Sabina Denkowska

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

1. Introduction

One of serious drawbacks in observational studies is the selection bias caused by the selection process to the treatment group. Propensity Score Matching (PSM) is a method recommended[1] to evaluate projects and programmes co-financed by the European Union, which allows for the reduction of the selection bias while estimating the average treatment effect on treated (ATT). Propensity Score Matching (PSM) refers to matching control units to treated units based on propensity scores estimated on the basis of observed characteristics. Matching methods (in particular PSM) rely on a strong assumption, called 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]. According to critics, the main drawback 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. It is not always an objection directed at the design stage and the data gathering stage. An unobserved confounder U may be unmeasurable or difficult to measure. And because this confounder has not been taken into consideration during the matching process, units from both groups may not be comparable in the sense of an unobserved confounder U. Thus, the estimated effect can be caused not necessarily by the treatment, but by the lack of balance[2] of an unobserved confounder U, which affects both the outcome and the

[1] The Programming Period 2014-2020, Guidance Document on Monitoring and Evaluation – European Regional Development Fund and Cohesion Fund – Concepts and Recommendations, 2014, pp. 6-7.
[2] Balancing of variables means the similarity of distributions understood as the lack of systematic differences in their distributions.

selection process into treatment[3]. That is why Rosenbaum [2005] recommends to complement observational studies based on matching with the sensitivity analysis, which allows for the assessment of robustness of the estimated effect to a potential unobserved confounder.

In the paper the primal and simultaneous Rosenbaum's approaches will be applied to assess robustness to unobserved confounder of the net effect of internships (estimated with PSM) for the young (up to 35) unemployed**,** organized by one of the biggest District Employment Offices in Małopolska.

1. Propensity Score Matching (PSM)

2.1 Notation, definitions, assumptions

Let X denote a vector of observable characteristics and $\mathrm{D}-$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: :

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

Evaluation studies usually aim at estimating an average treatment effect on treated (ATT), which allows to conclude whether the treatment was effective for units in the treatment group:

$$\tau_{ATT} = \mathrm{E}[Y^1 - Y^0 | \mathrm{D} = 1] \tag{2}$$

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

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

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

In PSM, matching is based on *propensity score* p, which is defined as conditional probability of being treated for a given vector x of observed characteristics X [Rosenbaum, Rubin 1983]:

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

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

The key assumption of PSM is Conditional Independence Assumption[4] (CIA) that treatment assignment D is independent of potential outcomes $\left(\mathrm{Y}^1, \mathrm{Y}^0\right)$ conditional on a given vector of covariates X (in notation of Rosenbaum, Rubin [1983]):

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

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 *propensity score* $p(\mathrm{x})$. The Conditional Independence Assumption is untestable, and what is more, easy to undermine in observational studies, which may lead to questioning the results obtained with the use of the PSM method.

The second assumption of PSM is called the common support or overlap assumption [Caliendo, Kopeing 2008] and can be written as:

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

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

Both, unconfoundedness and overlap assumptions constitute a property which is known as the "strong ignorability of assignment". The "strong ignorability" is necessary[5] for identifying the treatment effect [Rosenbaum and Rubin 1983].

2.2 Algorithm of PSM

In practice, the propensity score is 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. In the estimated logistic model all variables that simultaneously influence the selection process and the outcome variable should be included[6] [Stuart 2010]. In case of the PSM method, the model is only a means for achieving the goal, which is balancing variables, thus all attention should be focused not on estimating model parameters, but on the capability of this model to

[4] also known as ''ignorability'' [Rubin 1978] , "no hidden bias" or ''unconfoundedness". [Caliendo, Kopeinig 2008].

[5] However, for ATT, Heckman et al. [1998] show that the unconfoundedness assumption can be weakened to conditional mean independence [see also: Abadie, Imbens 2006]. Also the overlap assumption can be weakened because sufficient condition to identifying ATT is the right inequality in formula (6) [e.g. Caliendo,Kopeing 2008, Stawiński 2014].

