Central European Journal of Economic Modelling and Econometrics

Dynamic Caliper Matching

Paweł Strawiński*

Submitted: 20.10.2011, Accepted: 24.01.2012

Abstract

Matched sampling is a methodology used to estimate treatment effects. A caliper mechanism is used to achieve better similarity among matched pairs. We investigate finite sample properties of matching with caliper and propose a slight modification to the existing mechanism. The simulation study compare performance of both methods and show that standard caliper perform well only in case of constant treatment or uniform propensity score distribution. Secondly, in a case of non-uniform distribution and non-uniform treatment the dynamic caliper method outperform standard caliper matching.

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

JEL Classification: C14, C21, C52.

*University of Warsaw, email: pstrawinski@wne.uw.edu.pl

1 Introduction

Quasi-experimental methods are nowadays widely applied in evaluation studies. Their advantage, in comparison to fully controlled experimental design, is low cost. Matched sampling is a methodology for reducing bias due to observed covariates in comparative observational studies. However, even when matching on observable characteristics, it is necessary in order to estimate treatment effects to adjust for the difference in the distributions of those characteristics between treated and non-treated population. The most frequently used technique in application is pair matching, called also the nearest neighbour matching. The procedure seeks for each treated observation a non-treated counterpart with identical or very similar characteristics. In the adjustment process propensity score matching plays a fundamental role, since it reduces a course of dimensionality problem and allows for one dimension non-parametric regression (Rosenbaum and Rubin, 1983).

The caliper matching introduced in a work by Cochran and Rubin (1973) is a modification of the nearest neighbour matching procedure that impose a tolerance on the difference in characteristics between matched objects. Treated observations for which no matches can be found within a caliper are excluded from the analysis, which is one way of imposing a common support condition. A drawback of caliper matching is that it is difficult to know a priori what choice for tolerance level is reasonable (Todd, 2006).

In this paper we propose a slight modification of the caliper mechanism. We postulate that the size of the caliper should be retrieved from investigated data instead of choosing some ad hoc value. We call this procedure a dynamic caliper, as the size of the caliper depends solely on the estimated propensity score value. In other words, the size of the caliper is adjusted to the empirical data in the estimation process. A similar method was proposed by Rubin and Thomas (2000) but with a considerably larger caliper value on covariates. Their caliper value is 0.2 of the logit of the propensity score for the treated standarised in such way that variance in the control sample is equal to unity. This mean that the caliper value for instance is equal to $1/45$ for propensity score of 0.1, $1/5$ for propensity score of 0.5 and 2.25 for propensity score of 0.9, respectively. In other words, their caliper is very narrow for small value of propensity score and very wide for larger values. Furthermore, one-to-one matching estimators are widely used in empirical studies, and it is important to understand their properties. Thus, we analyse the properties of the dynamic caliper in comparison with the standard procedure, and show its strengths and weaknesses. Our main result is that a standard caliper performs poorly when treatment is not the same for all units. Secondly, we show that in case of non-uniform distribution of the propensity score and non-constant treatment the dynamic caliper method has a lower Root Mean Squared Error (RMSE) and hence is better than standard matching with a caliper.

The article is divided into four sections. Next section briefly introduces matching estimators. In the third section we describe Monte Carlo simulations for different

distributions of the propensity score and the outcome equations. In the fourth section we present our main results, while the fifth section summarises and concludes.

2 The 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 Y1i be an outcome when individual i receive a treatment and Y0i when he or she does not. The latter situation is called control treatment. Let Pi 0,1 be an indicator of treatment status. The average treatment effect on the treated (ATT) is defined as

$$ATT = E[Y_1|P = 1] - E[Y_0|P = 1] \tag{1}$$

Typical matching estimator has a form (Smith & Todd, 2005)

$$\frac{1}{N}\sum_{n=1}^{N}\left[Y_{1i} - E(Y_{0i}|P_i = 1)\right] \tag{2}$$

where $E(Y_{0i}|P_i = 1) = \sum W(i,j)Y_{0i}$ 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 causal effects can thus be thought of as a missing data problem (Rubin, 1973), where we are interested in predicting the unobserved potential outcomes.

