PRZEGLĄD STATYSTYCZNY
R. LX – ZESZYT 3 – 2013

PAWEŁ STRAWIŃSKI

SMALL SAMPLE PROPERTIES OF MATCHING WITH CALIPER[1]

1. INTRODUCTION

Matched sampling methodology is used to mimic control experiments and estimate causal treatment effects. Unlike in the experimental sciences, in the social sciences fully controlled experiments usually are not possible to conduct due to various reasons. This methodology can be applied in all situations when one has a treatment, for example particular social policy, a group of treated individuals and a group of untreated individuals. The ultimate goal is to measure the average difference in outcome between treated and non-treated individuals, so called, the average treatment effect. The problem is that participants and non-participants usually differ in observed pre-treatment characteristics X. The second problem arises because an individual can be treated or non-treated but not both at the same time.

The role of matching is to adjust a non-treated group in order to make it comparable with a treated group. It allows us to alter the differences in the distributions of characteristics between a treated and a non-treated group. Even when the differences are observed it is necessary to adjust for those differences to obtain unbiased estimates of treatment effects. Matching estimators link units from a treated group with those from a non-treated group. Matching is typically done without replacement, so each non-treated observation is used as a match only once and matches are independent (Abadie, Imbens, 2011). The procedure seeks for each treated observation a non-treated observation, also called control observation, identical or very similar to it in observed characteristics. The most widely used types of matching are Mahalanobis distance matching, propensity score matching and kernel based matching. Among them the propensity score matching plays a fundamental role, since it reduces the curse of dimensionality problem and allows for one dimensional non-parametric regression (Rosenbaum, Rubin, 1983). The propensity score itself is a probability of receiving a treatment. It is also important to mention that matching on the propensity score is sufficient to balance the observed covariates. More recently developed nonparametric matching estimators described, for example, in Heckman et al. (1997, 1998) use

[1] This research is partly financed by the Ministry of Science and Higher Education grant N111 109335 (50%) and the Faculty of Economic Sciences Warsaw University grant (50%).

weighted average over multiple observations to construct matches. However, those methods are computationally demanding.

In this paper we focus on a particular matching technique that is, matching with caliper and its applications in small samples. The caliper mechanism has been proposed to control for matching quality and prevent from inexact matches. There exists a practical trade-off when obtaining matched samples between desires to (i) find matches for all treated units and (ii) use only matched treated-control pairs that are extremely similar to each other (Rosenbaum, Rubin, 1985b). The former situation leads to inexact matching, as matched objects may differ substantially, while the latter leads to incomplete matching.

The main motivation is that the finite sample properties of caliper mechanism applied to the propensity score have not been subject of a comprehensive study. In our work we try to shed some light on the properties of matching with caliper used to control the differences in the propensity score. Our work is broader than Austin (2009). We consider different distributions of the propensity score and the outcome equation. We also compare the nearest-neighbour matching with and without caliper. It is also worth mentioning that the literature discusses the properties of caliper applied to covariates. The novelty of our approach is that we show properties of caliper imposed on the propensity score. Our main result is that using caliper may lead to biased estimates. Usually, bias is heavy, especially when the outcome value is non-linearly related to the propensity score. Only when either outcome equation is constant or the propensity score distribution is uniform the matching with caliper procedure is able to provide unbiased results. Secondly, the bias becomes smaller as the caliper value increases. Thirdly, to efficiently control for poor matches, the size of the caliper should be within the range recently proposed by Austin (2009).

The article is organised as follows. Section 2 contains short literature review. Section 3 introduces notation, matching estimators and caliper mechanism in detail. In the fourth section we describe Monte Carlo experiment for different distributions of the propensity score and the outcome equations. In the fifth we present our main results. The last section summarises and concludes.

2. LITERATURE REVIEW

Several studies looked into asymptotic properties of matching procedures, to mention Heckman et al. (1998), Hirano, Imbens and Ridder (2003), Abadie and Imbens (2011). All of them show that most matching techniques provide consistent estimates for the average treatment effect; however, only few of them are efficient, such as the class of reweighting estimators. Secondly, pair matching is asymptotically inefficient (Abadie, Imbens, 2006). In a recent article, Abadie and Imbens (2011) showed that simple matching estimators may include bias and that bias does not disappear in large samples. On the other hand, there is only a limited number of

studies dealing with finite sample properties of matching estimators. Here it is worth noting the works by Frölich (2004), Austin (2009) and Busso, DiNardo and McCrary (2009).