[6] In order to satisfy the assumption of conditional independence [Rubin, Thomas 1996].

balance variables [Caliendo, Kopeinig 2008, Stuart 2010]. Next, an algorithm of matching[7] control group to the treatment group, on the basis of estimated values of *propensity score,* is selected[8]. A very important stage of the PSM is the assessment of the quality of matching, because the effects of treatment can be evaluated only when the matching quality is satisfying. The assessment of the matching quality consists of checking and, if necessary, determining a region of common support and checking balance of variables included in the estimated logistic model. More information about determining the region of common support and indicators and tests used for assessing balance of variables can be found in e.g. Caliendo, Kopeing [2008], Stuart [2010], Strawiński [2006, 2014], Denkowska [2015]. If balance of variables cannot be considered satisfying, the researcher should consider other algorithms of matching or the return to the stage in which the model of logistic regression is estimated and introduce interactions and (or) squared qualitative variables into the model [Stuart 2010; Caliendo, Kopeinig 2008]. A tedious process of searching for the model and the best matching algorithm to satisfy balance of all variables, higher-order terms and interactions from the estimated logistic model will not necessarily be successful. It may indicate a failure of the CIA [Smith and Todd, 2005] and alternative evaluation approaches should be considered [Caliendo, Kopeing 2008].

The estimation of the treatment effect can be conducted only after obtaining satisfying balance of all variables, higher-order terms and interactions of covariates used in the model [Rubin 2001, p. 169].

3. Sensitivity analysis with Rosenbaum's approaches

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 the 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. These approaches differ in terms of finding the thresholds of the association between the unobserved confounder and the exposure or (and) the outcome that would render the test statistics of the study inference

[7] Detailed presentations of matching methods of the control group and variants of their application (with or without replacement, 1:*k* matching, with caliper, etc.) can be found in e.g. Caliendo, Kopeinig [2008], Stuart [2010].

[8] In Polish labour market evaluation studies the Nearest Neighbour Method with 1:1 matching is most frequently chosen (Wiśniewski, Maksim [2013], Konarski, Kotnarowski [2007], Trzciński [2009]).

insignificant [Stuart 2013]. In the primal approach, sensitivity parameter Γ relates unobserved confounder U to treatment D, while assuming that a confounder is a perfect predictor of the outcome. In the dual approach, sensitivity parameter Δ relates unobserved confounder U to outcome Y, while assuming that a confounder is a perfect predictor of the treatment. The simultaneous sensitivity analysis uses both sensitivity parameters Γ and Δ. From a practical point of view, primal and simultaneous approaches are the most important.

Rosenbaum's sensitivity analyses assume that matching has been done without replacement.

3.1 Primal Rosenbaum's approach

In the primal approach the question to be answered is how strongly an unobserved confounder associated with treatment should be to change the conclusions of the study, assuming that a confounder is a perfect predictor of the outcome.

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 ($\boldsymbol{\pi_k = \pi_l}$). Because of a potential unobserved confounder U, the odds that unit $\boldsymbol{k}$ receives treatment *de facto* is:

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

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

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

Rosenbaum [2002] shows the following bounds on the odds-ratio:

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

Let $\Gamma := exp(\gamma)$. "Two units which look the same at baseline before treatment, that is two units with the same observed covariates, may nonetheless differ in terms of unobserved covariates, so that one subject has the odds of treatment that are up to $\Gamma \geq 1$ times greater than the odds for another unit" [Rosenbaum 2005].

The sensitivity analysis with an unobserved confounder *U* proposed by Rosenbaum [2002] is based on several different randomization tests [see e.g..: Liu, Kuramoto, Stuart 2013, Keele 2010]. For a binary[9] outcome variable Y, the sensitivity analysis is based on McNemar's test. McNemar's test is used for checking if the confounder U has a significant impact on the result of the outcome variable *Y*. Information on paired units are presented in a contingency table (2x2). For units paired on the bases on *propensity score* the chances of being selected for a treated group are theoretically the same. In Rosenbaum's sensitivity analysis we analyse for which odds ratio of treatment of paired units (occurring due to unobserved confounder *U)* the conclusions of the study would change (i.e. by making 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 individual has a positive result of the outcome variable while a not-treated individual – a negative result. The lower and upper bounds on the p-value are calculated by analogy to binomial test *p*-value:

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

where probabilities:

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

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.

3.2 Rosenbaum's simultaneous approach

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

$$p^+ = \; p(treated) * \; p(outcome) \; + \; (1 - p(treated)\,) * (1 - p(outcome)\,) \quad (11)$$

where

[9] 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 2002, Keele 2010].

