ARGUMENTA
OECONOMICA
CRACOVIENSIA

No 15 • 2016
ISSN 1642-168X
AOC, 2016; 15: 57–74
DOI: 10.15678/AOC.2016.1504

Sabina Denkowska

ASSESSING THE ROBUSTNESS TO AN UNOBSERVED CONFOUNDER OF THE AVERAGE TREATMENT EFFECT ON THE TREATED ESTIMATED BY PROPENSITY SCORE MATCHING*

Abstract

One of the serious drawbacks of observational studies is the selection bias caused by the selection process to the treatment group. Propensity Score Matching (PSM), which allows for the reduction of the selection bias when estimating the average treatment effect on the treated (ATT), is a method recommended for the evaluation of projects and programmes co-financed by the European Union. PSM relies on a strong assumption known as the Conditional Independence Assumption (CIA) which implies that selection into the treatment group is based on observable variables, and all variables influencing both the selection process and outcome are observed by the researcher. If this does not hold, the estimated effect may be not so much the result of the treatment as of the lack of balance of an unobserved confounder, which affects both the selection process and the outcome. Rosenbaum's sensitivity analysis allows researchers to determine how strong the impact of such a potential unobserved confounder on selection into treatment and the outcome must be to undermine conclusions about ATT estimated by PSM. Rosenbaum's primal and simultaneous approaches are applied in the paper to assess robustness to an unobserved confounder of the net effect of internships for unemployed young people with a maximum age of thirty-five (estimated with PSM) organized by one of the biggest district employment offices in Małopolska.

Sabina Denkowska, Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: denkowss@uek.krakow.pl

* The author acknowledge the support from research funds, granted to the Faculty of Management of the Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

Keywords: Propensity Score Matching, sensitivity analysis, Rosenbaum's sensitivity analysis, labour market policy.
JEL Classification: C14, E24.

1. Introduction

One of the serious drawbacks of observational studies is the selection bias caused by the selection process to the treatment group. Propensity Score Matching (PSM), which allows for the reduction of the selection bias when estimating the average treatment effect on the treated (ATT), is a method recommended (see EC 2014, pp. 6–7) for the evaluation of projects and programmes co-financed by the European Union. Propensity Score Matching refers to matching control units to treated units based on propensity scores, which are estimated based on observed characteristics. In common with other matching methods, PSM relies on a strong assumption known as the Conditional Independence Assumption (CIA) which "implies that selection is solely based on observable characteristics, and that all variables that influence treatment assignment and potential outcomes simultaneously are observed by the researcher" (Caliendo & Kopeinig 2008). Critics argue that the main disadvantage of studies based on matching units, including the PSM method, is connected with not taking into account an important confounder, which influences both the selection process and the outcome. This objection is relevant not only at the design and data-gathering stages. It may be the case that an unobserved confounder U is unmeasurable or difficult to measure. If a confounder has not been taken into consideration during the matching process, the treatment and control groups may not be comparable. In this way, the estimated effect may not have been caused by the treatment but by the lack of balance[1] introduced by an unobserved confounder U, which affects both outcome and selection[2]. It is for this reason that Rosenbaum (2005, 2010) recommends that observational studies based on matching are complemented by sensitivity analysis, which assesses the robustness of the estimated treatment effect in respect of an unobserved confounder.

The paper applies the primal and simultaneous Rosenbaum approaches to assess the robustness in respect of unobserved confounders of the net

[1] The balancing of variables means the similarity of distributions understood as the lack of systematic differences in their distributions.

[2] This problem is non-existent in experimental studies based on randomization, which balances all observed and unobserved variables. The differences between the values of the outcome variables in experimental and control groups are thus the result of the treatment performed on units in the experimental group.

effect of internships (estimated with PSM) organized by one of the biggest district employment offices in Małopolska for unemployed people with a maximum age of thirty-five.

2. Propensity Score Matching

2.1 Notation, Definitions, Assumptions

Let X denote a vector of observable characteristics and let D denote treatment (exposure) ($\mathrm{D} \in \{0, 1\}$), where $\mathrm{D} = 1$ means that a unit was exposed to treatment and $\mathrm{D} = 0$ means that a unit was not exposed to treatment. For each i-th unit from an N-element population only one of two results for outcome variable Y is possible:

$$Y_i = \mathrm{D} \cdot Y_i^1 + (1 - \mathrm{D}) \cdot Y_i^0 = \begin{cases} Y_i^0, \text{gdy } \mathrm{D} = 0 \\ Y_i^1, \text{gdy } \mathrm{D} = 1 \end{cases}. \quad (1)$$

The usual aim of evaluation studies is to estimate an average treatment effect on the treated (ATT), which makes it possible to decide whether the treatment is effective for treated units:

$$\tau_{ATT} = \mathrm{E}[Y^1 - Y^0 \,|\, \mathrm{D} = 1]. \quad (2)$$

The average treatment effect on the treated can be expressed as the following difference:

$$\tau_{ATT} = (\mathrm{E}[Y^1 \,|\, \mathrm{D} = 1] - \mathrm{E}[Y^0 \,|\, \mathrm{D} = 0]) - (\mathrm{E}[Y^0 \,|\, \mathrm{D} = 1] - \mathrm{E}[Y^0 \,|\, \mathrm{D} = 0]) \quad (3)$$

in which the subtrahend is a selection bias resulting, among others things, from a lack of balance between the observed (and unobserved) variables in a treatment group and a control pool.