Frölich (2004) examined 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 ridge matching procedure is lower than that of one-to-one only if the optimal bandwidth is known. Usually, the optimal value of bandwidth is not known *a-priori* and has to be estimated. Austin (2009) also compared several matching techniques in a Monte Carlo study. He concentrated mainly on one-to-one matching. All examined estimators resulted in a similar number of matched pairs and similar balance of variables between treated and untreated samples. Moreover, matching on the propensity score with caliper size not exceeding 0.03 tends to result in estimates with negligible relative bias. Similarly, Busso, DiNardo and McCrary (2009) and Huber, Lechner and Wunch (2013) emphasise the role of trimming to account for common support. Controlling for common support condition effectively improves matching performance regardless of the estimator used.

Deheija and Wahba (1999, 2002) evaluate the performance of propensity score matching methods, including pairwise matching and caliper matching. They find that these simple matching estimators succeed in closely replicating the experimental results. Smith and Todd (2001) reconcile those findings and show that matching estimates cause substantial bias. More recently, Zhao (2004) studied small sample properties of propensity score matching versus covariate matching estimators and those of different matching metrics. He showed that propensity score matching is a good choice when the correlation between covariates and the participation indicator is high. On the other hand, propensity score matching does not perform well in small samples in comparison with other estimators.

Matching estimators frequently suffer from bias. In a seminal work by Rosenbaum and Rubin (1985b) three sources of bias in matching were identified. The first is departure from strong ignorable treatment assignment, which means that assignment to treatment is based on observable pre-treatment variables only. The second one is bias due to incomplete matching and, finally, the third component is due to inexact matching. The bias caused by incomplete matching can be severe and is much worse than the bias due to inexact matching. Assuming that the strong ignorability condition holds, we want to assess empirically which source of matching bias makes larger distortions. Recently, Busso, DiNardo and McCrary (2009) showed that pair matching performs best in terms of bias among all procedures usually applied.

3. CALIPER MATCHING

The main problem in treatment effect literature is the estimation of the average treatment effect on the treated. We follow a standard notation. Let Y_{1i} be an outcome

when individual i receives a treatment and Y_{0i} when he or she does not. The latter situation is called control treatment. Let X be a vector of individual characteristics, $P_i \in \{0,1\}$ be an indicator of treatment status. The treatment effect for an individual i can be written

$$\tau_i = Y_{1i} - Y_{0i}. \tag{1}$$

The problem arise, because for each individual i one can observe either Y_{1i} or Y_{0i}. Hence, estimating the individual treatment effect τ_i is not possible and one has to concentrate on the average treatment effect on the treated (Caliendo, Kopeing, 2008).

The latter is defined as

$$ATT = E\left(Y_1 \mid P = 1, X\right) - E\left(Y_0 \mid P = 1, X\right). \tag{2}$$

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

$$\frac{1}{N}\sum_{i=1}^{N}\left[Y_{1i} - E\left(Y_{0i} \mid P_i = 1, X\right)\right], \tag{3}$$

where $E(Y_{0i} \mid P_i = 1, X) = \sum_{j=1}^{N} W(i,j)Y_{0j}$ is an estimator of the counterfactual state, $W(i,j)$ is a matrix of distance between i and j, and N is a number of matched pairs. The fundamental problem of inference is that, for each individual we can observe only one of these potential outcomes, because each unit will receive either treatment or control, not both. The estimation of treatment effects can thus be thought of as a missing data problem (Rubin, 1973), where we are interested in replicating the unobserved potential outcomes.

It is assumed that, conditional on all factors X that influence the potential outcome and the decision to participate, P is independent of Y_0. This assumption has several names in the literature. It is called unconfoundness, conditional independence or overlap, or selection on observables (Imbens, 2004). The counterfactual mean can be identified, provided that the support of X for the treated sample is contained in the support for X in the non-treated one. This property is called common support condition. An additional assumption is the Stable Unit Treatment Value Assumption (Rubin, 1980), which states that the outcomes of one individual are not affected by treatment assignment of any other individual.

