PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 384

Taksonomia 24
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronie internetowej Wydawnictwa
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 384 • 2015
Taksonomia 24 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Sabina Denkowska
Uniwersytet Ekonomiczny w Krakowie
e-mail: sabina.denkowska@uek.krakow.pl

WYBRANE METODY OCENY JAKOŚCI DOPASOWANIA W *PROPENSITY SCORE MATCHING*

Streszczenie: Do przeprowadzania ewaluacji projektów i programów finansowanych ze środków unijnych coraz częściej zalecane są metody kontrfaktyczne, a wśród nich metoda *Propensity Score Matching* (PSM), która umożliwia redukcję obciążenia selekcyjnego przy szacowaniu przeciętnego efektu oddziaływania na jednostki poddane interwencji. Kluczowym etapem metody *Propensity Score Matching* jest ocena dopasowania grupy kontrolnej do grupy poddanej interwencji, gdyż od jakości dopasowania grupy kontrolnej zależy jakość szacowanych efektów oddziaływań. Celem artykułu jest zwrócenie uwagi na ten istotny etap procedury PSM wraz z propozycją uzupełnienia go o metody graficzne umożliwiające szybką diagnostykę i dające pełniejszy obraz zbalansowania zmiennych. Przykład empiryczny dotyczy zastosowania metody PSM do szacowania efektu netto staży zorganizowanych w 2013 r. przez Powiatowy Urząd Pracy w Tarnowie. Obliczenia zostały przeprowadzone w programie *R* z wykorzystaniem pakietów *Matching* oraz *MatchIt*.

Słowa kluczowe: *propensity score*, *Propensity Score Matching*, badania kontrfaktyczne.

DOI: 10.15611/pn.2015.384.06

1. Wstęp

Badania eksperymentalne oparte na randomizacji są niezwykle rzadko stosowane w badaniach społeczno-ekonomicznych. Zazwyczaj prowadzi się badania obserwacyjne, w których występuje obciążenie selekcyjne spowodowane wyborem jednostek do grupy poddawanej oddziaływaniu.

Do przeprowadzania ewaluacji projektów oraz programów współfinansowanych ze środków unijnych zalecane [The Programming Period 2014-2020 2014, s. 6-7] są metody kontrfaktyczne, a wśród nich metoda *Propensity Score Matching* (PSM). Metoda PSM umożliwia redukcję obciążenia selekcyjnego przy szacowaniu przeciętnego efektu oddziaływania na jednostki poddane interwencji. Kluczowym etapem metody *Propensity Score Matching* jest ocena dopasowania grupy kontrolnej do grupy poddanej interwencji, gdyż od jakości tego dopasowania zale-

ży jakość szacowanych efektów oddziaływań. Celem artykułu jest zwrócenie uwagi na ten istotny etap procedury PSM wraz z propozycją uzupełnienia go o metody graficzne, które umożliwiają szybką diagnostykę, dając jednocześnie pełniejszy obraz zbalansowania zmiennych.

2. Badania eksperymentalne a badania obserwacyjne

W badaniach eksperymentalnych opartych na randomizacji zmienne obserwowane w grupie eksperymentalnej oraz w grupie kontrolnej są „zbalansowane". Zbalansowanie zmiennych oznacza podobieństwo rozkładów, rozumiane jako brak systematycznych różnic w rozkładach zmiennych. Losowość doboru grupy eksperymentalnej balansuje również nieobserwowane zmienne w obu grupach. Randomizacja powoduje zatem, że grupy eksperymentalna i kontrolna są porównywalne, a tym samym różnice w wynikach zmiennej wyjściowej mogą być postrzegane jako skutek oddziaływania, jakiemu zostały poddane jednostki w grupie eksperymentalnej.

W badaniach społeczno-ekonomicznych przeprowadzanie eksperymentów opartych na randomizacji nie zawsze jest możliwe, a często byłoby po prostu nieetyczne (np. badanie wpływu spożywania alkoholu przez gimnazjalistów na wyniki w nauce). Przeprowadza się badania obserwacyjne, w których występuje obciążenie selekcyjne spowodowane wyborem jednostek do grupy poddawanej oddziaływaniu. Brak losowości przy doborze jednostek do grupy poddawanej oddziaływaniu skutkuje tym, że zmienne w obu grupach nie są zbalansowane, a zatem obie grupy nie są porównywalne.