Matching in PSM is based on the propensity score p, which is defined as the conditional probability of being treated for a given vector x of observed characteristics X (Rosenbaum & Rubin 1983):

$$p(\mathrm{x}) := P(\mathrm{D} = 1 \,|\, X = \mathrm{x}) = \mathrm{E}(\mathrm{D} \,|\, X = \mathrm{x}). \quad (4)$$

The underlying assumption of PSM is the Conditional Independence Assumption[3] (CIA) that treatment assignment D is independent of potential outcomes (Y^1, Y^0) conditional on a given vector of covariates X (in the notation of Rosenbaum & Rubin 1983):

$$(Y^1, Y^0) \perp \mathrm{D} \,|\, X. \quad (5)$$

[3] Also known as "ignorability" (Rubin 1978), "no hidden bias" or "unconfoundedness" (Caliendo & Kopeinig 2008).

Rosenbaum and Rubin (1983) show that if potential outcomes are independent of treatment conditional on vector x of covariates X, they are also independent of treatment conditional on the propensity score $p(x)$. That CIA is untestable and, moreover, easy to undermine in observational studies, may mean that questions are raised about the results obtained using PSM.

The second assumption of PSM is the common support assumption, which is also known as the overlap assumption (Caliendo & Kopeinig 2008). It can be written as follows:

$$0 < P(D = 1 | X = x) < 1 \text{ for all x in support of } X. \quad (6)$$

This means that each unit with the same vector x of observed characteristics X has some chance of being treated and some chance of not being treated.

Unconfoundedness and the overlap assumption both constitute a property known as the "strong ignorability of assignment", which is necessary[4] to identify the treatment effect (Rosenbaum & Rubin 1983).

2.2 Algorithm for Propensity Score Matching

In practice, propensity scores are usually estimated as the fitted probabilities of treatment derived from the estimated logistic model, in which treatment status is regressed on observed baseline characteristics X. All of the variables simultaneously influencing the selection process and the outcome should be included[5] in the estimated logistic model (Stuart 2010). In the case of PSM, the model is only a means to achieve the goal of balancing the variables. For this reason, attention should be focused on the model's capacity to balance the variables rather than on estimating its parameters (Caliendo & Kopeinig 2008, Stuart 2010). Next, a method for matching[6] the control group to the treatment group (on the basis of estimated propensity scores) is selected[7]. Because the effects of treatment should not

[4] For ATT, however, Heckman et al. (1998) show that the unconfoundedness assumption can be weakened to conditional mean independence (see also Abadie & Imbens 2006). The overlap assumption can also be weakened because the right inequality in formula (6) is a sufficient condition for identifying ATT (Caliendo & Kopeinig 2008, Strawiński 2014).

[5] To satisfy the assumption of conditional independence (Rubin & Thomas 1996).

[6] For details of methods for matching the control group and the different ways they can be applied (with or without replacement, 1:k matching, caliper and radius) see Caliendo and Kopeinig (2008) and Stuart (2010).

[7] The Nearest Neighbour Method with 1:1 matching is the commonest method employed in evaluations of the Polish labour market (Wiśniewski & Maksim 2013, Konarski & Kotnarowski 2007, Trzciński 2009).

be assessed unless the matching is satisfactory, the latter is evaluated by checking and, where necessary, by determining a region of common support and checking the balance of variables included in the estimated logistic model. More information about determining the region of common support and about the indicators and tests used for assessing the balance of variables is available in Caliendo and Kopeinig (2008), Stuart (2010), Strawiński (2008, 2014) and Denkowska (2015). If the balance of variables is found to be unsatisfactory, researchers should consider other methods for matching or return to estimating the model of logistic regression and introduce two-way interactions and/or higher-order terms (Stuart 2010, Caliendo & Kopeinig 2008). Unfortunately, the tedious process of searching for the model and the best matching method that will balance all of the variables, higher-order terms and interactions from the estimated logistic model will not necessarily be successful. This may mean that the CIA has failed (Smith & Todd 2005). If this is the case, alternative approaches to evaluation should be considered (Caliendo & Kopeinig 2008).

The estimation of the treatment effect should not be conducted until all of the variables, higher-order terms and interactions of covariates used in the model have been satisfactorily balanced (Rubin 2001).

3. Sensitivity Analysis with Rosenbaum's Approaches

3.1. General Remarks

Sensitivity analysis has been proposed to indicate the magnitude of a hidden bias that should be present to alter the conclusions of the study. The robustness of average treatment effects estimated with matching methods can be assessed with Rosenbaum's approaches. Gastwirth, Krieger and Rosenbaum (1998) distinguish primal, dual and simultaneous approaches, which differ with regard to finding the thresholds of the association between the unobserved confounder and the exposure and/or the outcome that would render the test statistics of the study inference insignificant (Liu, Kuramoto & Stuart 2013). In the primal approach, the sensitivity parameter Γ relates an unobserved confounder U to treatment D, while assuming that confounder U is a perfect predictor of the outcome. In the dual approach, sensitivity parameter Δ relates an unobserved confounder U to outcome Y, while assuming that a confounder U is a perfect predictor of the treatment. Simultaneous sensitivity analysis uses both sensitivity parameters: Γ and Δ. The primal and simultaneous procedures are the most important from a practical point of view.

Rosenbaum's sensitivity analyses assume that matching is performed without replacement.

3.2. Rosenbaum's Primal Approach

Assuming that an unobserved confounder is a perfect predictor of the outcome, the question to be answered in the primal approach is how strong its association with treatment has to be to change the conclusions of the study.