$p(treated) = \frac{\Gamma}{1+\Gamma}$ - probability of becoming a trainee, (12)

$p(outcome) = \frac{\Delta}{1+\Delta}$ - probability of getting a job, (13)

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 a point at which the result is sensitive to an unobserved confounder U [Liu, Kuramoto, Stuart 2013].

4. Empirical example

In the empirical example the net effect of internships organized by the District Employment Office in Tarnów for the young unemployed (up to 35) was estimated in order to verify how effective internships organized by District Employment Offices are as an activation method among the young unemployed[10]. The study was conducted with the use of PSM and the estimated effect was subjected to the sensitivity analysis with Rosenbaum's primal and simultaneous approaches in order to check its robustness to a potential unobserved confounder influencing both the inclusion to the group of interns and finding a job.

In 2013, 1409 unemployed persons up to 35 began internships and finished them at least 3 months before 10.08.2014. The source of data was a computer system of unemployment registration *Syriusz*.

Variables X used in the study can be divided into 4 categories[11]:

I. Socio-demographic characteristics and characteristics describing health (*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. Characteristics connected with employment, professional activity and educational activity (job – classification[12] (*gr00* - lack, *grX* - where *X* denotes the number according to classification), *staz_pr* –number of years in employment, *dl_bzr* – long-term

[10] The study of the net effect of the internships for all unemployed (regardless of age) can be found in Denkowska [2015]

[11] The preliminary selection of variables was based on the experience of the team working on the *Alternatywa II* project, who conducted the evaluation of the project with the use of the PSM method. The project was part of the latest edition of Phare SSG RZL 2003 (R. Trzciński, 2009), and its source of data was SI PULS. However, after consulting the employees from the District Employment Office, it turned out that not all features should and could be used in the study due to the limits of the *Syriusz* system. The paper describes the final set of variables used in the study.

[12] Classification in accordance with the ordinance of the Minister of Labour and Social Policy of 27.04.2010 on the Occupations and Specializations for Labour Market Needs and its scope.

unemployment (Yes/No), *szk* –trainings during two years before the internship (Yes/No), *l_prop* – the number of job offers during the last six months, *w_a* – activity indicator (community work, intervention jobs, trainings, internships, public work) during the last two years before the internship: 0 – no active days, 1 – up to 100 active days, 2 – up to 200 active days, etc.),

III. Characteristics referring to relative motivation to look for a job (*pr_zas* – eligibility for the unemployment benefit),

IV. Characteristics describing skills and abilities (*pr_B*- driving licence, category B, *angBG* – at least good knowledge of English, *angSL* –basic knowledge of English, *j_n* – knowledge of German).

The outcome variable *Y* was employment after 3 months after finishing the internship. It was assumed[13] that a person who was not registered on the verification day was employed.

The control pool consisted of 11568 young unemployed (up to 35) not subjected to activation in 2013. In order to establish the value of variables *X* and outcome variable *Y*, for each person from the control pool the date of 'starting' activation was randomly selected (measuring the values of variables *X*), next the average duration of the internship was added and after 3 months from the date of 'finishing' the internship, the registration of a person in the base was verified (the value of variable *Y*).

First, the logistic regression model was estimated, in which participation in the internship was the dependent variable. Numerous attempts directed at obtaining the best possible balance of variables included modifications of the regression model by introducing to it interactions, squares of variables and checking various matching algorithms without replacement[14]. The distributions of *propensity scores* in the group of interns and control group were analysed in order to check the region of common support, which influenced the decision to use a matching algorithm with caliper. Each time before and after matching, balance of variables, interactions and squares of variables were checked with the use of: standardized mean difference, *t*-tests for means in the interns group and in the control group. In case of continuous and discreet

[13] according to the methodology used by WUP (Regional Job Center) in Cracow
[14] Rosenbaum's sensitivity analysis can be applied only for matching methods without replacement.

variables, similarity of distributions in the interns group and in the control group was analysed with the use of the bootstrap KS test[15].