It is assumed that conditional on all factors that influence the potential outcome for untreated $(Y_0|P_i = 1)$ and the decision to participate, P_i is independent of $\{Y_0, Y_1\}$. In other words, decision of individual to receive treatment or control treatment can be considered as random. This assumption is called unconfoundness, conditional independence, or overlap or selection on observables (Imbens, 2004). The counterfactual mean can be identified, provided that the support of all factors that influence outcome X among the treated is contained in the support of X among the non-treated. This property is called common support condition. An additional assumption is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980), which states that the outcomes of one individual are not affected by treatment assignment of any other individual. This means that the outcome of the program does not depend on the number of participants.

The idea of matching is to compute similarity measure and use the algorithm to match observations from the treatment group with their closest counterpart from the control group. The aim is a construction adequate comparison group that replaces missing data and allows to estimate $E(Y_{0i}|P_i = 1)$ without imposing additional *a priori* assumptions (Blundell & 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 group an object with the same or very close value of the similarity measure p from the control group. Usually the propensity score which is probability of receiving the treatment is chosen for that purpose. Let define set A_i such that only one comparison unit i belongs to A_i:

$$A_i = \{j | j \in 1 \dots n : \min ||p_i - p_j||\} \tag{3}$$

where $||.||$ is a metric. In case of the nearest neighbour matching set A_i can be treated as weighting matrix. The weight matrix $P(i, j)$ is a square matrix with zeros and ones as elements. The value one is for the closest neighbour, and zeros for all remaining objects. This type of matching is called 1 to 1 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 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 pi-pj may be viewed as a measure of matching quality (Rosenbaum, 1985). The lower measure the better fit. For the ideal procedure all quality measures should equal 0. Relying on all matched pair regardless matching quality may affect the balance. The balance is 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 (1985) showed that balancing two samples on the propensity score is sufficient to equalise covariate distributions. On the other hand, if large number of poorly matched pairs would be left out, the size of the control group shrinks and for certain observations in the treatment group can be no adequate comparison in the control group. 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 among the bias and the variance of the estimator.
One to one or one to many matching is characterised by the risk having poorly matched pairs that is pair that are distant in terms of chosen similarity measure. The caliper matching (Cochran and Rubin, 1973) is a variation of the nearest neighbour matching that attempts to avoid "bad" matches (those for which p_j is far from p_i) by imposing a tolerance of the maximum distance $||p_i - p_j||$ allowed.

$$A_i = \{j | j \in 1 \dots n : \min ||p_i - p_j|| < \delta\} \tag{4}$$

The set A_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 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. A unresolved problem is choosing *a priori* reasonable value for tolerance level.

Rosenbaum and Rubin (1985) 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 deviations removes 98% of the bias in a normally distributed covariate. Rosenbaum and Rubin generally suggest the 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 one optimal value for the caliper. The literature suggests small number such as 0.005 or 0.001 (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 cut off the worst matched pairs and lowers the bias without significant increase in estimator variance. The caliper also lowers the value of matching quality measures. The cost is 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 precise point estimation of the ATT (Smith and Todd, 2005). The bias of the estimate is reduced at the cost of increased variance. On the other hand, Smith and Todd (2005) point out that the potential problem with a caliper is a lack of *a priori* knowledge about its optimal value. It is common practice to set the value by trial and error.
We postulate to use as matching procedure slightly modified caliper mechanism

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

In this setting the caliper value is directly linked with estimated propensity score. For the observations with low treatment probability modified mechanism requires better matches from the control group in order to be included in computation of the ATT estimator value. In practice, there is a few such observations, but on the other hand, it is very likely that there is good counterfactual in the control group for them. A large number of matched pairs with low treatment probability could cause the ATT estimator to be biased. Therefore, in our opinion influence of observation with low value of the propensity score should be limited, despite that for those observations it is relatively easy to find a good counterfactual observation. In a situation where probability of participation approaches 1 dynamic caliper will have no major differences from the standard one. As a result, we expect that a greater number of matched pairs is left aside in the computation, those with low participation probability.