Let us assume that there is an unobserved covariate U ($U \in <0; 1>$).

In matching methods we assume that a matched pair of units $\boldsymbol{k}$ and $\boldsymbol{l}$ with the same characteristics X ($\mathbf{x}_k = \mathbf{x}_l$) have the same probability of receiving treatment ($\pi_k = \pi_l$). But because of a potential unobserved confounder U, the odds that unit $\boldsymbol{k}$ receives treatment *de facto* are:

$$\frac{\pi_k}{1-\pi_k} = \exp(\kappa(\mathbf{x}_k) + \gamma u_k), \text{ where } 0 \le u_k \le 1. \tag{7}$$

So, for two units $\boldsymbol{k}$ and $\boldsymbol{l}$ with the same characteristics X ($\mathbf{x}_k = \mathbf{x}_l$) the odds ratio is:

$$\frac{\frac{\pi_k}{1-\pi_k}}{\frac{\pi_l}{1-\pi_l}} = \frac{\exp(\kappa(\mathbf{x}_k) + \gamma u_k)}{\exp(\kappa(\mathbf{x}_l) + \gamma u_l)} = \exp(\gamma(u_k - u_l)). \tag{8}$$

Rosenbaum (2002) shows the following bounds on the odds ratio:

$$\frac{1}{\exp(\gamma)} \le \frac{\frac{\pi_k}{1-\pi_k}}{\frac{\pi_l}{1-\pi_l}} \le \exp(\gamma). \tag{9}$$

Let $\Gamma := exp(\gamma)$. The units with the same values of observed covariates may nonetheless differ in terms of an unobserved confounder, so that one unit has odds of treatment that are up to $\Gamma \ge 1$ times greater than the odds for another unit" (Rosenbaum 2002, 2005).

Sensitivity analysis to an unobserved confounder proposed by Rosenbaum (2002) is based on several different randomisation tests (Liu, Kuramoto & Stuart 2013, Keele 2010). For a binary[8] outcome variable Y, the sensitivity analysis is based on McNemar's test, which is used to verify whether the confounder U has a significant impact on the result of the outcome variable Y. Information on paired units is presented in a contingency table (2×2).

[8] For other outcomes, the sensitivity test is based on the Wilcoxon sign rank test and the Hodges--Lehmann point estimate for the sign rank test (Rosenbaum 2005, Keele 2010).

The chances of being selected for a treated group are theoretically the same for units paired based on propensity score. When we employ Rosenbaum's sensitivity analysis we are seeking the odds ratio of treatment of the paired units (occurring due to unobserved confounder U) that would change the conclusions of the study in such a way as to render it insignificant.

Let T denote the number of all pairs in which the results of the outcome variable Y differ, and let a denote the number of pairs in which a treated unit has a positive result for the outcome variable and a not-treated unit has a negative result. The lower and upper bounds on the p-value are calculated by analogy with the binomial test p-value:

$$p_{lower} = \sum_{i=a}^{T} \binom{T}{i} (p^{-})^{i} (1-p^{-})^{T-i} \text{ and } p_{upper} = \sum_{i=a}^{T} \binom{T}{i} (p^{+})^{i} (1-p^{+})^{T-i}, \quad (10)$$

where the probabilities:

$$p^{-} = \frac{1}{1+\Gamma} \text{ and } p^{+} = \frac{\Gamma}{1+\Gamma} \quad (11)$$

are lower and upper bounds on the probability of being treated and are determined for different, hypothetical values of Γ. The lower bound p_{lower} is always lower than the observed p-value and, thereby, less important and rarely taken into account. Calculations are repeated with different values of Γ to find the value of parameter Γ in which p_{upper} becomes greater than 0.05.

3.3. Rosenbaum's Simultaneous Approach

By analogy with the primal analysis (Gastwirth, Krieger & Rosenbaum 1998), the upper bound on the p-value, p_{upper} in formula (10) is calculated in Rosenbaum's simultaneous approach with (Liu, Kuramoto & Stuart 2013):

$$p^{+} = p(treated) \cdot p(outcome) + (1 - p(treated)) \cdot (1 - p(outcome)), \quad (12)$$

where

$$p(treated) = \frac{\Gamma}{1+\Gamma}, \quad (13)$$

$$p(outcome) = \frac{\Delta}{1+\Delta} \quad (14)$$

are determined for different, hypothetical values of Γ and Δ.

A combination of values of Γ and Δ for which $p_{upper} \geq 0.05$, is the point at which the result is sensitive to an unobserved confounder U (Liu, Kuramoto & Stuart 2013).

4. Application of Rosenbaum's Sensitivity Analysis to the Study on the Net Effect of Internships

The net effect of the internships for unemployed people with a maximum age of thirty-five organized by the Tarnów District Employment Office was estimated. The purpose was to gauge the general effectiveness of internships organized by district employment offices in activating people who are young and unemployed[9]. The study was conducted using PSM. Rosenbaum's primal and simultaneous sensitivity analyses were applied to the estimated effect to check its robustness to a potential unobserved confounder influencing both the inclusion to the group of interns and finding a job.

In 2013, 1,409 unemployed people with a maximum age of thirty--five began internships. They were completed at least three months before 10 August 2014. The data were obtained from the Syriusz computer system, which is used to register unemployment.