The best balance of variables was obtained for the logistic model to which interactions and $wiek^2$ variable were introduced (Table 1). The Nearest Neighbour Method used in the study (1:1, without replacement, with caliper[16] ($caliper = 0{,}5$)) led to the removal of two interns, for whom there were no good matches in the control group.

Table 1 presents standardized mean differences[17] obtained from the formulas:

$$SDiff_{before} = \frac{\bar{X}_B - \bar{X}_{CP}}{S_B} 100\%, \qquad SDiff_{after} = \frac{\bar{X}_{BM} - \bar{X}_{CM}}{S_B} 100\% \qquad (13)$$

where: $\bar{X}_B, \bar{X}_{CP}$ - denote means in the beneficiaries (interns) group and in the control pool before matching, $\bar{X}_{BM}, \bar{X}_{CM}$ - denote means in the beneficiaries group and in the control group after matching, while S_B - stands for standard deviation in the beneficiaries group before matching. The analysis of standardized differences is based on checking whether the values of standardized differences for all variables (per module) decreased after matching, and whether the values obtained after matching can be considered satisfying. In most empirical studies a standardized mean difference below 3% or 5% is considered sufficient [Caliendo, Kopeinig 2008].

Table 1 presents also p-values from t-tests for means for all variables, interactions and $wiek^2$ variable and p-values from KS bootstrap test[18] for all continuous and discreet variables and interactions.

All standardized differences after matching decreased and did not exceed 4.3%, t-tests did not reveal significant differences between means, and Smirnov-Kolmogorov bootstrap test 'confirmed' similarity of distributions for continuous and discreet variables.

---

[15] Bootstrap version of the Smirnov-Kolmogorov test can be used both in case of continuous and discreet random variables [Abadie 2002, Sekhon 2011].

[16] Rubin and Thomas [1996] recommend to keep the limit at the $0{,}25\,s$ or $0{,}5\,s$, where: $s = \sqrt{\frac{S_B^2 + S_{PK}^2}{2}}$, and S_B^2 and S_{PK}^2 denote variance respectively in beneficiaries group and in the control pool. It is worth noticing that some units from the beneficiaries group may stay without matching.

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

[18] Kolmogorov-Smirnov bootstrap test allows for testing similarity of distributions of both continuous and discreet random variables [Abadie 2002, Sekhon 2011].

Table 1. Standardized mean differences, p-values from t-test for means and Kolmogorov-Smirnov bootstrap test before and after matching.

Variables	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-0	-	-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	-
*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

* Variable *st_pr* is a continuous variable, so similarity of distributions was also verified with the classic Kolmogorov-Smirnov test, which failed to reject the null hypothesis about similarity of distributions after matching (p=0.11908).
Source: own calculations in *Matching* package in *R*.

All standardized differences after matching decreased and did not exceed 4.3%, t-tests did not reveal significant differences between means, and Smirnov-Kolmogorov bootstrap test 'confirmed' similarity of distributions for continuous and discreet variables.

After all variables, interactions and variable $wiek^2$ were considered balanced, the net effect of internships for the unemployed (up to 35) was estimated. The estimated net effect of internships for the unemployed up to 35 was 10,945%, with standard error[19] 1,87% ($p = 5.1905\text{e} - 09$).

Despite all the effort put into exhaustive use of information collected in the *Syriusz* system, doubts appeared whether observed causality between participation in internships and employment is not *de facto* caused by an unobserved confounder. For example, certain personality features, such as entrepreneurship or communication skills, definitely have a strong impact on employment, hence a question arises: how strong the impact of this unobserved factor on employment and the selection process should be to make the results statistically insignificant?

In order to conduct sensitivity analysis using Rosenbaum's primal approach, the results obtained for 1407 pairs are presented in the contingency table (Table 2). The number of all pairs in which the results of outcome variable Y differed among 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

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

Source: own calculations in Matching package *in R.*

[19] See: Imbens, Ambadie [2006].

During the next step, for hypothetical values of Γ, probabilities p^- and p^+ were calculated, which were next used to obtain a lower and upper bounds for p-value following formulas (9) and (10). Table 3 presents the results of calculations for selected values of Γ.