3. Model przyczynowy Neymana-Rubina

Model przyczynowy Neymana-Rubina[1] [Sekhon 2008] stanowi fundament wnioskowania przyczynowego. Podstawowymi pojęciami modelu przyczynowego są: jednostki, oddziaływanie oraz potencjalne wyniki zmiennej wyjściowej.

Niech X oznacza wektor obserwowanych charakterystyk jednostki, zaś D – oddziaływanie ($D \in \{0,1\}$), przy czym $D = 1$ oznacza, że jednostka została poddana oddziaływaniu, a $D = 0$ oznacza, że nie została ona poddana oddziaływaniu. Dla każdej i-tej jednostki z N – elementowej populacji możliwy jest tylko jeden z dwóch wyników zmiennej wyjściowej Y:

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

[1] W literaturze tematu występuje również pod nazwą modelu przyczynowego Rubina [Holland 1986].

Efekt przyczynowy (*causal effect*) jest zdefiniowany jako różnica pomiędzy dwoma potencjalnymi wynikami zmiennej wyjściowej, przy czym tylko jeden z wyników zmiennej Y jest obserwowany [Sekhon 2008].

Fundamentalnym problemem wnioskowania przyczynowego (*causal inference*) jest fakt, iż dla każdej *t*-tej jednostki obserwujemy tylko jeden z wyników zmiennej wyjściowej Y [Holland 1986]. Nieobserwowany wynik zmiennej wyjściowej Y nazywany jest wynikiem kontrfaktycznym.

4. Przeciętny efekt oddziaływania na jednostki poddane interwencji

Ewaluacja oddziaływania (*impact evaluation*) ma na celu dokonanie pomiaru efektu oddziaływania D na zmienną wyjściową Y. W badaniach społeczno-ekonomicznych oddziaływaniem może być na przykład interwencja polegająca na przeznaczaniu środków na organizację programów, projektów czy szkoleń dla pewnych grup społecznych. Jednostkami poddawanymi interwencji[2] mogą być osoby, gospodarstwa domowe lub instytucje. Efekt interwencji mierzony jest za pomocą zmiennej wyjściowej Y.

Najczęściej w badaniach ewaluacyjnych celem jest estymacja przeciętnego efektu oddziaływania na jednostki poddane interwencji ATT (*Average Treatment Effect on Treated*):

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

który umożliwia ocenę, czy podjęte działania interwencyjne były opłacalne. Przeciętny efekt oddziaływania dla jednostek poddanych oddziaływaniu można przedstawić następująco:

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

Odjemnik w rozwinięciu wzoru (3) to tzw. obciążenie selekcyjne wynikające z niepokrywających się obszarów określoności oraz z tego, że wektor X oraz zmienne nieobserwowane w grupie poddanej interwencji i w puli kontrolnej nie są zbalansowane.

5. Metoda *Propensity Score Matching* (PSM)

Metoda *Propensity Score Matching* umożliwia redukcję obciążenia selekcyjnego przy szacowaniu przeciętnego efektu oddziaływania na jednostki poddane interwencji ATT. Polega na dopasowywaniu do grupy poddanej interwencji takiej grupy kontrolnej, wyselekcjonowanej z puli kontrolnej osób niepoddanych oddziaływaniu, że rozkłady charakterystyk wektora X w obu grupach będą zbalansowane. Rosenbaum

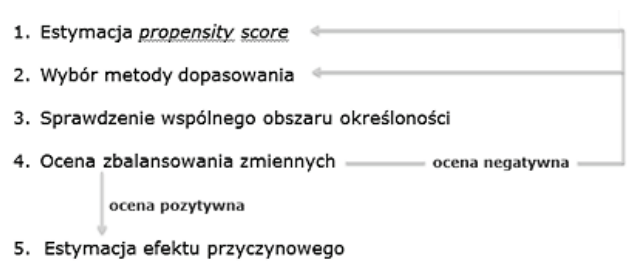
[2] Jednostki poddane interwencji w dalszej części artykułu nazywane są również beneficjentami.

i Rubin [1983] zaproponowali, aby do dopasowywania jednostek wykorzystywać funkcję balansującą *propensity score p*:

$$p(X) := P(D = 1 \mid X). \tag{4}$$

Dopasowywanie na podstawie wartości *Propensity Score* (PS) pozwala uniknąć problemów z wielowymiarowością wektora charakterystyk X.