The *X* variables employed in the study can be divided into four categories[10]:

I. Socio-demographic and health variables: *plec* – sex, *wiek* – age in years, *s_w* – single parenthood, *n_p* – disability, education (*w_brak* – lack, *w_sp* – elementary, *w_gim* – junior high school, *w_zaw* – vocational, *w_sr* – high school, *w_pm* – post-high school, *w_w* – university).

II. Employment, educational activity and activity on the labour market: job – classification[11] (*gr00* – lack, *grX* – where *X* denotes the classification number), *staz_pr* – number of years in employment, *dl_bzr* – long-term unemployment (Yes/No), *szk* – training during the two years preceding the internship (Yes/No), *l_prop* – number of job offers during the last six months, *w_a* – indicator of activity (community work, intervention jobs, training, internships, public work) in the two years preceding the internship: 0 – no active days, 1 – up to 100 active days, 2 – up to 200 active days. And so forth.

[9] A study of the net effect of the internships for all unemployed people – regardless of age – can be found in Denkowska (2015, 2016).

[10] The preliminary selection of variables was based on the experience gained from the Alternatywa II project, which the team evaluated using PSM. The project was part of the latest edition of Phare SSG RZL 2003 (Trzciński 2009). The source of data was SI PULS. After consulting the employees of the District Employment Office, however, it became clear that – due to the limitations of the Syriusz system – not all of the features could and should be used in the study. The paper describes the final set of variables used in the study.

[11] Classification in accordance with the ordinance of the Minister of Labour and Social Policy of 27 April 2010 on the Classification of Occupations and Specializations for Labour Market Needs.

III. Relative motivation to look for a job: *pr_zas* – eligibility for unemployment benefit.

IV. Skills and abilities: *pr_B* – driving licence, category B, *angBG* – at least a good knowledge of English, *angSL* – basic knowledge of English, *j_n* – knowledge of German.

The outcome variable *Y* was employment three months after finishing the internship. It was assumed[12] that a person not registered on the day the data were checked was employed.

The control pool consisted of 11,568 young people with a maximum age of thirty-five who had not been involved in an activation in 2013. To establish the values of variables *X* and outcome variable *Y*, for each person from the control pool the date of "starting internship" was randomly selected (measuring the values of the *X* variables). The average duration of the internship was added next. A check of whether or not the person was registered in the database was performed three months after the date the internship was "completed" (the value of variable *Y*).

The logistic regression model, in which the dependent variable was participation in the internship, was estimated first. There followed numerous attempts to obtain the best possible balance of variables, including modifying the regression model by introducing interactions and squares of variables, and checking various matching methods without replacement[13]. The distributions of propensity scores in the group of interns and the control group were analysed to check the region of common support, which influenced the decision to use a matching method with a caliper. The balance of variables, interactions and squares of variables were checked using standardized mean difference, and with *t*-tests for means in the interns group and in the control group, each time before and after matching. In the case of continuous and discreet variables, the similarity of distributions in the interns group and in the control group was verified using the bootstrap KS test[14].

The best balance of variables was obtained for the logistic model to which interactions and the $wiek^2$ variable were introduced (Table 1). The Nearest Neighbour Method used in the study (1:1, without replacement and with

[12] According to the methodology used by WUP (the regional employment office) in Kraków.

[13] Rosenbaum's sensitivity analysis can be applied only to matching methods without replacement.

[14] A bootstrap version of the Kolmogorov-Smirnov test can be used for both continuous and discreet random variables (Abadie 2002, Sekhon 2011).

a caliper[15] (*caliper* = 0.5)) led to the removal of two interns, for whom there were no good matches in the control group.

Table 1 presents the standardized mean differences[16] obtained from the formulas:

$$SDiff_{before} = \frac{\overline{X}_T - \overline{X}_{CP}}{S_T} \cdot 100\%,\ SDiff_{after} = \frac{\overline{X}_{TM} - \overline{X}_{CM}}{S_T} \cdot 100\%, \qquad (15)$$

where: $\overline{X}_T, \overline{X}_{CP}$ denote the means in the treatment (interns) group and in the control pool before matching, $\overline{X}_{TM}, \overline{X}_{CM}$ denote the means in the treatment (interns) group and in the control group after matching, while S_T stands for standard deviation in the treatment (interns) group before matching. The analysis of standardized differences is based on checking whether the values of standardized differences for all variables (per modulus) decrease after matching, and whether the values obtained after matching can be considered satisfactory. A standardized mean difference of at or below 3%, or at or below 5%, is considered sufficient in the majority of empirical studies (Caliendo & Kopeinig 2008).

Table 1 presents the standardized differences before and after matching, the *p*-values from the *t*-tests for the means of all variables, the interactions and $wiek^2$ variable, and the *p*-values from the KS bootstrap test[17] for all continuous and discreet variables and interactions.

All of the standardized differences decreased after matching, and none exceeded (per modulus) 4.3%. The *t*-tests did not reveal significant differences between the means. The Smirnov-Kolmogorov bootstrap test "confirmed" that the distributions for the continuous and discreet variables were similar.

After establishing that all of the variables, interactions and the $wiek^2$ variable were balanced, the net effect of the internships was estimated as the difference between the percentage of employed interns and the percentage of employed persons in the control group. The net effect of the internships for unemployed people with a maximum age of thirty-five was 10.945% with a standard error (see Imbens & Ambadie 2006) of 1.87% (p = 5.1905e – 09).