3 Monte Carlo study

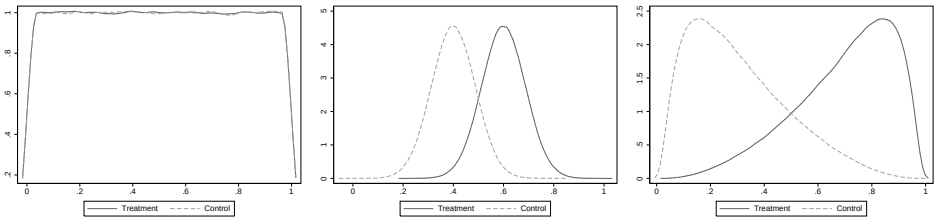
In this section we describe the Monte Carlo simulation conducted to examine the properties of the propensity score matching with dynamic caliper in comparison with

standard matching with caliper procedure. Since the propensity score is unknown in general, it is assumed that is estimated in a semi-parametric way.

The design of experiment involves several assumptions and pre-set parameters values. At the beginning we decided to work with moderate sample sizes, and we establish this parameter on 500. A number of that range is very common in this type of simulations founded in the literature. The next pre-set parameter value is a ratio of treated observations to control observations. Frölich (2004) has shown that the mean squared error of matching is lower and hence the quality of matching procedure is higher when control to treated ratio is higher than one to one and is low in cases where there are more treated observations than those in the control group. Relying on those results we decided to set a constant relation between the number of treated observation and the number of controls, and set this parameter to 1:2. The precise number is each simulation is determined stochastically. For each observation we draw a random number from standard uniform distribution and we include observation in the treated group if this random number is below 1/3. Otherwise, this particular observation is located in the control group. In this way, we receive on average 165 treated observations and 335 control observations. The following step involves setting the distribution of propensity score values. We considered three different distributions: uniform, normal and Johnson SB distribution. In a case of two latter distributions, the distribution in treatment group is concentrated at the right tail, while in the control group at left tail (see Figure 1).

The uniform distribution of the propensity score vector, presented on the left panel of Figure 1, is just used as a benchmark. The normal distributions, presented on the middle panel of Figure 1, are a picture of a rather ideal case in which most of characteristics follow a normal distribution. The normal distribution of several personal characteristics is a common assumption is social sciences. On the right panel the propensity scores follow a Johnson SB distribution. This is very flexible distribution, described by four parameters, and has a closed form. Those properties cause the fact that this distribution is frequently used in simulation based studies. The distributions are parameterised in such a way that propensity score values belongs to (0,1) interval.

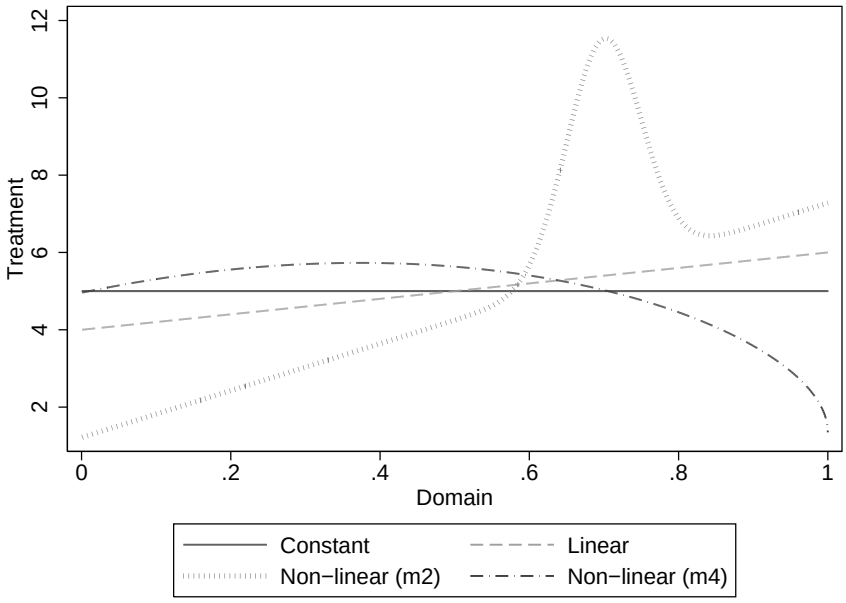
Figure 1: Propensity score distributions



