STATISTICS IN TRANSITION-new series, December 2011
Vol. 12, No. 3, pp. 595—607

AN IMPROVEMENT OF QUALITY OF STATISTICAL MATCHING FOR SURVEY DATA USING DYNAMIC CALIPER

Paweł Strawiński[1]

ABSTRACT

Nowadays, matching is a widely used technique to estimate program net effects. The goal of the method is to establish a counterfactual state by choosing from the control pool a group that is similar to those in the treatment group. In this article we propose a modification of the matching with caliper procedure. The novelty in our approach is setting the caliper value as a fraction of estimated propensity score. The simulation results and examples are presented. Using Deheija and Wahba (1999) data benefits of the proposed approach are stressed. The obtained results indicate that proposed approach is more efficient than the one traditionally used.

Key words: matching, propensity score, caliper, evaluation.

1.Introduction

The objective of many empirical works is a replication of controlled experiment via non-experimental techniques. In such circumstances quasi-experimental methods are employed. Despite that they are frequently used in program evaluations, there is no consensus in the econometric literature about its efficiency (Smith and Todd, 2005).

Evaluation of the program impact always includes speculation how one would function without a program. At every moment one could be in only one alternative state, not in both. The one can be a program participant or non-participant. If the data would be experimental data, it would be sufficient to compare the outcome for experimental and control group. Therefore, evaluation problem is a missing data problem (Heckman et al., 1997). Propensity score matching methods are increasingly being used in observational studies to reduce the impact of treatment-selection bias in the estimation of causal treatment effects (Austin, 2009).

[1] Warsaw University. Email: pstrawinski@wne.uw.edu.pl.

In the statistical and econometric literature various matching methods are described that allow for reconstruction of control group from survey, that is non-experimental data. The most widely applied technique is a propensity score matching (Rosenbaum & Rubin, 1983). Its popularity arises from the implementations in various statistical packages. The propensity score itself is a probability of being a member of the treated group, for instance a program participant. Deheija and Wahba (1999, 2002) showed that propensity score could be used to match observations from experimental and non-experimental pool of potential controls. The role of matching procedure is to pick up from control pool a reference control group. In this way one is able to replicate the results of fully controlled experiment, particularly the outcome of the National Support Work Demonstration (NSW) experiment, and the obtained results would be slightly biased in comparison with the controlled experiment. The NSW was a temporary employment program designed to help disadvantaged workers lacking basic jobs skills move into labour market providing them work experience at subsidised positions (LaLonde, 1986). On the other hand, Smith and Todd (2005) also analysed the impact of the NSW on its participants and showed that the matching results may be sensitive to the specification of the propensity score vector and the comparison group.

Our work is located in this empirical research area. We show that with the use of Strawiński (2009) modified matching algorithm we are able to replicate more accurately the experimental results than using standard techniques. The aim of this research is not a result replication, but to point out that modified matching algorithm outperforms others that are well-established in the literature.

The remaining part of the article is organised as follows. In the next section we present the related literature. In the third section we briefly present the matching methods and explain Strawiński (2009) modification. In the fourth we present data and empirical results, while section five summarizes and concludes.

2. A brief review of the relevant literature

The literature describing various approaches to matching is enormous. The seminal paper we refer to is LaLonde (1986). The author evaluated economically the NSW program with experimental and quasi-experimental techniques. In the latter the control groups were chosen from Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID). Various estimation methods were employed from difference in mean values, though regression analysis, difference-in-differences to selection methods. The author showed that various estimation techniques provide different results that fundamentally differ from the experimental ones. LaLonde (1986) concludes that non-experimental methods are not efficient in replication of the experimental results and researchers should be aware of the potential specification errors in similar evaluations.

In similar work based on the same data Heckman and Hotz (1989) used wider range of estimation tools and obtained reverse results. They include in the models only pre-treatment covariates and successfully replicated the experimental results. Deheija and Wahba (1999) also explored the same dataset. Their aim was to make a judgement about quality of the propensity score matching in the evaluation context. The authors showed that it is sufficient to use simple matching estimators to closely replicate the experimental results. This is possible even when some assumptions of matching method are not satisfied (i.e. random participation).