[15] Rubin and Thomas (1996) recommend keeping the limit at 0.25*s* or 0.5*s* where: $s = \sqrt{\frac{S_T^2 + S_{CP}^2}{2}}$, and S_T^2 and S_{CP}^2 denote variance in the treatment group and in the control pool respectively. It is worth noting that some participants from the treatment group may remain unmatched.

[16] The dichotomous variables were treated as continuous variables and standardized mean differences were obtained from the same formulas (Stuart 2010).

[17] A bootstrap version of the Kolmogorov-Smirnov test makes it possible to test the similarity of the distributions of both continuous and discreet random variables (Abadie 2002, Sekhon 2011).

Table 1. Standardized Mean Differences, *P*-values from *T*-tests for Means, and *P*-values from the Kolmogorov-Smirnov Bootstrap Test Before and After Matching

Variable	Before			After		
	$SDiff_{before}$	Test *t* *p*-value	KS bootstrap *p*-value	$SDiff_{after}$	Test *t* *p*-value	KS bootstrap *p*-value
Plec	–41.907	< 2.22e – 16	–	–2.1782	0.49353	–
wiek	–30.870	< 2.22e – 16	< 2.22e – 16	–2.2438	0.52446	0.556
s_w	–18.797	2.476e – 10	–	2.6609	0.45812	–
n_p	–6.4606	0.02512	–	0.46946	0.89976	–
w_sp	–38.026	< 2.22e – 16	–	–3.9018	0.30348	–
w_gim	–45.542	< 2.22e – 16	–	2.8634	0.41111	–
w_zaw	–70.761	< 2.22e – 16	–	–2.298	0.43519	–
w_sr	9.5860	0.0006772	–	2.7338	0.37518	–
w_pm	8.7961	0.001535	–	–0.87107	0.81645	–
w_w	46.639	< 2.22e – 16	–	–0.86959	0.75514	–
gr00	1.6127	0.56733	–	1.2729	0.69139	–
gr1	–0,5808	0.8388	–	0.0000	1.00000	–
gr2	45.166	< 2.22e – 16	–	–1.4605	0.60023	–
gr3	1.0624	0.7062	–	–0.21171	0.95084	–
gr5	–18,362	2.5091e – 10	–	2.7545	0.44239	–
gr6	–10.601	0.00058511	–	2.3879	0.47954	–
gr7	–59.703	2.22e – 16	–	1.5431	0.61887	–
gr8	–17.722	4.8343e – 09	–	–1.194	0.73892	–
gr9	–20.829	1.5147e – 11	–	–1.2311	0.74564	–
staz_pr	–47.266	< 2.22e – 16	< 2.22e – 16	–1.726	0.59530	0.01*
dl_bzr	–3.5778	0.20556	–	3.0175	0.38321	–
l_prop	9.3155	0.00086544	< 2.22e – 16	–1.9491	0.57823	0.796
pr_zas	–5.1164	0.074724	–	0.0000	1.00000	–
w_a	–3.9585	0.1599	0.056	0.0000	1.00000	0.876
szk	7.0995	0.0098951	–	–1.9673	0.58626	–
pr_B	31.148	< 2.22e – 16	–	–3.4191	0.26518	–
angBG	35.752	< 2.22e – 16	–	0.89826	0.76306	–
angSL	8.3716	0.0030145	–	–4.2772	0.19080	–
j_niem	15.494	3.6026e – 08	–	2.5881	0.42853	–
$wiek^2$	–33.603	< 2.22e – 16	< 2.22e – 16	–2.3053	0.51359	0.85997
*gr1*plec*	1.6906	0.53474		0.0000	1.00000	–
*gr00*plec*	–13.428	3.2142e – 06	–	3.1465	0.31151	–
*gr00*w_a*	2.9061	0.29779	0.308	2.5506	0.44051	0.628
*gr00*w_sp*	–21.093	3.54e – 11	–	–0.84576	0.81858	–
*wiek*d_bezr*	6.5890	0.020431	0.004	2.6035	0.45437	0.766
*plec*s_w*	–20.003	2.3416e – 09	–	–2.6688	0.52713	–
*dl_bzr*w_sp*	–20.269	0.29959	–	0.0000	1.00000	–
*n_p*w_zaw*	–6.1322	0.037246	–	1.8898	0.59303	–
*gr3*s_w*	–7.5494	0.012138	–	–1.194	0.73892	–
*gr3*w_w*	3.5677	0.19516		0.89119	0.79629	–
*gr00*wiek*	–2.2709	0.42444	< 2.22e – 16	1.2302	0.70606	0.95
*gr2*w_w*	45.344	< 2.22e – 16		–1.462	0.59719	
*gr3*l_prop*	5.9199	0.032278	0.03	2.5361	0.47542	0.7
*gr5*staz_pr*	–24.615	6.2172e – 15	< 2.22e – 16	–1.579	0.65040	0.25
*gr1*pr_zas*	2.0151	0.45614	–	0.0000	1.00000	–

Table 1 cnt'd

Variable	Before			After		
	$SDiff_{before}$	Test t p-value	KS bootstrap p-value	$SDiff_{after}$	Test t p-value	KS bootstrap p-value
*w_w*gr3*	3.5677	0.19516	–	0.89119	0.79629	–
*gr5*staz_pr*	–24.615	6.2172e – 15	< 2.22e – 16	–1.579	0.65040	0.25

Note: * variable *st_pr* is a continuous variable, so the similarity of distributions was also verified with the classic Kolmogorov-Smirnov test, which failed to reject the null hypothesis about the similarity of distributions after matching ($p = 0.11908$).