Table 3. Bounds for selected values of Γ

Gamma	*Probabilities*	
	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: own calculations in Matching package *in R.*

In tab. 3 we can notice 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. It means that one person in a pair may have odds 1,36 times greater to become an intern as the other because of different values of confounder *U* (which has a strong influence on employment), but there is still strong evidence that internships have an impact on employment ($p = 0{,}04578$). On the other hand, when Γ=1,37, the relation between internships and employment would no longer be significant ($p = 0{,}05581$). Parameter Γ=1,36 indicates[20] moderate robustness to the occurrence of an unobserved variable *U*.
The results obtained (Table 2) were confirmed by the analysis conducted with the use of *rbound* package in R [Keele 2010]. In this package Rosenbaum's primal approach is available for binary, ordinal and continuous variables for the matching variant[21] 1:*k*.

In Rosenbaum's simultaneous approach we look for the smallest values[22] of parameters Γ oraz Δ for which $p_{upper} \geq 0.05$ (calculated from formulas (11) - (13)).

[20] In social sciences values of Γ are usually from 1 to 2 [Keele 2010].
[21] cran.r-project.org/web/packages/rbounds/rbounds.pdf (2.07.2014)

This way we obtain points (Γ, Δ) at which the result is sensitive to an unobserved confounder U [Liu, Kuramoto, Stuart 2013]. The results of Rosenbaum's simultaneous analysis are presented in Table 4.

Table 4. The results (p_{upper}) of the simultaneous approach for different values of Γ and Δ.

Γ \ Δ	**1.0**	**1.5**	**2.0**	**2.25**	**2.3**	**2.5**	**3.0**	∞
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: own calculations in R.

Analysing the results in Tab. 4, we can notice that for $\Delta = \Gamma = 2.25$ p_{upper} is 0.035, so one person in a pair may be 2.25 times as likely to become an intern and 2.25 times as likely to get employment as the other because they have different values of U, but there is still strong evidence that internships have an impact on employment ($p = 0.035$). On the other hand, when $\Delta = \Gamma = 2.3$, the association between internships and employment would no longer be significant ($p = 0.0530$).

The 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 as likely to get employment as the other because of different values of U, but there is still strong evidence that internships have an impact on employment ($p = 0.0257$). On the other hand, when Γ=2 and Δ=3, causality between internships and employment would no longer be significant ($p = 0.0957$). And, by analogy, one person in a pair may be 2.5 times as likely to become an intern and twice as likely to get employment as the other because they have different values of U, but there is still strong evidence that internships have an impact on employment ($p = 0.0257$). On the other hand, for Γ=3

[22] to one or two decimal places

and Δ=2, causality between internships and employment would no longer be significant ($p = 0.0957$).

The results presented in Table 5 indicate that the primal analysis of sensitivity is a particular case of the simultaneous approach. Probabilities (bold) for different values of Γ when $\Delta \to \infty$ (or values of Δ when $\Gamma \to \infty$) are the same as in Table 3.

Table 5. The results (p_{upper}) of the simultaneous approach for different values of Γ and Δ.

Γ \ Δ	**1.0**	**1.3**	**1.36**	**1.37**	**1.4**	**1.5**	**2.0**	**2.5**	**∞**
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: own calculations in R.

5. Conclusions

The sensitivity analysis is very important because the researcher conducting an observational study can never be sure that all confounders have been taken into account. Rosenbaum [2002] recommends conducting such studies in two stages. The 'classical' matching based on Conditional Independence Assumption should always be conducted with the sensitivity analyses on the occurrence of an unobserved confounder. The application of sensitivity analysis can 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 indicate that the obtained result is sensitive to deviations from the unconfoundedness assumption, and some caution is advised while interpreting.

In the empirical example presented in the paper, Rosenbaums's primal and simultaneous sensitivity analyses were applied to the estimated (with PSM) net effect of internships for the young unemployed (up to 35) organised by one of the biggest

District Employment Offices in Małopolska. Unfortunately, robustness of the results obtained in the study cannot be related to other similar studies, because analyses (based on PSM) of the labour market in Poland[23] are not complemented with the sensitivity analyses. However, it does not mean that the analysis of robustness of the estimated results should be abandoned, on the contrary, it should become an important element of all observational studies. The knowledge of robustness of the estimated results to an unobserved confounder can be very helpful for decision-makers when drawing conclusions from such studies.