Rosenbaum i Rubin [1983], proponując metodę PSM, przedstawili również jej założenia modelowe. Pierwszym z nich jest *założenie warunkowej niezależności* (*Conditional Independence Assumption*), dotyczące statystycznej niezależności mechanizmu doboru od potencjalnych wyników zmiennej wyjściowej *Y*, pod warunkiem wektora charakterystyk *X*, czyli $\left(Y^1, Y^0\right) \perp D \mid X$ dla każdego wektora charakterystyk *X*. Powyższe założenie oznacza, że cały proces selekcji musi być oparty jedynie na obserwowanych charakterystykach wektora *X* oraz wszystkie zmienne wpływające na udział w programie oraz na wynik zmiennej wyjściowej *Y* są obserwowane przez badacza [Caliendo, Kopeinig 2008]. Niestety, to kluczowe założenie jest w praktyce nietestowalne. Założenie drugie dotyczy *wspólnego obszaru określoności* (*common support*) i oznacza, że każda jednostka populacji musi mieć dodatnie prawdopodobieństwo trafienia do grupy beneficjentów[3]: $0 < P(D = 1 \mid X) < 1$. Spełnienie obu powyższych założeń jest warunkiem niezbędnym do szacowania efektu oddziaływania [Rosenbaum, Rubin 1983; Sekhan 2008]. Rosenbaum i Rubin [1983] wykazali, że jeśli spełnione jest założenie statystycznej warunkowej niezależności mechanizmu doboru od potencjalnych wyników zmiennej wyjściowej *Y* dla danego wektora charakterystyk *X*, to również jest ono spełnione dla *propensity score p*(*X*). Do założeń modelowych metody *PSM* należy również założenie SUTVA (*Stable Unit Treatment Assumption*), mówiące, że wynik zmiennej wyjściowej *Y* dla poszczególnych jednostek nie może zależeć od wyników pozostałych jednostek [Rubin 1978].



Rys. 1. Algorytm metody *Propensity Score Matching*

Źródło: opracowanie własne na podstawie wielu źródeł.

[3] Nie może to również być zdarzenie pewne, gdyż byłby problem z dopasowaniem jednostki z puli kontrolnej.

Algorytm metody *Propensity Score Matching* przedstawia rys. 1. W praktyce wartości *propensity score* (etap 1 na rys. 1) najczęściej wyznaczane są jako wartości teoretyczne estymowanego w tym celu modelu logistycznego. W przypadku metody PSM model jest jedynie środkiem do osiągnięcia celu, którym jest zbalansowanie zmiennych, zatem cała uwaga powinna być skoncentrowana nie na estymacji parametrów modelu, ale na zdolności tego modelu do balansowania zmiennych [Augurzky, Schmidt 2001, Stuart 2010]. W estymowanym modelu logistycznym powinny zostać uwzględnione[4] wszystkie zmienne wpływające zarówno na selekcję, jak i na ostateczny wynik zmiennej wyjściowej *Y* [Stuart 2010].

W kolejnym kroku (etap 2 na rys.1) wybierana jest metoda dopasowania[5] grupy kontrolnej do grupy beneficjentów na podstawie oszacowanych wartości *propensity score*. W polskiej praktyce ewaluacyjnej[6] najczęściej stosowana jest Metoda Najbliższego Sąsiada[7] z dopasowywaniem 1:1.

Kluczowym etapem procedury PSM jest ocena dopasowania grupy kontrolnej, bowiem do szacowania efektów interwencji można przystąpić dopiero, gdy jakość dopasowania grupy kontrolnej można uznać za satysfakcjonującą. W ramach oceny jakości dopasowania grupy kontrolnej wyróżniamy sprawdzenie i ewentualnie wyznaczenie[8] wspólnego obszaru określoności (etap 3 na rys.1) oraz ocenę zbalansowania zmiennych (etap 4 na rys.1).