Source: author's own calculations in Matching package in R.

Despite the enormous effort invested in making exhaustive use of the information in the Syriusz system, doubt arose as to whether the causality observed between participation in internships and employment was not in fact caused by an unobserved confounder. We may be almost certain, for example, that certain personality features, such as entrepreneurship or communication skills, have a strong impact on employment. The decisive question here is how strong the impact of the unobserved factor on the selection process and employment should be to render the results statistically insignificant.

In order to conduct sensitivity analysis using Rosenbaum's primal approach, the information for 1407 pairs is presented in the contingency table[18] (Table 2). The number of pairs in which the results of outcome variable Y were different to each other was 712 ($T = 433 + 279$), and the number of pairs in which only interns were employed was 433 (a).

Table 2. Contingency Table for Paired Individuals

Group		Interns		Sum
		Employment	Lack	
Control group	Employment	431	279	710
	Lack	433	264	697
Sum		864	543	1407

Source: author's own calculations in Matching package in R.

During the next stage, for hypothetical values of Γ, probabilities p^- and p^+ were calculated, which were then used to obtain lower bounds and upper

[18] We may note, incidentally, that the odds ratio is 1.562, which means that for unemployed people with a maximum age of thirty-five, the odds for getting a job are 1.562 times greater for interns than for non-interns. In other words, the internship increases the odds for securing employment 1.562 times.

bounds for *p*-value according to formulas (10) and (11). The results of the calculations for selected values of Γ are presented in Table 3.

Table 3 informs us that the largest value (to two decimal places) of parameter Γ for which the probability p_{upper} was lower than 0.05, was 1.36. This means that the odds of one person in a pair becoming an intern can be 1.36 times greater than those of the other person in a pair because of different values for the confounder *U*, which has a powerful influence on employment, but there is still strong evidence that internships have an impact on employment ($p = 0.04578$). On the other hand, when $\Gamma = 1.37$, the relationship between internships and employment is no longer significant ($p = 0.05581$). $\Gamma = 1.36$ indicates[19] moderate robustness to the occurrence of an unobserved variable *U*.

Table 3. Bounds for Selected Values of Γ

Gamma	Probability	
	p_{lower}	p_{upper}
1.00	0.0000	0.00000
1.10	0.0000	0.00000
1.15	0.0000	0.00005
1.20	0.0000	0.00042
1.30	0.0000	0.01128
1.35	0.0000	0.03719
1.36	**0.0000**	**0.04578**
1.37	**0.0000**	**0.05581**
1.38	0.0000	0.06740
1.39	0.0000	0.08066
1.40	0.0000	0.09568
1.50	0.0000	0.34337

Source: author's own calculations in Matching package in R.

The results obtained (Table 3) were confirmed by an analysis conducted using the rbounds R package (Keele 2010). In this package Rosenbaum's primal approach is available for binary, ordinal and continuous variables for the matching variant 1:*k* (Keele 2014).

In Rosenbaum's simultaneous approach we look for the smallest values[20] of parameters Γ and Δ for which $p_{upper} \geq 0.05$ (calculated from formulas (10), (12)–(14)). We thus obtain points (Γ, Δ), at which the result is sensitive to

[19] Values of Γ (in the primal version of Rosenbaum's approach) in the social sciences are usually from 1 to 2 (Keele 2010).

[20] To one decimal place or two decimal places.

an unobserved confounder *U* (Liu, Kuramoto & Stuart 2013). The results of Rosenbaum's simultaneous analysis are presented in Table 4.

We are informed by the results in Table 4 that for $\Delta = \Gamma = 2.25$, p_{upper} is 0.035. This means that one person in a pair may be 2.25 times more likely to become an intern, and 2.25 times more likely to gain employment, than the other because they have different values of *U*. Yet there remains strong evidence that internships have an impact on employment ($p = 0.035$). Given $\Delta = \Gamma = 2.3$, on the other hand, the association between internships and employment would no longer be significant ($p = 0.0530$).

Table 4. Results (p_{upper}) of the Simultaneous Approach for Different Values of Γ and Δ

		Δ							
		1.0	1.5	2.0	2.25	2.3	2.5	3.0	+∞
Γ	1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1.5	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0010	0.3434
	2.0	0.0000	0.0000	0.0026	0.0096	0.0120	**0.0257**	**0.0957**	0.9995
	2.25	0.0000	0.0001	0.0096	**0.0350**	0.0432	0.0876	0.2701	1.0000
	2.3	0.0000	0.0001	0.0120	0.0432	**0.0530**	0.1060	0.3136	1.0000
	2.5	0.0000	0.0003	**0.0257**	0.0876	0.1060	0.1986	0.4945	1.0000
	3.0	0.0000	0.0010	**0.0957**	0.2701	0.3136	0.4945	0.8334	1.0000
	+∞	0.0000	0.3434	0.9995	1.0000	1.0000	1.0000	1.0000	1.0000

Source: author's own calculations in R.

Analysis of Table 4 also leads to the conclusion that, for example, one person in a pair may be twice as likely to become an intern and 2.5 times more likely to gain employment than the other because of different values of *U*. However, internships have a significant impact on employment ($p = 0.0257$). Were we to have $\Gamma = 2$ and $\Delta = 3$, though, the causality between internships and employment would no longer be significant ($p = 0.0957$). By analogy, furthermore, one person in a pair may be 2.5 times more likely to become an intern, and twice as likely to secure employment than the other because they have different values of *U*, but there is still strong evidence that internships have an impact on employment ($p = 0.0257$). Were we to have $\Gamma = 3$ and $\Delta = 2$, on the other hand, the causality between internships and employment would no longer be significant ($p = 0.0957$).