The idea of matching is to compute a similarity measure and use the algorithm to match observations from the treatment group with their closest counterpart from the control sample. The aim is to construct an adequate comparison group that replaces missing data and allows us to estimate $E(Y_{0i}|P=1,X)$ without imposing additional *a-priori* assumptions (Blundell and Costa-Dias, 2009). Objects are matched according

to the estimated value of the similarity measure. The straightforward algorithm is to choose for each object in the treatment sample an object with the most proximal value of the similarity measure p_i from the control sample. Usually the propensity score, $p_i = \Pr(P = 1 \mid X_i)$, which is the probability of receiving the treatment by individual i with characteristics X_i, is chosen for that purpose. Let us define a set C_i such that for each unit i only one comparison unit j belongs to C_i:

$$C_i = \left\{ X_{j,j \neq i} \mid j \in \{1 \ldots n\} : \min \left\| p_i - p_j \right\| \right\}, \tag{4}$$

where ||.|| is a metric. In case of the nearest neighbour matching, set C_i can be treated as weighting matrix. The weight matrix $W(i,j)$ is a square matrix with zeros and ones as elements. The value one is assigned to the closest neighbour, and zeros to all remaining units. This type of matching is called one-to-one matching. Each unit from the treatment group is linked with only one element in the control group.

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 evaluator is to decide how to treat poorly matched observations (Lee 2005, pp. 89). The total distance, the average distance or the median distance between matched pairs p_i-p_j may be viewed as a measure of matching quality (Rosenbaum, 1985). The lower the measure, the better the fit. For the ideal procedure all quality measures should equal 0. Relying on all matched pairs regardless of matching quality may affect the balance. The balance is a weaker condition than close matching within each pair, and since it is weaker it can often be attained when close matching within pairs is not possible. Rosenbaum and Rubin (1983) showed that balancing two samples on the propensity score is sufficient to equalise covariate distributions. On the other hand, if a large number of poorly matched pairs were left out, the size of the control sample shrinks which means that for certain observations in the treatment group there is no adequate comparison in the control sample. As a result, they are dropped from the analysis. This would help with the balance but at the cost of efficiency, because some information is not used. The evaluator has to choose between the bias due to inexact matching and bias and increased variance due to incomplete matching.

One-to-one or one-to-many matching is characterised by the risk of having poorly matched pairs, that is, pairs distant in terms of the chosen similarity measure. The caliper matching (Cochran, Rubin, 1973) is a variation of the nearest neighbour matching that attempts to avoid poor quality matches by imposing a tolerance of the maximum distance $||p_i$-$p_j||$ allowed. The impact of the caliper may be compared to the focus in a camera. When attention is paid to a specific point, other distant points are not visible. The procedure simply drops objects without a close match in the control group

$$C_i = \left\{ X_{j,j \neq i} \mid j \in \{1 \ldots n\} : \min \left\| p_i - p_j \right\| < \delta \right\}. \tag{5}$$

The set C_i is made of such objects j, that their distance from the nearest match is not greater than δ. That is, a match for person i is selected only if $||p_i-p_j|| < \delta$, where δ is the pre-specified tolerance. Treated persons for whom no matches can be found within caliper are excluded from the analysis, which is one way of imposing a common support condition. Implementation of caliper matching may lead to a smaller bias in regions where similar controls are sparse. An unresolved problem is choosing an *a-priori* reasonable value for tolerance level.

Rosenbaum and Rubin (1985b) discuss the choice of the caliper size, generalizing the results from Table 2.3.1 of Cochran and Rubin (1973). When variance of the linear propensity score in the treatment group is twice as large as that in the control group, a caliper of 0.2 standard deviation removes 98% of the bias in a normally distributed covariate. Rosenbaum and Rubin generally suggest a caliper of 0.25 standard deviation of the linear propensity score. However, in the analysis they considered matching on the Mahalanobis distance not on the propensity score.

Unfortunately, there is no single optimal value for the caliper. The literature suggests small numbers such as 0.005 or 0.001 (see Austin, 2009). The caliper reduces the bias of the average treatment effect estimator at the cost of an increased variance (Heckman et al., 1997). In a special case, when the propensity score distribution is the same in the treatment and the control group, the caliper cuts off the worst matched pairs and lower the bias without a significant increase in the estimator variance. The caliper also lowers the value of matching quality measures. The cost is a lower number of successfully matched pairs. As a consequence, the variance of the average treatment effect may increase. However, this is not a major concern as long as one is interested in a precise estimation of the ATT (Smith, Todd, 2005). On the other hand, Smith and Todd (2005) point out that the potential problem with a caliper is the lack of *a-priori* knowledge about its optimal value. It is common practice to set the value by trial and error.

4. MONTE CARLO STUDY DESIGN

In this section we describe a Monte Carlo simulation conducted to examine the properties of the propensity score matching with caliper and compare with one-to-one matching on the propensity score. Since the propensity score is unknown in general, it is assumed, that it is estimated in a semi-parametric way. In practice, logit, probit or linear probability model is used.