Legend: Solid line represent distribution in treated groups, dashed in controls ones.

Another parameter that we control in simulation is a shape of the outcome in the treated population conditional on the propensity score value. We consider four different distributions; they are presented on Figure 2, and in Table 1.

Figure 2: Distribution of treatment effect



The uniform distribution mirrors the ideal case, when the value of the treatment is the same for all objects. This distribution will be also used as a benchmark. The linear distribution reflects the situation in which objects that are more likely to take a 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 non-linear m2 curve might represent situation where outcome depends discontinuously on object characteristic that is strongly related to the propensity score. The non-linear m4 curve could be thought as a reversal of linear curve. The program pays the most for those participants that are less likely to participate. Consider 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 last assumption involves the outcome value in the non-treated population and it is set to 0 for simplicity. Knowing the propensity score value and the outcome for all observations we were able to compare the result of standard caliper matching with our proposition of dynamic caliper matching. The construction of caliper mechanism

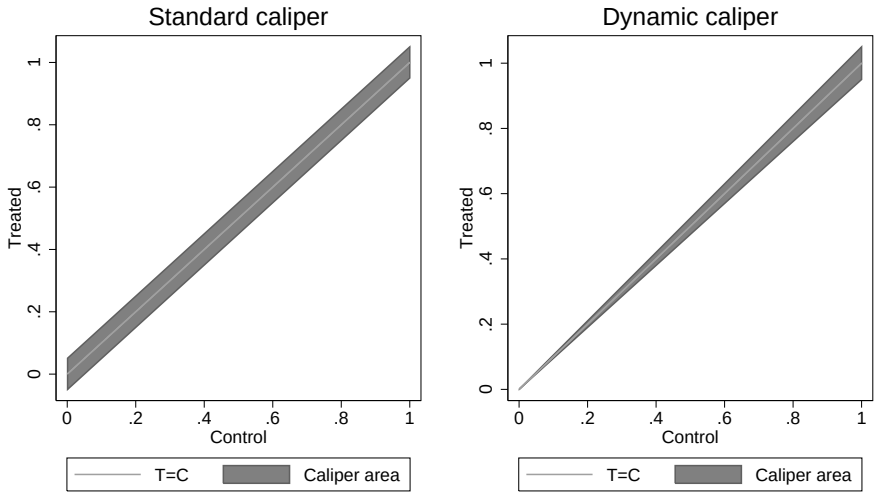
Table 1: Outcome equations for treated population

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

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

is different in both methods, as it is shown in equation (4) and (5). For the same numerical value of caliper parameter standard method seeks for comparison units in larger area. The shape of the area for allowed matches is rectangular in case of standard method, and triangular for dynamic caliper (see Figure 3). Thus, with the same parameter value in both mechanisms the size of the area for possible matches using dynamic caliper is a half of those in standard method. To neglect this difference in simulation, the caliper size in dynamic setting is going to be twice of that for standard caliper. The simulation is carried for all distributions of the propensity score vector and the functional forms for outcome equation with 10,000 replications.

Figure 3: Effect of caliper



Before moving to the result it is worth to note, that numerical experiment is designed in such a way that "true" value of the average treatment effect should be 5 regardless of the distribution of the propensity score vector and the functional form of the outcome equation. The small error added to the outcome equation causes that deviation from value of 5 no greater than 0.01 should be regarded as purely random. Conversely, larger deviations would be an indication of biasness of particular estimation technique. We also run simulations with larger errors of 0.1 and 0.5 but it has no impact on the final results.