Bibliography

Abadie A., Imbens G.W., 2006*, Large sample properties of matching estimators for average treatment effects*, Econometrica, vol. 74(1), 235-267.

Caliendo, M., Kopeinig S., 2008*, Some Practical Guidance for the Implementation of Propensity Score Matching*, Journal of Economic Surveys, 22(1), 31–72.

Denkowska S., 2015, *Wybrane metody oceny jakości dopasowania w Propensity Score Matching*, [w:] Jajuga K., M. Walesiak (red.) *Taksonomia 24. Klasyfikacja i analiza danych – teoria i zastosowania,* Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 384, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2015, s. 60-74.

Heckman J.J., Ichimura H., Smith J., 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

Konarski R., Kotnarowski M., 2007 *Zastosowanie metody propensity score matching w ewaluacji ex-post*, [w:] *Ewaluacja ex-post. Teoria i praktyka badawcza*, red. A. Huber, PARP, Warszawa.

Liu W., Kuramoto S.K., Stuart E.A., 2013, *An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research*, Prevention Science 14(6), p. 570-580.

Rosenbaum P. R., Rubin D. B., 1983, *The central Role of Propensity Score in Observational Studies for Casual Effects*, Biometrika, 70(1), 41-55.

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

Rosenbaum, P. R., 2005, *Observational Study*, in Encyclopedia of Statistics in Behavioral Science, ed. Brian S. Everitt and David C. Howell. Vol. 3 John Wiley and Sons.

[23] Zob. np.: Wiśniewski, Maksim [2013], Konarski, Kotnarowski [2007], Trzciński [2009].

Rubin, D., (1990), *Formal Modes of Statistical Inference for Causal Effects*, Journal of Statistical Planning and Inference, 25, 279-292.

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.

Strawiński P., 2008, *Quasi-eksperymentalne metody ewaluacji*, Środowisko i warsztat ewaluacji, red. A. Haber, RARP, Warszawa, s. 193-220.

Strawiński P., 2014, Propensity Score Matching. Własności małopróbkowe, Wyd. Uniwersytetu Warszawskiego, Warszawa.

Stuart E.A., 2010, *Matching methods for causal inference: a review and a look forward,* Statistical Science, Vol. 25, No. 1, s. 1-21.

Stuart E.A., 2010, *Methods for Assessing Sensitivity of Non-Experimental Study Results to Unobserved Confounding*, http://academyhealth.org/files/2013/tuesday/stua.pdf (11.2015)

Trzciński R., 2009, *Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych*, PARP, Warszawa, http://www.parp.gov.pl/index/more/13335 (12.2014).

Wiśniewski Z., Maksim M., 2013, *Polityka rynku pracy w Polsce – wyniki badań ewaluacyjnych prowadzonych za pomocą metody propensity score matching,* [w:] *Rola Funduszy Unijnych w Rozwoju społeczno-gospodarczym regionu*, ZN nr 753, wyd. US, Szczecin, s. 93-110.

The Programming Period 2014-2020, Guidance Document on Monitoring and Evaluation – European Regional Development Fund and Cohesion Fund – Concepts and Recommendations, 03.2014.

Sabina Denkowska

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

(abstract)

One of serious drawbacks of observational studies is the selection bias caused by the selection process to the treatment group. Propensity Score Matching (PSM) is one of counterfactual methods more and more frequently recommended in the evaluation of projects and programmes financed by the European Union, which allows for the reduction of the selection bias while assessing the Average Treatment Effect on Treated (ATT). The key assumption of PSM is Conditional Independence Assumption, which means 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, Kopeing 2008]. If it does not hold, the estimated effect may be not so much the result of the treatment as the lack of balance of an unobserved confounder, which affects both the selection process and the outcome. Rosenbaum's

sensitivity analysis allows the researcher to determine how strongly an unobserved confounder must affect selection into treatment and/or the outcome in order to undermine the conclusions about ATT estimated with the PSM analysis. In the article Rosenbaum's primal and simultaneous approaches are be applied to assess robustness to an unobserved confounder of the net effect of internships (estimated with PSM) for the young unemployed (up to 35)**,** organised (in 2013) by one of the biggest District Employment Offices in Małopolska.

Key words: propensity score, Propensity Score Matching, sensitivity analysis, Rosenbaum’s approaches, labour market policy.