W celu sprawdzenia i ewentualnego wyznaczenia wspólnego obszaru określoności w literaturze tematu można spotkać różne zalecenia: wizualne porównanie rozkładów *propensity score* w obu grupach, wyznaczenie wartości maksymalnych oraz minimalnych w obu grupach lub też zastosowanie metody *trimming,* w której wspólny obszar określoności zdefiniowany jest tylko w przedziałach, gdzie gęstość rozkładu PS w obu grupach przekracza pewną dodatnią wartość progową [Smith, Todd 2005]. Innym sposobem na uniknięcie problemów ze wspólnym obszarem określoności jest zastosowanie metody z limitem, czyli z maksymalną dopuszczalną odległością między wartościami *propensity score* w obu grupach. Rubin i Thomas [1996] zalecają, by limit był na poziomie 0,25s lub 0,5s, gdzie: $s = \sqrt{\frac{S_B^2 + S_{PK}^2}{2}}$, a S_B^2 oraz S_{PK}^2 oznaczają wariancje odpowiednio w grupie beneficjentów oraz w puli kontrolnej. Warto zauważyć, że tym samym część jednostek z grupy beneficjentów może pozostać bez dopasowania.

[4] W celu spełnienia założenia warunkowej niezależności [Rubin, Thomas 1996].

[5] Szczegółowe prezentacje metod dopasowania grupy kontrolnej, jak również wariantów ich stosowania (ze zwracaniem lub bez, dopasowywanie 1:k, z limitem lub bez itd.) można znaleźć np. u Caliendo, Kopeinig [2008], Stuart [2010].

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

[7] Zob.: [Rubin1978].

[8] Caliendo, Kopeinig [2008] podkreślają, że przeciętne efekty oddziaływań można szacować tylko we wspólnym obszarze określoności.

Do oceny zbalansowania zalecane jest stosowanie zarówno numerycznych, jak i graficznych metod diagnostycznych [Stuart 2010]. Do podstawowych numerycznych metod oceny zbalansowania należy zaproponowana przez Rosenbauma i Rubina [1983] metoda badania standaryzowanych różnic średnich[9]:

$$SDiff_{przed} = \frac{\left(\overline{X}_B - \overline{X}_{PK}\right)}{\sqrt{\frac{S_B^2 + S_{PK}^2}{2}}} \cdot 100\%, \qquad SDiff_{po} = \frac{\left(\overline{X}_{BM} - \overline{X}_{KM}\right)}{\sqrt{\frac{S_{BM}^2 + S_{PK}^2}{2}}} \cdot 100\%, \qquad (5)$$

gdzie $\overline{X}_B, \overline{X}_{PK}, S_B^2, S_{PK}^2$ oznaczają średnie oraz wariancje odpowiednio w grupie beneficjentów oraz w puli kontrolnej przed dopasowaniem, zaś $\overline{X}_{BM}, \overline{X}_{KM}$ oznaczają średnie po dopasowaniu w grupie beneficjentów oraz w grupie kontrolnej, a S_{BM}^2 oznacza wariancję w grupie beneficjentów po dopasowaniu. Zauważmy, że we wzorach (5) zaproponowanych przez Rosenbauma i Rubina [1983] zarówno przed dopasowaniem, jak i po dopasowaniu, w mianowniku występuje wariancja w puli kontrolnej S_{PK}^2. W późniejszych pracach wzór na standaryzowane różnice średnich po dopasowaniu pojawia się w wersji zmodyfikowanej, gdzie we wzorze $SDiff_{po}$ wariancja w puli kontrolnej S_{PK}^2 została zastąpiona wariancją S_{KM}^2 w dopasowanej grupie kontrolnej. W 2006 roku Austin i Mamdani [2006] zaproponowali modyfikację wzorów na standaryzowane różnice średnich, zalecaną[10] w opracowaniach na temat PSM, w której S_B oznacza odchylenie standardowe w grupie beneficjentów przed łączeniem:

$$SDiff_{przed} = \frac{\left(\overline{X}_B - \overline{X}_{PK}\right)}{S_B} \cdot 100\%, \qquad SDiff_{po} = \frac{\left(\overline{X}_{BM} - \overline{X}_{KM}\right)}{S_B} \cdot 100\%. \qquad (6)$$