Smith and Todd (2005) investigated influence of the participation in NSW on beneficent and showed that propensity score based results are sensitive to variable selection for score vector and for the comparison group. In their opinion, there is no well-established methodology and therefore the results are influenced by subjective decisions of the evaluators. They conclude that matching methods are useful, but their capability should not be exaggerated.

Relying on the cited literature one could conclude that there is no consensus between micro evaluators and many scientifically interesting aspects of the empirical evaluation methods require further investigation.

3. Matching methods

The goal of matching is estimation of program impact on participants. Let P be a variable that describes state of the involvement in the program and $P_i=1$ for person i being participant (member of treatment group) and $P_i=0$ for out-of the program person (member of control group). Let Y_{1i} be a value of the outcome variable when person i is participant, and let Y_{0i} be a value of the outcome variable for non-participant. The average treatment effect estimate is

$$Y_{ATE} = \frac{1}{N}\left[\sum_{i=1}^{N}\left(P_i Y_{1i} + (1-P_i)Y_{0i}\right)\right] \tag{1}$$

The program effect according to (1) is an effect for the average person from the population regardless his participation status. In practice the difference between the average value of the outcome variable for all treated and all non-treated unit is treated as (1). This may lead to bias, because treated units may systematically differ from the control units.

To circumvent the aforementioned problem another statistics is usually computed. It is the average treatment effect on the treated units (ATT). This is a measure of treatment limited to the treated group

$$Y_{ATT} = E\left(Y_{1i} - Y_{0i} \mid P_i = 1\right) = E\left(Y_{1i} \mid P_i = 1\right) - E\left(Y_{0i} \mid P_i = 1\right) \tag{2}$$

The Y_{ATT} is averaged over participants and shows quantitative impact of the program to them. Its value is equal to the average change in the outcome variable caused by participation. The value informs about a profit or a loss from participation for program beneficent. It is worth to note, that $E(Y_{1i}|P_i=1)$ is an expected value of the outcome variable for the program participants and is observed while $E(Y_{0i}|P_i=1)$ is non-observed counterfactual state.

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 one to estimate $E(Y_{0i}|P_i=1)$ without imposing additional a-priori assumptions (Blundell & Costa-Dias, 2000). The role of matching is a replication of experimental conditions from cross-sectional data by selection of the proper comparison units. The successfully constructed quasi-experimental reference group provides bias reduction and more precise results (Rubin, 1973).

Conditioning on all covariates available is not possible due to the course of dimensionality. Rosenbaum and Rubin (1983) proposed balancing scores $b(x)$, i.e. functions based on significant covariates, such that the conditional distribution of X on $b(X)$ is independent from the state of participation. Among balancing scores is propensity score, i.e. conditional probability of the program participation on observable characteristics X. Matching procedures that rely on balancing scores are known as propensity score matching (PSM).

Propensity score matching is a general method with different implementations and similarity measures. Firstly, the different functional forms for the propensity score can be chosen. Secondly, various similarity measures can be applied. Typical matching estimator has a form (Smith & Todd, 2005)

$$\frac{1}{N}\sum_{i=1}^{N} Y_{1i} - E(Y_{0i} \mid P_i = 1) \tag{3}$$

where

$E(Y_0 \mid P_i = 1) = \sum_{i=1}^{N} W(i,j) Y_{0i}$ 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. Objects are matched according to estimated value of similarity measure. The straightforward algorithm is to choose for each object in treatment group an object with the same or very close value of the similarity measure from the control group. Let us define set A_i such that only one comparison unit i belongs to A_i:

$$A_i = \left\{ j \mid j \in \{1 \ldots n\} : \min \left\| w_i - w_j \right\| \right\} \tag{4}$$

where ||.|| is a metric. Then, weight matrix *W(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 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 w_i and w_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 sum of distance, the average distance or the median distance between matched pairs w_i-w_j may be viewed as quality of matching measure. The lower measure the better fit. For the ideal procedure all quality measures should equal 0. When the nearest neighbour matching uses distance or propensity score as a similarity measure then in the infinite samples procedure has a zero variance (Orazio et al., 2009, p. 46). Relying on all matched pair regardless matching quality may affect the balance. On the other hand, if large number of poorly matched pairs would be left out, the size of the control group would shrink and for certain observations in the treatment group could 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 pairs that are distant in terms of chosen similarity measure. To ensure that only well-matched pairs are compared the caliper is used (Cochran & Rubin 1973). The impact of the caliper may be compared to the focus in the camera. When attention is paid to specific point, other distant points are not visible. The procedure simply drops objects without close match.