In the Monte Carlo experiment several characteristic of matching can be examined. For instance, different variants of matching procedure, different functional forms of the outcome variable, (in)dependency of the outcome from the propensity score, different distributions of covariates, different sample sizes and different proportion of treated observations to non-treated ones. This complexity makes a general experiment cumbersome. We decided to design the Monte Carlo experiment in such a way that

is as simple as possible on one hand, and as comprehensive as possible on the other. Therefore, we concentrate our attention on the most important features of the matching procedure. We decided not to estimate the propensity score; we instead assume that the propensity score values are drawn from known distributions.

The second pre-set element is the sample size and the proportion of treated to non-treated observations. We decided to work with moderate sample sizes of 500 observations. A number of that range is very common in literature and enables analysis of small sample properties. As shown by Frölich (2004), different proportions of treated and controls may result in different conclusions. We decided to hold this parameter constant and set its value to 1/3. This means that we have twice as many control observations as the treated ones. These proportions are usually found in empirical studies. Usually the control group is larger than the treated group but the difference is not large (see Frölich, 2004).

We considered three different distributions: uniform, normal and Johnson S_B. In the case of two latter distributions, the distribution in the treatment sample is concentrated at the right tail, while in the control sample it is concentrated at the left tail. This setup is used to mimic real differences between the treatment and the control group. The distributions are parameterised and rescaled in such a way that the support is always on the (0,1) interval. They are presented in Figure 1.

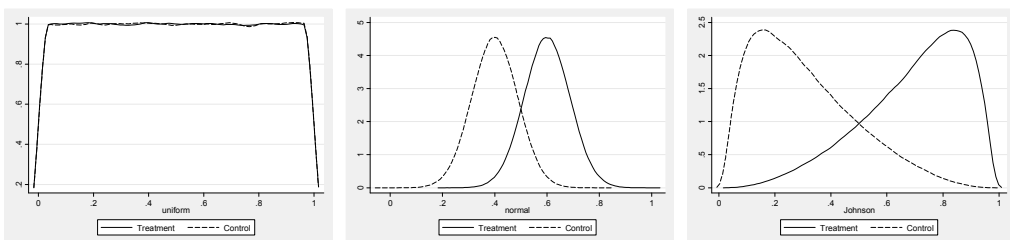

Figure 1. Propensity score distributions
Legend: solid line represent distribution in treated samples, dashed in controls ones.
Source: own computations.

Different distributions are used to replicate the behaviour of real data. The uniform distribution of the propensity score vector, presented on the left panel of Figure 1, is just used as a benchmark. The normal distribution, presented on the middle panel of Figure 1, is a picture of a rather ideal case in which linear combinations of object characteristics follow a normal distribution. In practice, normal distribution is only an approximation and true distribution may possess heavy tails or outliers. Nevertheless, the normal distribution of several characteristics is a common assumption in social sciences. On the right panel the propensity scores follow a Johnson S_B distribution. This is a very flexible distribution, described by four parameters, with a closed analytical form. Depending on the specific parameterisation, Johnson S_B distribution

can be similar to normal distribution, to asymmetric distribution, to distribution with heavy tails, to distribution with probability mass concentrated at the edge of support and to many others. Due to those properties it is frequently used in simulation based studies.

The next element in our numerical experiment is the functional form of the outcome equation. We depart from uniform curve and end up with a highly non-linear outcome. The outcome equations are summarized in Table 1 and presented on Figure 2. The uniform distribution mirrors the ideal case, when the value of treatment is the same for all objects. This distribution will also be used as a benchmark. The linear distribution reflects the situation in which objects that are more likely to take part in a program will benefit more. For instance, this is very common in social support programs. Two other non-linear curves are adapted from Frölich (2004). The m2 curve might represent a situation where the outcome depends discontinuously on an object characteristic strongly related to the propensity score. The m4 curve could be thought of as a reversal of the linear curve. The program pays outmost for those participants that are less likely to participate. Consider, for example, a job training program and education as a key determinant of the propensity score. Usually, well educated persons do not need such programs and are able to find a job without external help. The outcome for the non-treated population is set to be 0 for the reasons of simplicity.

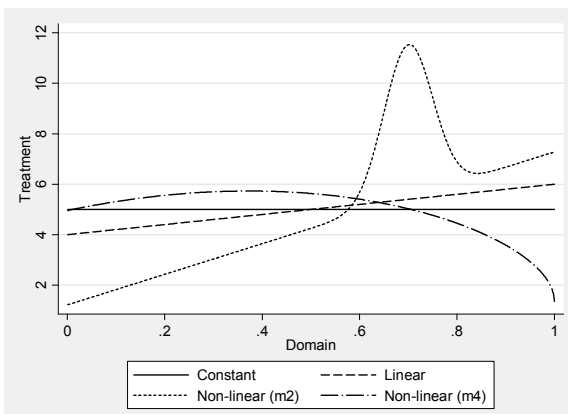

Figure 2. Distribution of the treatment effect

Source: own computations.