4 Empirical results

The main results of our numerical experiment are presented in three separate tables. Each table consist outcomes for only one distribution of the propensity score and all possible combinations of other parameters are considered. The values in caliper size column refer to the size of caliper in standard approach. In case of dynamic caliper they are simply doubled.

The results presented in Table 2 are kind of benchmark to the further results. They are obtained under assumption of identical distribution of the propensity score in the treatment and the control group. In this case dynamic caliper method should be neither no better nor no worse than standard caliper matching. In case of the constant impact of treatment in fact there is no difference. However, when the impact of treatment is not uniform and depends on the value of the propensity score results show different pattern. With linear outcome equation standard caliper technique gives still unbiased results, while the results from dynamic caliper method are positively biased. Nevertheless, as the size of the caliper increases the bias is smaller, due to greater number of successfully matched pairs (see Table 5). The sizes of standard errors for both methods are on the same level. Similar results are observed for both non-linear specifications. Standard methods provide unbiased estimates, while results of estimation with the dynamic caliper mechanism are biased and the bias disappear as the caliper size increases.

In a situation in which distribution of the propensity score in the treatment group differs from those in the control group the results are different. Table 3 shows the situation when propensity score follows a normal distribution in both groups but with different mean value. As the size of the treatment is the same for all objects, both methods, that is, caliper and dynamic caliper, provide identical and unbiased results. In a situation with linear dependence between treatment value and propensity score value both methods result in downward biased estimates, and again results from both methods do not differ statistically from one another. For each pair we perform two-sided t-test. Beside that, the difference in estimates do not exceed one standard deviation. In simulations with nonlinear outcome equations both methods perform rather poorly and it is hard to decide which one is better. However, the results of the dynamic caliper mechanism are closer to the "true value" of 5 than those obtained

Table 2: The ATT estimated with uniform distribution of propensity score

Treatment	constant		linear		m2		m4	
Caliper size	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper
0.001	5.000	5.000	5.000	5.265	5.010	6.082	4.997	4.750
	0.000	0.001	0.064	0.057	0.307	0.295	0.099	0.119
0.005	5.000	5.000	5.000	5.123	5.011	5.508	4.998	4.941
	0.000	0.000	0.046	0.044	0.219	0.224	0.071	0.080
0.010	5.000	5.000	5.000	5.069	5.010	5.278	4.997	4.981
	0.000	0.000	0.045	0.044	0.215	0.218	0.069	0.075
0.020	5.000	5.000	5.000	5.036	5.010	5.148	4.997	4.994
	0.000	0.000	0.045	0.044	0.215	0.216	0.069	0.072
0.025	5.000	5.000	5.000	5.029	5.010	5.121	4.997	4.995
	0.000	0.000	0.045	0.044	0.215	0.215	0.069	0.072
0.050	5.000	5.000	5.000	5.015	5.010	5.066	4.997	4.997
	0.000	0.000	0.045	0.045	0.215	0.215	0.069	0.070

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

Table 3: The ATT estimated with normal distribution of propensity score

Treatment	constant		linear		m2		m4	
Caliper size	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper
0.001	5.000	5.000	4.065	4.072	3.462	3.518	5.211	5.202
	0.001	0.001	0.014	0.014	0.098	0.103	0.015	0.015
0.005	5.000	5.000	4.107	4.113	3.796	3.869	5.159	5.150
	0.001	0.001	0.010	0.010	0.097	0.101	0.013	0.013
0.010	5.000	5.000	4.124	4.129	3.996	4.083	5.135	5.126
	0.001	0.001	0.010	0.010	0.104	0.108	0.013	0.013
0.020	5.000	5.000	4.138	4.145	4.224	4.335	5.112	5.101
	0.001	0.001	0.010	0.010	0.113	0.119	0.013	0.013
0.025	5.000	5.000	4.143	4.150	4.307	4.430	5.104	5.092
	0.001	0.001	0.010	0.010	0.118	0.125	0.013	0.014
0.050	5.000	5.000	4.160	4.170	4.611	4.781	5.074	5.055
	0.001	0.001	0.010	0.010	0.137	0.146	0.014	0.016

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

from the standard method.