$$A_i = \left\{ j \mid j \in \{1 \ldots n\} : \min \left\| w_i - w_j \right\| < \delta w_i \right\} \tag{5}$$

The set A_j is made of such objects *j* that their distance from the nearest match is not greater than δ. Unfortunately, there is no one optimal value for the caliper. The literature suggests small number, such as 0.005 or 0.001 (see Austin (2009) for Monte Carlo results). The caliper reduces the bias of average treatment effect estimator at the cost of increased variance (Heckman et al., 1997). In a special case, when propensity score distribution is the same in the treatment and the control group, the caliper cuts off the worst matched pairs and lowers the bias without significant increase in the 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 estimation of ATT (Smith & Todd, 2005). On the other hand, Smith and Todd (2005) pointed out that the potential problem with caliper is a lack of *a-priori* knowledge about its optimal value. It is common practice to set the value by try and error method.

In Strawiński (2009) the modified caliper mechanism is proposed

$$A_j = \left\{ j \mid j \in \{1..n\} : \min \left\| w_i - w_j \right\| < \delta w_i \right\} \tag{6}$$

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 practice, there is a few such observations but it is very likely that there is good counterfactual state in the control group for them. A large number of matched pairs with low treatment probability could cause ATT estimator to be biased. Therefore, their influence should be limited despite that for those observations it is relatively easy to find a match. In a situation where probability of participation approaches 1 dynamic caliper will have no major differences from standard one. As a result, greater number of matched pairs is left aside in the computation, those with low participation probability.

4. Empirical results

Strawiński (2009) performed simulations and showed that dynamic caliper in some circumstances outperforms standard procedure. We will try to confirm these results and apply to the real empirical data. We use data from the National Supported Work Demonstration (NSW) survey. We choose these data because they were examined on many occasions in similar works (i.e. LaLonde, 1986; Deheija & Wahba, 1999; Smith & Todd, 2005) and their results are often cited. The NSW program took place in mid 1970's in the United States. It focused on providing first working experience. The recipients were disadvantaged economically (young mothers) or socially (with criminal record) on the labour market. Among the applicants program participants were chosen randomly.

The random assignment to the program and the survey after program are important elements from the evaluation process point of view. The original control group were influenced by attrition problem. To reduce attrition aside original sample of two additional group were drawn: one from the Current Population Survey (CSP) and the other from the Panel Study of Income Dynamics (PSID).

Table 1. The average values of covariates

Sample	N	age	educ	Black	hisp	nodg	marr	re74	Re75
NSWre74_T	185	25.82	10.35	0.84	0.06	0.71	0.19	2096	1532
		7.16	2.01	0.36	0.24	0.46	0.39	4887	3219
NSWre74_C	260	25.05	10.09	0.83	0.11	0.83	0.15	2107	1267
		7.06	1.61	0.38	0.31	0.37	0.36	5688	3103
PSID	2490	34.85	12.12	0.25	0.03	0.31	0.87	19429	19063
		10.44	3.08	0.43	0.18	0.46	0.34	13407	13597

Table 1. The average values of covariates (cont.)

Sample	N	age	educ	Black	hisp	nodg	marr	re74	Re75
PSID2	253	36.09	10.77	0.39	0.07	0.49	0.74	11027	7569
		12.08	3.18	0.49	0.25	0.50	0.44	10815	9042
PSID3	128	38.26	10.30	0.45	0.12	0.51	0.70	5567	2611
		12.89	3.18	0.50	0.32	0.50	0.46	7255	5572
CPS	15992	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651
		11.05	2.87	0.26	0.26	0.46	0.45	9570	9270
CPS2	2369	28.25	11.24	0.11	0.08	0.45	0.46	8728	7397
		11.70	2.58	0.32	0.28	0.50	0.50	8968	8112
CPS3	429	28.03	10.24	0.20	0.14	0.60	0.51	5619	2466
		10.79	2.86	0.40	0.35	0.49	0.50	6789	3292

Age = age in years; educ = years of education; marr=1 married, else 0; nodg=1 not finished school, else 0; black = 1 for Afro-American, else 0; hisp=1 for Latino, else 0; re74 = real income in 1974, re75 = real income in 1975, u74 =1 unemployed in 1974, else 0; u75=1 unemployed in 1975, else 0.

