todd *Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme*, Review of Economic Studies, 64 (1997), pp. 605-654.
9. D. Huber, M. Lechner and C. Wunch *How to Control for Many Covariates? Reliable Estimators Based in the Propensity Score*, IZA DP, 5268 (2010).
10. G. Imbens *Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review*, Review of Economic and Statistic. 86 (2004), pp. 4-29.
11. M-J. Lee *Micro-Econometrics for Policy, Program, and Treatment Effects*, Oxford University Press (2005).
12. P. Rosenbaum and D. Rubin *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika, 70 (1983), pp. 41-55.
13. D.Rubin and N.Thomas *Matching Using Estimated Propensity Score: Relating Theory to Practice*, Biometrics, 52 (1996), pp. 249-264.
14. J. Smith and P. Todd *Does Matching Overcome LaLonde's Critique of nonexperimental estimators?*, Journal of Econometrics, 125 (2005), pp. 305-353.
15. P. Strawiński *Dynamic Caliper Matching*, Central European Journal of Economic Modelling and Econometrics, 3 (2011), pp. 97-110.
16. Z. Zhao *Using Matching to Estimate treatment Effects: Data Requirements, Matching Methods, and Monte Carlo evidence*, The Review of Economic and Statistic, 86 (2008), pp. 91-107.



FACULTY OF ECONOMIC SCIENCES
UNIVERSITY OF WARSAW
44/50 DŁUGA ST.
00-241 WARSAW
WWW.WNE.UW.EDU.PL