In the simulations we use one-to-one procedure and concentrate on caliper size and comparison of matching with and without caliper. The literature suggest rather narrow caliper, but then results are prone to be biased because many potential pairs would be discarded from analysis. The wider caliper allows for greater imbalance at the individual level, however it is easier to achieve balance in the propensity score

distribution between the treated and the control group. Therefore, we considered wide range of caliper sizes from 0.001 to 0.05. Our aim is to check the influence of a particular value of the caliper size on the estimate of ATT.

Table 1.

Outcome equations for the treated sample

Distribution	Outcome equation for the treated group
Constant	y = 5+e, e~U(0, .01)
Linear	y = 4+2*p+e, e~U(0, .01)
Non-linear m2	$y = 0.1+0.5*p+1/2*(\exp(-200*(p-0.7)^2))+e$, e~U(0, .01)
Non-linear m4	$y = 0.2+(1-p)^{0.5}-0.6*(0.9 - p)^2+e$, e~U(0, .01)

Please note that the non-linear curves are adjusted by a linear transformation to have a mean value of 5.

Having assigned specific values to all experiment parameters we are ready to compare the results of the simulation. In each numerical experiment we apply standard one-to-one matching and one-to-one matching with caliper to the same data set. The numerical experiments are designed in such a way that the "true" value of the ATT should be equal to 5 regardless of the distribution of the propensity score, functional form of outcome equation and the estimation technique. The small error term added to the outcome equation is purely random, and hence deviations from 5 no grater than 0.01 are meaningless. As the robustness check we ran simulations with larger random errors of 0.05 and 0.5, but the size of error turned out to have no impact on the final results.

5. EMPIRICAL RESULTS

The main results of our numerical experiment are presented in three separate tables. Each table consist of outcomes for only one distribution of the propensity score and all possible combinations of other parameters. The "1:1" column presents simple one-to-one nearest neighbour matching estimates and the "1:1 caliper" contains the result for matching with caliper. For each caliper size the number in the top row is an estimate of the ATT, while the number in the bottom row is the standard error of that ATT estimate. The results presented in Table 2 serve as kind of a benchmark for further results. They are obtained under the assumption of identical distribution of the propensity score in the treatment and the control group.

When estimated propensity scores have a uniform distribution, identical in the treatment and the control group, neither the shape of the outcome equation nor the size of the caliper have influence on estimates. Both methods, with and without caliper, provide identical results. It is worth emphasising that different caliper values cause a different number of matched pairs (see Table 5).

Table 2.

The ATT estimated with uniform distribution of propensity score

Treatment	constant		linear		m2		m4	
Caliper size	1:1	1:1 caliper	1:1	1:1 caliper	1:1	1:1 caliper	1:1	1:1 caliper
0.001	5.000	5.000	5.000	5.000	5.010	5.010	4.997	4.997
	0.000	0.000	0.045	0.064	0.215	0.307	0.069	0.099
0.005	5.000	5.000	5.000	5.000	5.010	5.011	4.997	4.998
	0.000	0.000	0.045	0.046	0.215	0.219	0.069	0.071
0.010	5.000	5.000	5.000	5.000	5.010	5.010	4.997	4.997
	0.000	0.000	0.045	0.045	0.215	0.215	0.069	0.069
0.020	5.000	5.000	5.000	5.000	5.010	5.010	4.997	4.997
	0.000	0.000	0.045	0.045	0.215	0.215	0.069	0.069
0.025	5.000	5.000	5.000	5.000	5.010	5.010	4.997	4.997
	0.000	0.000	0.045	0.045	0.215	0.215	0.069	0.069
0.050	5.000	5.000	5.000	5.000	5.010	5.010	4.997	4.997
	0.000	0.000	0.045	0.045	0.215	0.215	0.069	0.069

Please note that for each caliper size the number in the top row is an estimate of ATT and the one in the bottom row is its standard error.

Source: own computations.

Table 3.