The last set of simulations deals with propensity score that follows Johnson S_B distribution. Again, when the treatment is a simple constant value there are no significant differences between two methods of estimations. In a case of linear distribution of the propensity score the ATT estimates obtained via dynamic caliper method are closer to the "true values" than those from a standard caliper method.

On the other hand, the differences are within one standard error with the exception for the smallest caliper value where is larger. Under non-linear outcome equation the picture is somewhat blur. For the m2 equation all but two results for dynamic caliper are closer to the "true" value than standard method. The similar results are observed for m4 equation, in most cases the dynamic caliper performs better than its standard counterpart. However, the estimates are significantly and positively biased.

Table 4: The ATT estimated with Johnson distribution of propensity score

Treatment	constant		linear		m2		m4	
Caliper size	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper
0.001	5.000	5.000	4.694	4.777	4.197	4.604	5.686	5.582
	0.000	0.000	0.052	0.048	0.322	0.315	0.073	0.079
0.005	5.000	5.000	4.809	4.851	4.796	4.952	5.523	5.439
	0.001	0.001	0.032	0.030	0.213	0.197	0.055	0.056
0.010	5.000	5.000	4.863	4.897	5.018	5.076	5.418	5.336
	0.001	0.001	0.029	0.028	0.189	0.174	0.054	0.055
0.020	5.000	5.000	4.906	4.935	5.099	5.099	5.314	5.231
	0.001	0.001	0.027	0.027	0.170	0.158	0.054	0.056
0.025	5.000	5.000	4.918	4.947	5.103	5.097	5.281	5.197
	0.001	0.001	0.027	0.027	0.165	0.153	0.055	0.058
0.050	5.000	5.000	4.952	4.983	5.094	5.089	5.180	5.074
	0.001	0.001	0.026	0.026	0.151	0.141	0.058	0.064

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

The last element of the simulation is to check the influence of the caliper method and it size on the number of successfully matched pairs, that is the number of those objects in the treated group for with there is a pair within a caliper distance in the control group.

As the number of matched pairs depend only on the distribution of propensity score, the table is common for all outcome equation specifications. With the uniform distribution of the propensity score, the caliper value equal or larger than 0.01 has no impact on the number of matched pairs. The dynamic version of caliper is, as it is expected, more conservative and prevents a greater number of poor matches.

With the normal distribution of the propensity score dynamic version of caliper allows for about 5% of possible matches more in comparison with standard procedure. However, as the caliper size increase the difference between two methods in term of the number of matched pairs become smaller. When propensity score follow a Johnson SB distribution the situation is very similar to those for normal distribution, except that in each cell there is greater number of successfully matched pairs.

The comparison of Root Mean Squared Error (RMSE) for both estimation methods confirms our results. To conserve space we show in Table 5 results for caliper of 0.005 only; other results are similar to those presented. When propensity scores

Table 5: Number of successfully matched pairs

	Propensity score distribution					
	uniform		normal		Johnson S_B	
Caliper size	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper
0.001	81	74	52	55	83	78
0.005	159	140	96	101	93	105
0.010	165	153	112	117	115	127
0.020	165	159	126	131	132	143
0.025	165	160	130	136	136	147
0.050	165	163	143	150	149	159

follow uniform distribution or treatment is constant, the standard caliper procedure provides unbiased results with low variance. If the value of treatment depends on the value of propensity score, a dynamic mechanism that adjusts caliper to the data has lower RMSE. The difference between the two methods is significant in the case of non uniform propensity score distribution and nonlinear outcome equation.