Source: Own calculations based on Deheija i Wahba (1999)

The NSW sample is an original survey sample. It consists of two subsets: experimental one (NSWre74_T) and control one (NSWre74_C). In LaLonde (1986) work only those with available income information for 1974 were considered. The author draws two additional control groups, one from the PSID, second from the CPS. The eye examination of the average values of each variable indicates significant differences between the samples in terms of age, ethnicity variables, martial status and salary received before the program. In order to reduce disparities between program participants and members of the control group LaLonde (1986) drew further two control groups from each survey called PSID2, PSID3 and CPS2, CPS3, respectively, with the structure that is more similar to the treatment group.

In the first step of the analysis the experimental results were replicated. This step is necessary to provide comparability with findings of previous researches. In the next step three matching methods and three specifications for functional form of the propensity score, those proposed by Deheija and Wahba (1999), were used. They are presented in Table 2.

Table 2. Propensity score specification

Vector	Covariates
1	Age, age^2, educ, educ^2, marr, nodg, black, hisp, re74, re75, re74^2, re75^2, u74*black
2	Age, age^2, educ, educ^2, marr, nodg, black, hisp, re74, re75, re74^2, re75^2, u74, u75
3	Age, age^2, educ, educ^2, marr, nodg, black, hisp, re74, re75, u74, u75, educ*re74, age^3

Age = age in years; educ = years of education; marr=1 married, else 0; nodg=1 not finished school, else 0; black = 1 for Afro-American, else 0; hisp=1 for Latino, else 0; re74 = real income in 1974, re75 = real income in 1975, u74 =1 unemployed in 1974, else 0; u75=1 unemployed in 1975, else 0.

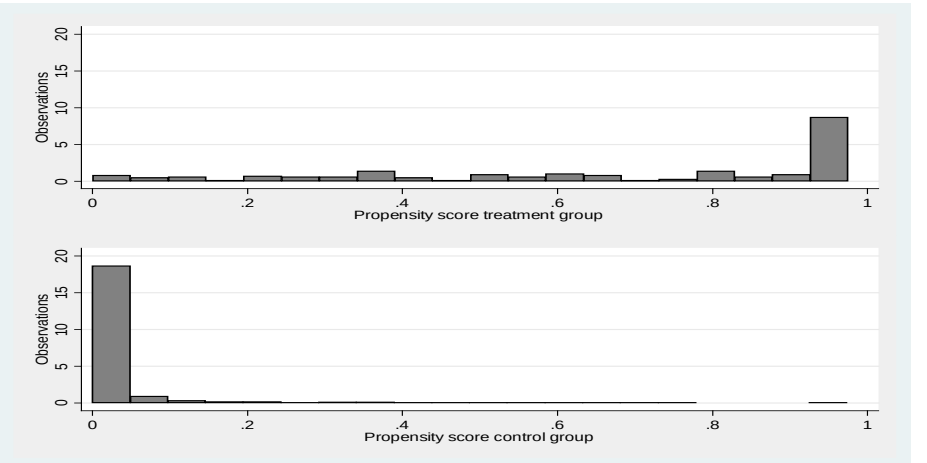
Source: Own calculations based on Deheija i Wahba (1999).

In the next step the propensity scores were estimated for each sample and their distributions were examined. The latter step was to check if there is a potential advantage from dynamic caliper mechanism. The dynamic caliper provides better estimates of the average treatment effect when the distribution of the propensity score values is different in the treatment and the control group. To conserve the space we present graphically the distribution for the first specification only.

Figure 1 presents propensity score distributions for the PSID sample and the first specification of the propensity score vector. For all remaining samples and specifications graphs are similar. There is evident difference in the distribution of the propensity score between treatment and control group. For the treatment group the distribution is skewed to the left and concentrated on (0.95,1) interval, while in the control group it is skewed to the right and concentrated on (0,0.05) interval. This implies major differences between treatment and control group in terms of observed characteristics. In addition, there is a lack of comparison observations for individuals with propensity score value around 0.9. As a result matching without imposing caliper may lead to comparison of incomparable individuals. Therefore, it is advisable to use caliper and dynamic caliper (Strawinski, 2009).