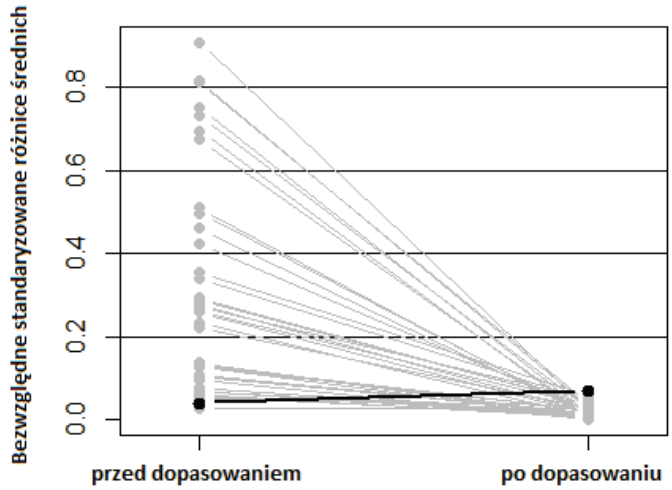
W przypadku zmiennych dychotomicznych standaryzowane różnice średnich mogą być wyznaczane z powyższych wzorów (traktując te zmienne jako ciągłe) [Stuart 2010] lub ze wzorów opartych na frakcjach [Austin 2011].

Analiza standaryzowanych różnic polega na sprowadzeniu, czy po dopasowaniu zmalały wartości standaryzowanych różnic dla wszystkich zmiennych (na moduł) oraz czy wartości otrzymane po dopasowaniu pozwalają uznać dopasowanie za satysfakcjonujące. Bardzo pomocny w ocenie zmian standaryzowanych różnic średnich przed dopasowaniem i po dopasowaniu jest przykładowy wykres przedstawiony na rys. 2.

---

[9] Standaryzowane różnice średnich występują w literaturze również pod nazwą *standaryzowane obciążenia* (np. u Rosenbauma, Rubina [1983], Caliendo, Kopeinig [2008], Strawińskiego [2008]).

[10] Np. Stuart [2010].



Rys. 2. Przykładowy wykres wartości bezwzględnych standaryzowanych różnic średnich

Źródło: opracowanie własne w *R* w pakiecie *MatchIt.*

Niestety, brak jest jasno ustalonych progów dla standaryzowanych różnic średnich, które ułatwiłyby i zobiektywizowały decyzję o zbalansowaniu cech lub o braku zbalansowania. W większości badań empirycznych redukcja standaryzowanych różnic średnich poniżej 3% lub 5% jest traktowana jako wystarczająca do stwierdzenia zbalansowania zmiennych [Caliendo, Kopeinig 2008]. Natomiast w polskojęzycznej literaturze prezentującej wyniki badań ewaluacyjnych opartych na metodzie PSM występują zredukowane wartości standaryzowanych różnic wynoszące [Trzciński 2009, s. 63] nawet ponad 20%. Takie wyniki autorzy komentują [Trzciński 2009, s. 63] jako wyniki na „niezadowalającym poziomie", niemniej jednak nie przeszkadza im to stwierdzić, że "poziom zbalansowania zmiennych można uznać za satysfakcjonujący" i przystąpić do szacowania efektu netto oddziaływania. W badaniach ewaluacyjnych często ocena zbalansowania ogranicza się właśnie do porównania standaryzowanych różnić średnich [Konarski, Kotnarowski 2007], czasami uzupełnionego o test *t*-Studenta dla średnich (np. u Trzcińskiego [2009]) zaproponowany przez Rosenbauma i Rubina [1985]. Niemniej jednak do stwierdzenia satysfakcjonującego zbalansowania zmiennych nie wystarczy jak u Trzcińskiego [2009, s. 63] zauważyć, że „wykonany test *t* nie wykazał statystycznie istotnych różnic w średnich większości zmiennych"[11]. Aby przystąpić do szacowania efektów oddziaływań, konieczne jest (podobnie jak w badaniach eksperymentalnych) zbalansowanie wszystkich zmiennych obserwowanych (ich interakcji, potęg wyższych rzędów uwzględnionych w modelu regresji logistycznej),

[11] Zob.: Trzciński [2009, s. 63].

a nie tylko zbalansowanie „większości zmiennych". W literaturze tematu można spotkać zarówno głosy zalecające [Rosenbaum, Rubin 1985; Caliendo, Kopeinig 2008], jak i głosy krytycznie oceniające wykorzystywanie testów statystycznych[12] (zob. np. [Imai, King, Stuart 2008; Austin 2011]) do oceny zbalansowania zmiennych. Dla pełniejszej diagnostyki Rubin [2001] zaleca porównanie ilorazu wariancji w grupie poddanej interwencji oraz grupie kontrolnej przed dopasowaniem i po dopasowaniu. Stuart [2010] podkreśla znaczenie diagnostyki opartej na wykresach w ocenie dopasowania zmiennych, która umożliwia szybką ocenę zbalansowania nawet w przypadku dużej liczby zmiennych. Zaleca m.in. porównanie rozkładów *propensity score* przed dopasowaniem i po dopasowaniu, m.in. w celu sprawdzenia wspólnego obszaru określoności, oraz porównanie rozkładów zmiennych ciągłych na wykresach kwantyl-kwantyl. W literaturze można spotkać jeszcze inne ciekawe propozycje[13] oceny zbalansowania.