The ATT estimated with normal distribution of propensity score

Treatment	constant		linear		m2		m4	
Caliper size	1:1	1:1 caliper	1:1	1:1 caliper	1:1	1:1 caliper	1:1	1:1 caliper
0.001	5.000	5.000	5.000	4.846	4.997	3.462	4.999	5.211
	0.001	0.001	0.013	0.016	0.148	0.098	0.021	0.015
0.005	5.000	5.000	5.000	4.897	4.997	3.796	4.999	5.159
	0.001	0.001	0.013	0.012	0.148	0.097	0.021	0.013
0.010	5.000	5.000	5.000	4.917	4.997	3.996	4.999	5.135
	0.001	0.001	0.013	0.012	0.148	0.104	0.021	0.013
0.020	5.000	5.000	5.000	4.934	4.997	4.224	4.999	5.112
	0.001	0.001	0.013	0.011	0.148	0.113	0.021	0.013
0.025	5.000	5.000	5.000	4.940	4.997	4.307	4.999	5.104
	0.001	0.001	0.013	0.011	0.148	0.118	0.021	0.013
0.050	5.000	5.000	5.000	4.959	4.997	4.611	4.999	5.074
	0.001	0.001	0.013	0.012	0.148	0.137	0.021	0.014

Please note that for each caliper size the number in the top row is an estimate of ATT and the one in the bottom row is its standard error.

Source: own computations.

In case of a normal distribution of the propensity score the results for one-to-one matching are similar to those in Table 2, despite the fact that the propensity score distribution in the treatment and the control group is different. Unfortunately, the caliper mechanism seems to cause bias to the results. The caliper estimates are unbiased for constant treatment. The effect of bias is moderate in case of linear treatment and non-linear m4 treatment. For non-linear m2 treatment equation the bias is heavy. In all cases the bias becomes smaller as the caliper size rises.

Those results indicate that the caliper mechanism can potentially distort the estimation results.

The last set of estimates present the results for a Johnson S_B distribution of the propensity score (Table 4.). In this simulation the pre-matching differences between the treated and the control group are the greatest. Despite that, one-to-one matching is able to provide unbiased estimates for all but one distribution of the outcome. The estimates for a non-linear m2 curve are biased and the bias significantly exceeds randomness of simulation. The results seem to indicate that matching with caliper is a much worse choice. Estimates are biased for all non-constant outcomes. For a linear and a non-linear m4 specification the bias of caliper matching is relatively large. In case of a non-linear m2 outcome the bias is larger than that of one-to-one matching.

Table 4.

ATT estimated with Johnson S_B distribution of propensity score

Treatment	constant		linear		m2		m4	
Caliper size	1:1	1:1 caliper	1:1	1:1 caliper	1:1	1:1 caliper	1:1	1:1 caliper
0.001	5.000	5.000	5.001	4.694	5.089	4.197	5.003	5.686
	0.000	0.000	0.026	0.052	0.135	0.322	0.068	0.073
0.005	5.000	5.000	5.001	4.809	5.089	4.796	5.003	5.523
	0.001	0.001	0.026	0.032	0.135	0.213	0.068	0.055
0.010	5.000	5.000	5.001	4.863	5.089	5.018	5.003	5.418
	0.001	0.001	0.026	0.029	0.135	0.189	0.068	0.054
0.020	5.000	5.000	5.001	4.906	5.089	5.099	5.003	5.314
	0.001	0.001	0.026	0.027	0.135	0.170	0.068	0.054
0.025	5.000	5.000	5.001	4.918	5.089	5.103	5.003	5.281
	0.001	0.001	0.026	0.027	0.135	0.165	0.068	0.055
0.050	5.000	5.000	5.001	4.952	5.089	5.094	5.003	5.180
	0.001	0.001	0.026	0.026	0.135	0.151	0.068	0.058

Please note that for each caliper size the number in the top row is an estimate of ATT and the one in the bottom row is its standard error.

Source: own computations.

Those results indicate that controlling for strict similarity of the estimated propensity score is not sufficient to obtain unbiased estimates of the ATT. This in turn implies that, in small samples, bias arising from inexact matching is relatively low in comparison with the one caused by incomplete matching, This is not an extraordinary finding, since in small samples the number of successfully matched pairs is low and each matched pair approximately accounts for 1% of total pairs. If one removes 20% or even 50% of total initial pairs this may cause a spectacular bias.

Table 5 illustrates the problem of diminishing matched pairs. As the caliper size shrinks, the number of successfully matched pairs decreases dramatically if the propensity scores are differently distributed in the treatment and in the control group. With the caliper value of 0.005, which is the one most frequently suggested in the literature, one would lose over 40% of matched pairs in case of a normal distribution of the propensity score, and nearly 45% in case of a Johnson S_B specification. The bias of the estimate is the result of a non-symmetric distribution of the outcome.

Table 5.