Having in mind the aforementioned properties of the propensity score vector for all specifications three different ATT estimators were used, namely simple one-to-one matching, one-to-one matching with caliper and one-to-one matching with dynamic caliper where a caliper value is a function of estimated propensity score value. The caliper was arbitrary set to 0.005 as this value is usually used in empirical research (Smith & Todd, 2005; Austin. 2009). For models with dynamic caliper mechanism it is assumed that the propensity score for matched comparison unit may differ at most 0.5% from the value of the propensity score for treated unit.

Figure 1. Propensity score distribution in treatment and control group for PSID sample



Top panel present distribution of propensity score in treatment group, bottom panel in control group.

Source: Own calculations in Stata 10 using PSMATCH2.

Table 3. Panel A. The PSID sample

Method/vector		Linear regression	Matching	Caliper Matching	Dynamic caliper matching
treatment	ATT	-15205			
	SE	1155			
Vector 1	ATT	217	1655	1107	1599
	SE	1106	1864	1871	2254
Vector 2	ATT	275	618	-160	595
	SE	1016	1858	1734	2174
Vector 3	ATT	243	1469	1190	1187
	SE	1023	1614	1424	1533

Table 3. Panel B. The CPS sample

Method/vector		Linear regression	Matching	Caliper Matching	Dynamic caliper matching
treatment	ATT	-8498			
	SE	712			
Vector 1	ATT	834	2393	2268	2426
	SE	599	932	923	988
Vector 2	ATT	1199	1991	1603	1355
	SE	555	1013	1009	1066
Vector 3	ATT	1567	1702	1849	1685
	SE	557	1037	1059	1139

ATT stands for the Average Treatment on Treated and SE for standard error.

Source: Own calculations.

The results presented in Table 3 are point estimates of the average treatment effects on the treated. They show a net benefit from taking part in the NSE program. The results include a reference linear regression model and three matching based models. The reference model is a simple regression of earnings in 1978 on treatment dummy. The first column contains the propensity score vector specification. In a seminal LaLonde (1986) article for the experimental sample the estimate of average treatment on treated is 1974 and after correction with linear regression 1962, both numbers statistically significant at 0.01 level. When the original reference group was replaced by the one from another survey the values and signs of the estimates changed dramatically.

When groups are controlled for the distribution of covariates the picture is different (compare results in the first row of each panel with other results). For the PSID control linear regression underestimates the true effect, while matching overestimates. Considering different matching algorithms, the simple matching is closest to the experimental results, and also dynamic caliper results are acceptable, while traditional caliper results are significantly biased. The reason is a vast number weakly matched pairs in the extreme regions of the propensity score distribution. For the CPS control group the picture is slightly different. The linear regression underestimates the effect, while matching overestimates. In case of CPS sample, the use of caliper changes the estimation results. For each propensity score vector specification the result close to experimental one is obtained from caliper or dynamic caliper matching. Therefore, leaving out of the analysis poorly matched pairs is empirically justified. However, one is not able to choose definitely a better method between caliper and dynamic caliper.

In addition, the distance between distributions of the propensity scores for matched pairs were calculated as well as their medians. Usually, when the aim is estimation of the ATT appropriate measure of accuracy is mean squared error or other measures that accounts for bias and variance (D'Orazio et al. 2009, p. 9). In our case we concentrate on reducing the bias. Therefore, we use a median value of difference in propensity scores between treated and control groups. For ideal matching of the treatment and the control group the median of the difference should be equal to zero. The higher median, the worse adjustment in terms of estimated propensity score. The results are presented in Table 4.

Table 4. Matching adjustment quality

Method/vector	Matching	Caliper Matching	Dynamic Caliper Matching
PSID			
vector1	2.4239	1.0216	0.8249
vector2	3.7038	1.0459	0.7542
vector3	3.4756	0.9434	0.5722
CPS			
vector1	0.3820	0.1988	0.1196
vector2	0.3859	0.1891	0.1129
vector3	0.3227	0.2348	0.1191

The numbers in the table are medians of differences in propensity scores between treatment and control group multiplied by 1000.

Source: Own calculations.

Independently from chosen sample and functional form of the propensity score, the worst matched are pairs matched by one-to-one without caliper. Setting the caliper improves the results. The median is approximately halved. The dynamic caliper improves matching further. The gain is significant in both samples. After implementation of dynamic caliper the median of unadjustment is a quarter of that for one-to-one matching without caliper.