Jeżeli zbalansowania zmiennych nie można uznać za satysfakcjonujące, to wówczas należy zastosować inne metody dopasowywania grupy kontrolnej (etap 2 na rys.1) bądź ewentualnie powrócić do etapu estymacji modelu regresji logistycznej (etap 1 na rys.1), wprowadzając do modelu interakcje oraz zmienne ilościowe podniesione do kwadratu [Stuart 2010; Caliendo, Kopeinig 2008]. Żmudny proces poszukiwania modelu i metody dopasowania grupy kontrolnej dającej satysfakcjonujące zbalansowanie wszystkich zmiennych uwzględnionych w modelu wraz z włączonymi interakcjami czy też zmiennymi w wyższych potęgach niekoniecznie musi zakończyć się powodzeniem. W takim przypadku problemem może być niespełnione założenie o warunkowej niezależności [Smith, Todd 2005], czyli istota problemu leży w danych, którymi badacz dysponuje.

Zdaniem Rubina [2001, s. 169] bardzo istotne jest to, żeby badacz miał dostęp do danych wyjściowych dopiero po otrzymaniu satysfakcjonującego zbalansowania, co pozwala uniknąć sytuacji, gdy wartość oszacowanego efektu wpływa na decyzję o uznaniu zbalansowania za satysfakcjonujące. Dopiero w przypadku otrzymania satysfakcjonującego zbalansowania wszystkich zmiennych, interakcji oraz zmiennych w wyższych potęgach uwzględnionych w modelu można przystąpić do szacowania efektu oddziaływania (etap 4 na rys. 1).

---

[12] Do oceny zbalansowania wykorzystywany jest również test Kołmogorowa-Smirnowa dla zmiennych ciągłych lub jego bootstrapowa wersja (również dla zmiennych dyskretnych).

[13] Na przykład Sianesi [2004] zaproponował, by do oceny dopasowania wykorzystywać test weryfikujący łączną istotność parametrów w modelu regresji logistycznej (ewentualnie pseudo-R^2) dla połączonych grup poddanej interwencji oraz kontrolnej otrzymanej w wyniku dopasowywania (zob.: [Sianesi 2004; Caliendo, Kopeinig 2008]). Z kolei Deheija i Wahba [1999; 2002] zaproponowali, by ocenę dopasowania przeprowadzać w warstwach, otrzymanych w wyniku podziału jednostek z obu grup na podstawie oszacowanych wartości *propensity score* w ten sposób, że dla każdej z warstw nie występuje istotna różnica pomiędzy średnimi wartościami *propensity score* dla grupy poddanej interwencji oraz grupy kontrolnej [Caliendo, Kopeinig 2008].

6. Estymacja efektu netto staży dla bezrobotnych – przykład empiryczny

Przykład empiryczny dotyczy ewaluacji efektu netto staży dla bezrobotnych, zorganizowanych w roku 2013 przez Powiatowy Urząd Pracy w Tarnowie. W roku 2013 staż rozpoczęło 1640 osób[14]. Ostatecznie jednak w badaniu wzięły udział 1623 osoby, które zakończyły staż co najmniej trzy miesiące przed 10.08.2014. Źródłem danych był system informatyczny Syriusz.

Zmienne uwzględnione w badaniu można podzielić na 4 kategorie[15]:

I. Cechy społeczno-demograficzne oraz dotyczące stanu zdrowia (*plec* – płeć, *wiek* – wiek w latach, *sc* – stan cywilny, *s_w* – samotne wychowywanie dzieci, *n_p* – stwierdzona niepełnosprawność, wykształcenie (*w_brak* – brak, *w_podst* – podstawowe, *w_gim* – gimnazjalne, *w_zaw* – zawodowe, *w_sr* – średnie, *w_pm* – pomaturalne, *w_w* – wyższe)).