Number of successfully matched pairs

	Propensity score distribution					
	uniform		normal		Johnson S_B	
Caliper size	1 to 1 matching	caliper matching	1 to 1 matching	caliper matching	1 to 1 matching	caliper matching
0.001	165	81	165	52	165	83
0.005	165	159	165	96	165	93
0.010	165	165	165	112	165	115
0.020	165	165	165	126	165	132
0.025	165	165	165	130	165	136
0.050	165	165	165	140	165	149

Source: own computations.

To compare efficiency of both methods we decided to compute the RMSE statistics, presented in Table 6. With uniform distribution of the propensity score the RMSE for caliper matching is larger than for no caliper for all considered outcome equations. Results of simulations are similar for a normal and a Johnson S_B distribution of the propensity score. The caliper mechanism provides better results than matching without caliper for constant treatment. For all other functional forms of treatment equation the results showed that simple matching provides more precise estimates.

We derive two important implications from the above results. First, we confirm the theoretical results which implicitly assume constant character of treatment. Secondly, and more importantly, the caliper mechanism induces significant rise in the bias and the variance of the estimates in case of a non-uniform treatment. In empirical practice

it is very unlikely that the impact of treatment can be treated as uniform. Therefore, the usage of caliper introduces variance and this variance is larger than the reduction of variance due to lowering the bias.

Table 6.

Root Mean Squared Error

Treatment	constant		linear		m2		m4	
Distribution	no caliper	standard caliper	no caliper	standard caliper	no caliper	standard caliper	no caliper	standard caliper
uniform	0.000376	0.000378	0.045101	0.046571	0.216150	0.223129	0.070145	0.072525
normal	0.001139	0.000659	0.013216	0.155195	0.146735	1.223241	0.020911	0.160712
Johnson	0.000906	0.000571	0.026516	0.311774	0.163154	0.334864	0.068231	0.531482

RMSE computed for caliper size of 0.005.

Source: own computations.

6. CONCLUSIONS

In this article we studied properties of matching and matching with caliper in small samples. Despite that not much is known, both methods are widely used in applied evaluation research and it is important to understand their properties. We focused our attention on the properties of the ATT estimator. There are relatively few studies on the small-sample properties of different matching estimators. We tried to bridge the gap and shed some light on the problem.

To achieve that, we have used distributions that are either assumed in theoretical papers or employed in similar simulations. We confirmed the theoretical results and showed that simple one-to-one matching estimators are unbiased in most cases. At the same time the caliper method which is theoretically designed to control for the differences in distributions of the propensity score may introduce a substantial bias to the ATT estimator in small samples.

It turned out that in small samples the bias due to inexact matching is relatively small in comparison with that of incomplete matching. There are several reasons for this. First of all, there are only few unmatched observations. Secondly, if the covariates have the same distribution among matched and unmatched units the bias is limited. The third possibility is a constant value of treatment. Our empirical results suggest that even if the treatment is not constant, the value of bias is not large. On the other hand, the bias due to incomplete matching turned out to be substantial in our simulations, up to 15% for very conservative caliper size, and about 10% in case of the most popular 0.005 caliper.

Another practical problem with using a caliper is the loss of a significant number of possible matched pairs. This is of particularly great importance when the pool

of possible matched pairs is limited. A decrease in the number of matches caused by caliper is the primary reason why the caliper matching becomes incomplete. Moreover, the results show that this lack of completeness causes a significant bias to the ATT estimates.

There are several limitations to our simulations. First of all, we assumed that we work on propensity scores, not covariates. Our results are valid as long as distributions used in a simulation to mimic behaviour of the real data are close to empirical realisations. To achieve that, we have used distributions that are either assumed in empirical works or used in similar studies. To provide robustness of our results we alter the parametrisation of the distributions, and the results seem to be robust. Secondly, we concentrate on small sample behaviour of the ATT estimator. It is worth noting that 500 observations used in simulations give a maximum of 165 matched pairs. This number is rather low. Therefore, we replicate the part of the experiment for Johnson S_B distribution and a sample size of 1,500 (about 500 matched pairs) and the general picture does not change, however biases are much lower, about a half of those presented in Table 4. The replication of full experiment is hardly possible due to computational complexity. Nevertheless, the robustness is in our opinion confirmed with these results.

University of Warsaw

REFERENCES

[1] Abadie A., Imbens G., (2006), Large Sample Properties of Matching Estimators for Average Treatment Effects, *Econometrica*, 74 (1), 235–267.

[2] Abadie A. Imbens G., (2011), Bias-Corrected Estimates for Average Treatment Effects, *Journal of Business and Economic Statistics*, 29, 1–11.