The primal approach provides a more sensitive indication than the simultaneous approach because it assumes a perfect relationship between

Table 5. Results (p_{upper}) of the Simultaneous Approach for Different Values of Γ and Δ

		Δ								
		1.0	1.3	1.36	1.37	1.4	1.5	2.0	2.5	+∞
Γ	1.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0000**
	1.3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0113**
	1.36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0458**
	1.37	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0558**
	1.4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	**0.0957**
	1.5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	**0.3434**
	2.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0026	0.0257	**0.9995**
	2.5	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0257	0.1986	**1.0000**
	+∞	**0.0000**	**0.0113**	**0.0458**	**0.0558**	**0.0957**	**0.3434**	**0.9995**	**1.0000**	**1.0000**

Source: author's own calculations in R.

the unobserved variable *U* and the outcome variable *Y*. The results presented in Table 5 demonstrate that the primal analysis of sensitivity is a particular case of the simultaneous approach. The probabilities (bold) for different values of Γ when $\Delta \to +\infty$ (or values of Δ when $\Gamma \to +\infty$) are the same as those in Table 3.

5. Conclusions

Because researchers conducting observational studies can never be sure that all confounders have been taken into account, sensitivity analysis is very important. Rosenbaum (2005, 2010) recommends a two-stage procedure for studies of this kind. What may be termed "classical" matching, which involves propensity scores estimated based on observed characteristics, should always be complemented with sensitivity analysis to asses robustness to a potential unobserved confounder. This practice will help increase confidence in the results obtained in observational studies. Higher values of Γ and Δ indicate robustness of the estimated effect to an unobserved confounder, while smaller values tell us that the result is sensitive to deviations from unconfoundedness, and remind us to proceed with caution in our interpretation.

The paper has set out the results of an empirical application of Rosenbaums's primal and simultaneous sensitivity analyses to the net effect of internships (estimated with PSM) for unemployed young people

with a maximum age of thirty-five organized by one of the biggest district employment offices in Małopolska. It is unfortunate that PSM-based analyses of the labour market in Poland (see e.g. Wiśniewski & Maksim 2013, Konarski & Kotnarowski 2007, Trzciński 2009) have not been complemented by sensitivity analyses. Had they been performed, it would have been possible to relate the results of this study to other, similar studies. This is not, however, a signal to abandon analyses of the robustness of the estimated results. Quite the contrary. Robustness analysis should be incorporated as an important element of all observational studies. If decision-makers are armed with knowledge of the robustness of the estimated results, they are better equipped to draw conclusions from these studies.

Bibliography

Abadie, A. (2002) "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models". *Journal of the American Statistical Association* 97 (457): 284–92, https://doi.org/10.1198/016214502753479419.

Abadie, A. and Imbens, G. W. (2006) "Large Sample Properties of Matching Estimators for Average Treatment Effects". *Econometrica* 74 (1): 235–67, https://doi.org/10.1111/j.1468-0262.2006.00655.x.

Caliendo, M. and Kopeinig, S. (2008) "Some Practical Guidance for the Implementation of Propensity Score Matching". *Journal of Economic Surveys* 22 (1): 31–72, https://doi.org/10.1111/j.1467-6419.2007.00527.x.

Denkowska, S. (2015) "Wybrane metody oceny jakości dopasowania w Propensity Score Matching" [Selected quality assessment methods in propensity score matching] in K. Jajuga and M. Walesiak (eds) *Taksonomia 24. Klasyfikacja i analiza danych – teoria i zastosowania* [Taxonomy 24. Classification and analysis of data – theory and application]. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu No. 384. Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, pp. 60–74.

Denkowska, S. (2016) "Zastosowanie analizy wrażliwości do oceny wpływu nieobserwowanej zmiennej w Propensity Score Matching" [The application of sensitivity analysis in assessing the impact of an unobserved confounder in propensity score matching] in K. Jajuga and M. Walesiak (eds) *Taksonomia 27. Klasyfikacja i analiza danych – teoria i zastosowania* [Taxonomy 27. Classification and analysis of data – theory and application]. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu No. 427. Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, pp. 66–75.

EC (2014) *The Programming Period 2014–2020. Guidance Document on Monitoring and Evaluation. European Regional Development Fund and Cohesion Fund. Concepts and Recommendations*, March. Bruxelles: European Commission.

Gastwirth, J., Krieger, A. and Rosenbaum, P. (1998) "Dual and Simultaneous Sensitivity Analysis for Matched Pairs". *Biometrika* 85 (4): 907–20, https://doi.org/10.1093/biomet/85.4.907.

Heckman, J. J., Ichimura, H., Smith, J. and Todd, P. (1998) "Characterizing Selection Bias Using Experimental Data". *Econometrica* 66: 1017–98.

Keele, L. (2010) *An Overview of Rebounds: An R Package for Rosenbaum Bounds Sensitivity Analysis with Matched Data*, personal.psu.edu/ljk20/rbounds%20vignette.pdf. Accessed: 15 January 2016.

Keele, L. (2014) *Package 'Rbounds'*, https://cran.r-project.org/web/packages/rbounds/rbounds.pdf. Accessed: 15 January 2016.