Table 6: Root Mean Squared Error

Treatment	constant		linear		m2		m4	
Distribution	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper	standard caliper	dynamic caliper
uniform	0.000378	0.000398	0.046571	0.131745	0.223129	0.585890	0.072525	0.100656
normal	0.000659	0.000673	0.893121	0.887302	1.223241	1.157174	0.160712	0.152604
Jonhson S_B	0.000571	0.000585	0.195951	0.155794	0.334864	0.244080	0.531482	0.452099

RMSE computed for caliper size of 0.005

5 Conclusions

The influence of the caliper mechanism on the estimation of the Average Treatment Effect on the Treated is not well recognised in the literature. On the other hand, the caliper is frequently used in applications to control for the balance between treated and non-treated population. In this paper we tried to shed some light on impact of caliper on the properties of the ATT estimator. We also have proposed a modification of the caliper mechanism and conduct a comparative study. We call our method the dynamic caliper. The name is rooted in fact that we postulate that the size of the caliper should be retrieved empirically from available data.

We show that the standard caliper matching provide unbiased estimates in specific situations. Namely, when the treatment is constant, that is in situation in which the influence of the treatment is the same for every treated subject, or the probability

of being treated is the same for all objects. With the propensity score distribution that is closer to the real empirical data our simulations indicate that the estimates of the ATT are biased and the RMSE's are quite large. Also we observe that the smaller caliper size comes along with the higher bias. This means that usually trade-off between achieving balance between the treated and the control group and unbiased estimates of the ATT is present.
The dynamic caliper is characterised by lower bias and lower variance in non-uniform propensity score and non-linear outcome setting. On the other hand, the dynamic caliper method performs poorly when the propensity score follows uniform distribution. The estimates are severely biased and have significantly larger RMSE in most cases. In simulations in which we assumed propensity score distribution that is close to the real data realizations, in most cases the dynamic caliper is better, in the sense that using that technique causes a lower bias and mean squared error. This result shows that the likelihood of obtaining a closer estimate to the true value is larger when using the dynamic caliper.

6 Acknowledgements

This research is partly financed by Ministry of Science and Higher Education grant N111 109335 (50%) and Faculty of Economic Sciences Warsaw University grant (50%). I am especially grateful to all comments to the earlier versions of paper during seminars at the Univeristy of Warsaw, PhD workshop in Krakow and Microeconometric conference at University of Szczecin. And finally I would like to thank the anonymous reviewer for final suggestions. All remaining errors are mine. The views expressed in the article are those of author and shall not be connected with the institutions.

References

[1] 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*, vol. 5, pp. 171-184.

[2] Blundell R., Costa-Diás M. (2009) Alternative Approaches to Evaluation in Empirical Microeconometrics, *Journal of Human Resources*, vol. 44, pp. 565-640.

[3] Cochrane W., Rubin D. (1973) Controling Bias in Observational Studies. A Review, *Sankhya*, vol. 35, pp. 417-466.

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

[5] Imbens G. (2004) Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review, *Review of Economics and Statistics* 86, 4-29.

[6] 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, 605-654.

[7] Lee M-J. (2005) *Micro-Econometrics for Policy, Program, and Treatment Effects*, Oxford University Press.

[8] Rosenbaum P. (1985) Optimal Matching for Observational Studies, *Journal of the American Statistical Association* 84, 1024-1032.

[9] Rosenbaum P., Rubin D. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika* 70, 41-55.

[10] Rosenbaum P., Rubin D. (1985) Constructing Control Group using Multivariate Matched Sampling Methods that Incorporate Propensity Score, *The American Statistician* 39, 33-38.

[11] Rubin D. (1973) Matching to Remove Bias in Observational Studies, *Biometrics* 29, 159-183.

[12] Rubin (1980) Bias Reduction using Mahalanobis Metric Matching, *Biometrics* 36, 293-298.

[13] Rubin D., Thomas N. (2000) Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates, *Journal of American Statistical Association* 95, 573-585.

[14] Smith J., Todd P. (2005) Does Matching Overcome LaLonde's Critique of nonexperimental estimators?, *Journal of Econometrics*, vol. 125, str. 305-353.

[15] Todd P. (2006) *Matching estimators*, unpublished manuscript.