The results presented in the Table 4 indicate that previously obtained for the ATT should be reinterpreted. Taking into consideration the ATT and the imbalance measures simultaneously we conclude that the proposed mechanism of matching observations with dynamic caliper outperforms one-to-one matching and one-to-one matching with caliper. In dynamic method, the variance of the ATT estimate is reduced and simultaneously the bias is controlled for. Furthermore, we are able to replicate the experimental results in two dimensions. Firstly, in aspect of covariates balance between treatment and control group. Secondly, in terms of the ATT estimate. The results themselves are very close to experimental and the standard deviation of the estimators is only slightly higher than in the standard procedure.

It is important to note that the results are more sensitive to swap of the reference group than to estimation method. In all but one case, where the CPS group is used and the second specification of propensity score vector, the estimates of the average treatment effect on treated are closest to experimental one when caliper or dynamic caliper is used. Therefore, we conclude that dynamic caliper is an effective tool to find unbiased results. The dynamic caliper reduces the variance of the matching procedure at the cost of small increase in the variance of the treatment estimator.

4. Conclusions

The obtained results indicate that in non-experimental methods balancing the treatment and the control group in terms of covariates is at least as important as correct specification of propensity score vector's functional form. Up to now, the process of balancing tedious work and it has been shown that dynamic caliper can ease the problem, because it controls simultaneously for the bias of ATT estimator and the balance of treatment and control group. However, one has to bear in mind that the best estimation tool will fail when there is a lack of candidates for good counterfactuals.

The aim of the current research was a comparison of dynamic caliper with other estimation techniques based on the propensity score. In the literature rather different functional forms for propensity scores were considered or kernel matching. Therefore, we put more emphasis on the process of pairing observations. It has been shown that the procedure modified by Strawiński (2009) is an efficient tool to replicate experimental results.

Relying on the results obtained in this empirical exercise we conclude that matching with dynamic caliper is an efficient method. The variance of matching procedure measures as median of difference in propensity score is reduced (see Table 4), at the cost of the variance of ATT estimator, they are slightly higher (compare standard errors in column matching and dynamic caliper matching in Table 3). But the latter is negligible.

Acknowledgements

The research is co-founded by Polish Ministry of Science and Higher Education grant N111 109335.

I would like to thank you, anonymous referees, for your useful comments and suggestions.

REFERENCES

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.

BLUNDELL R., COSTA-DIÁS M. (2000) „Evaluation Methods for Non-Experimental Data", Fiscal Studies, vol. 21/4, pp. 427-468.

COCHRANE, RUBIN (1973) „Controling Bias in Observational Studies. A Review", Sankhya, vol. 35, pp. 417-466.

DEHEJIA R., WAHBA S. (1999) „Causal Effects in Nonexperimental studies: Reevaluating the Evaluation of Training Program", Journal of American Statistical Association, vol. 94, no 448.

DEHEJIA R., WAHBA S. (2002) „Propensity score matching methods for nonexperimental causal studies", Journal of the American Statistical Association, vol. 84, pp. 151-161.

HECKMAN J., HOTZ J. (1989) „Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training", Journal of the American Statistical Association, vol. 84(408), pp. 862-880.

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, vol. 64/4, pp. 605-654.

LEE M-J. (2005) „Micro-Econometrics for Policy, Program, and Treatment Effects", Oxford University Press.

ROSENBAUM P., RUBIN D. (1983) „The Central Role of the Propensity Score in Observational Studies for Causal Effects“, Biometrika, vol. 70/1, pp. 41-55.

RUBIN D. (1973) „Matching to Remove Bias in Observational Studies”, Biometrics, vol. 29, pp. 159-183.

SMITH J., TODD P. (2005) „Does Matching Overcome LaLonde’s Critique of nonexperimental estimators?”, Journal of Econometrics, vol. 125, str. 305-353.

STRAWIŃSKI P. (2007) „Causality, selection and engodenuous effects”[Przyczynowość, selekcja i endogeniczne oddziaływanie], Przegląd Statystyczny nr 4/2007, pp. 49-61.

STRAWIŃSKI P. (2009) „Matching with Dynamic Caliper. Preliminary Results”[Łączenie danych z dynamicznym obcięciem. Wyniki wstępne], Metody Ilościowe w Badaniach Ekonomicznych X, pp. 232-242.