[3] Austin P. (2009), Some Methods of Propensity Score Matching Had Superior Performance to Others: Result of an Empirical Investigation and Monte Carlo Simulations, *Biometrical Journal*, 5, 171–184.

[4] Blundell R., Costa-Diás M., (2000), Evaluation Methods for Non-Experimental Data, *Fiscal Studies*, 21 (4), 427–468.

[5] Busso M., DiNardo J., McCrary J., (2009), New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators, *IZA Discussion Paper*, 3998.

[6] Cochrane W., Rubin D., (1973), Controling Bias in Observational Studies. A Review, *Sankhya*, 35, 417–466.

[7] Dehejia R., Wahba S., (1999), Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Program, *Journal of American Statistical Association*, 94, 1053–1062.

[8] Dehejia R., Wahba S., (2002), Propensity Score Matching Methods for Nonexperimental Causal Studies, *Journal of the American Statistical Association*, 84, 151–161.

[9] Frölich (2004), Finite Sample Properties of Propensity-Score Matching and Weighting Estimators, *The Review of Economics and Statistics*, 86 (1), 77–90.

[10] Heckman J., Ichimura H., Smith J., Todd P., (1998), Characterizing Selection Bias Using Experimental Data, *Econometrica*, 66 (5), 1017–1098.
[11] Heckman J., Ichimura H., Todd P., (1997), Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme, *The Review of Economic Studies*, 64 (4), 605–654.
[12] Hirano K., Imbens G., Ridder G., (2003), Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, *Econometrica*, 71 (4), 1161–1189.
[13] Huber D., Lechner M., Wunch C., (2013), The Performance of Estimators Based on the Propensity Score, *Journal of Econometrics*, 175 (1), 1–21.
[14] Imbens G., (2004), Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics*, 86 (1), 4–29.
[15] Lee M-J., (2005) *Micro-Econometrics for Policy, Program, and Treatment Effects*, Oxford University Press.
[16] Rosenbaum P., Rubin D., (1983), The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70 (1), 41–55.
[17] Rosenbaum P., Rubin D., (1985a), Constructing Control Group Using Multivariate Matched Sampling Methods That Incorporate Propensity Score, *The American Statistician*, 39 (1), 33–38.
[18] Rosenbaum P., Rubin D., (1985b), Bias Due to Incomplete Matching, *Biometrics*, 41 (1), 103–116.
[19] Rubin D., (1973), Matching to Remove Bias in Observational Studies, *Biometrics*, 29, 159–183.
[20] Rubin D., (1980), Bias Reduction Using Mahalanobis Metric Matching, *Biometrics*, 36 (2), 293–298.
[21] Smith J., Todd P., (2001), Reconciling Conflicting Evidence on the Performance of Propensity--Score Matching Methods, *The American Economic Review*, 91 (2), 112–118.
[22] Smith J., Todd P., (2005), Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?, *Journal of Econometrics*, 125, 305–353.
[23] Zhao Z., (2004), Using Matching to Estimate Treatment Effects: Data Requirements, Matching Methods, and Monte Carlo Evidence, *The Review of Economics and Statistics*, 86 (1), 91–107.

WŁASNOŚCI MAŁOPRÓBKOWE ESTYMACJI PRZEZ DOPASOWANIE

Streszczenie

Mechanizm obcięcia (suwmiarki) jest szeroko stosowanym narzędziem zabezpieczającym przed słabo dopasowanymi połączeniami. W literaturze opisywane są własności asymptotyczne estymatorów z obcięciem. W artykule opisany jest eksperyment symulacji numerycznej badający własności tych estymatorów w małych i przeciętnie liczebnych próbach. Pokazujemy, że mechanizm obcięcia (suwmiarki) powoduje znaczne obciążenie estymatora przeciętnego efektu oddziaływania wobec jednostek poddanych oddziaływaniu (*ang*. ATT) i wzrost wartości jego wariancji w porównaniu ze standardowym estymatorem 1:1.

Słowa kluczowe: łączenie według prawdopodobieństwa, obcięcie (suwmiarka), eksperyment Monte Carlo, własności małej próby

SMALL SAMPLE PROPERTIES OF MATCHING WITH CALIPER

A b s t r a c t

A caliper mechanism is a common tool used to prevent from inexact matches. The existing literature discusses asymptotic properties of matching with caliper. In this simulation study we investigate properties in small and medium sized samples. We show that caliper causes a significant bias of the ATT estimator and raises its variance in comparison to one-to-one matching.

Keywords: propensity score matching, caliper, Monte Carlo experiment, finite sample properties