Konarski, R. and Kotnarowski, M. (2007) "Zastosowanie metody propensity score matching w ewaluacji ex-post" [The use of propensity score matching in ex-post evaluations" in A. Huber (ed.) *Ewaluacja ex-post. Teoria i praktyka badawcza* [Ex-post evaluation. Theory and research practice]. Warszawa: PARP.

Liu, W., Kuramoto, S. K. and Stuart, E. A. (2013) "An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research". *Prevention Science* 14 (6): 570–80, https://doi.org/10.1007/s11121-012-0339-5.

Rosenbaum, P. R. and Rubin, D. B. (1983) "The Central Role of Propensity Score in Observational Studies for Casual Effects". *Biometrika* 70 (1): 41–55, https://doi.org/10.1093/biomet/70.1.41.

Rosenbaum, P. R. (2002) *Observational Studies*. New York: Springer.

Rosenbaum, P. R. (2005) "Observational Study" in B. S. Everitt and D. C. Howell (eds) *Encyclopedia of Statistics in Behavioral Science*, vol. 3. New York: John Wiley & Sons.

Rosenbaum, P. R. (2010) *Design of Observational Studies*. New York: Springer.

Rubin, D. (1978) "Bayesian Inference for Causal Effects: The Role of Randomization". *The Annals of Statistics* 6 (1): 34–58, https://doi.org/10.1214/aos/1176344064.

Rubin, D. (2001) "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation". *Health Services & Outcomes Research Methodology* 2: 169–88.

Rubin, D. and Thomas, N. (1996) "Matching Using Estimated Propensity Scores: Relating Theory to Practice". *Biometrics* 52 (1): 249–64, https://doi.org/10.2307/2533160.

Sekhon, J. S. (2011) "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R". *Journal of Statistical Software* 42 (7): 1–52, https://doi.org/10.18637/jss.v042.i07.

Smith, J. and Todd, P. (2005) "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?". *Journal of Econometrics* 125 (1–2): 305–53, https://doi.org/10.1016/j.jeconom.2004.04.011.

Strawiński, P. (2008) "Quasi-eksperymentalne metody ewaluacji" [Quasi-experimental evaluation methods] in A. Haber (ed.) *Środowisko i warsztat ewaluacji* [The environment and technique of evaluation]. Warszawa: PARP, pp. 193–220.

Strawiński, P. (2014) *Propensity Score Matching. Własności małopróbkowe* [Propensity score matching. Small sample properties]. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.

Stuart, E. A. (2010) "Matching Methods for Causal Inference: A Review and a Look Forward". *Statistical Science* 25 (1): 1–21, https://doi.org/10.1214/09-sts313.

Trzciński, R. (2009) *Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych* [The use of propensity score matching in evaluation research]. Warszawa: PARP, http://www.parp.gov.pl/index/more/13335. Accessed: December 2014.

Wiśniewski, Z. and Maksim, M. (2013) "Polityka rynku pracy w Polsce – wyniki badań ewaluacyjnych prowadzonych za pomocą metody propensity score matching" [Labour market policy in Poland – the results of evaluation research carried out using propensity score matching] in *Rola funduszy unijnych w rozwoju społeczno-gospodarczym regionu* [The role of EU funds in the region's socio-economic development], Zeszyty Naukowe Uniwersytetu Szczecińskiego No. 753. Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, pp. 93–110.

Abstract

Ocena odporności na występowanie nieobserwowanej zmiennej przeciętnego efektu oddziaływania na jednostki poddane oddziaływaniu oszacowanego za pomocą Propensity Score Matching

Jedną z poważnych wad badań obserwacyjnych jest obciążenie selekcyjne spowodowane selekcją jednostek do grupy poddawanej oddziaływaniu. Metoda Propensity Score Matching (PSM), która umożliwia redukcję obciążenia selekcyjnego podczas szacowania przeciętnego efektu odziaływania na jednostki poddane oddziaływaniu (ATT), jest metodą coraz częściej zalecaną przy ewaluacji projektów oraz programów współfinansowanych przez Unię Europejską. PSM opiera się na mocnym założeniu, zwanym założeniem warunkowej niezależności (CIA), które implikuje, że selekcja do grupy poddawanej oddziaływaniu musi być oparta wyłącznie na zmiennych obserwowanych i że wszystkie zmienne wpływające na poddanie oddziaływaniu oraz na potencjalne wyniki zmiennej wyjściowej są obserwowane przez badacza. Jeżeli założenie to nie jest spełnione, to oszacowany efekt może być nie tyle wynikiem oddziaływania, co skutkiem braku zbalansowania nieuwzględnionej (nieobserwowanej) w badaniu zmiennej, która wpływa zarówno na proces selekcji, jak i zmienną wyjściową. Analiza wrażliwości Rosenbauma umożliwia badaczom ocenę, jak silny musiałby być wpływ takiej potencjalnej nieobserwowanej zmiennej na proces selekcji oraz na zmienną wyjściową, aby podważyć wnioski na temat efektu ATT oszacowanego za pomocą PSM. Podejścia podstawowe oraz jednoczesne Rosenbauma są zastosowane w artykule do oceny odporności na występowanie nieobserwowanej zmiennej, efektu netto staży dla młodych bezrobotnych w wieku do 35 roku życia (oszacowanego za pomocą PSM), zorganizowanych przez jeden z największych powiatowych urzędów pracy w Małopolsce.

Słowa kluczowe: Propensity Score Matching, analiza wrażliwości, metody analizy wrażliwości Rosenbauma, polityka rynku pracy.