II. Cechy związane z zatrudnieniem, aktywnością zawodową oraz aktywnością szkoleniową (zawód – klasyfikacja[16] (*gr00* – brak, *grX* – gdzie *X* – numer zgodny z klasyfikacją), *staz_pr* – staż pracy, *dl_bzr* – długotrwałe bezrobocie (Tak/Nie), *szk* – szkolenia w ostatnich 2 latach przed stażem (Tak/Nie), *lprop* – liczba propozycji pracy w ciągu ostatnich 6 miesięcy, *w_a* – wskaźnik aktywności (prace społecznie użyteczne, prace interwencyjne, szkolenia, staże, roboty publiczne) w ciągu dwóch lat przed stażem: 0 – brak dni aktywnych, 1 – do 100 dni, 2 – do 200 dni itd.).

III. Cechy odnoszące się do względnej motywacji osób do poszukiwania pracy (*pr_zas* – prawo do zasiłku).

IV. Cechy dotyczące posiadanych umiejętności (*pr_B* – prawo jazdy kat. B, *angBG* – znajomość języka angielskiego w stopniu co najmniej dobrym, *angSL* – znajomość języka angielskiego słaba lub podstawowa, *j_n* – znajomość języka niemieckiego).

Zmienną wyjściową *Y* było zatrudnienie po 3 miesiącach od zakończenia stażu. Przyjęto (zgodnie z metodologią stosowaną przez WUP w Krakowie), że osoba niezarejestrowana w dniu weryfikacji w bazie jako bezrobotna ma pracę.

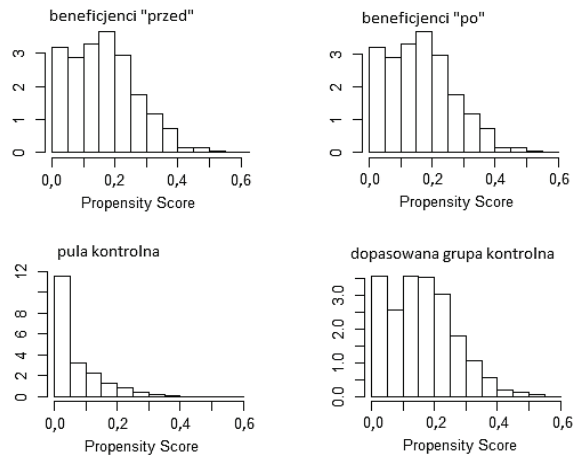
---

[14] Dla 16 osób z tej grupy to nie była jedyna aktywizacja, jakiej zostali poddani w 2013 r. (dwie osoby uczestniczyły w dwóch stażach rozpoczętych w 2013 i osoby te potraktowano w badaniu tak, jakby uczestniczyły w jednym „dłuższym" stażu, przy czym weryfikację zatrudnienia przeprowadzono po upływie trzech miesięcy od zakończenia drugiego stażu). Udział osób poddanych dodatkowej aktywizacji był poniżej 1%.

[15] Przy wstępnym wyborze zmiennych korzystano z doświadczeń zespołu prowadzącego ewaluację za pomocą metody PSM projektu Alternatywa II, realizowanego w ramach ostatniej edycji Phare SSG RZL 2003, dla których źródłem danych był SI PULS [Trzciński 2009]. W artykule prezentowany jest już ostateczny zbiór zmiennych wykorzystywanych w badaniu.

[16] Klasyfikacja zgodna z rozporządzeniem Ministra Pracy i Polityki Społecznej z dnia 27.04.2010 r. w sprawie klasyfikacji zawodów i specjalności na potrzeby rynku pracy oraz zakresu jej stosowania.

Pulę kontrolną stanowiło 19 217 osób nieobjętych aktywizacją w roku 2013. W celu ustalenia dla osób z puli kontrolnej wartości zmiennych X oraz zmiennej wyjściowej Y dla każdej osoby z puli kontrolnej wylosowana[17] została data „rozpoczęcia" aktywizacji (wprowadzono wartości zmiennych X), dodano przeciętny czas trwania stażu, a po trzech miesiącach od daty „zakończenia" zweryfikowano rejestrację osoby w bazie (wprowadzano wartość zmiennej Y).



Rys. 3. Rozkłady PS w grupach przed dopasowaniem i po dopasowaniu za pomocą MNS (1:1, *caliper* = 0,25)

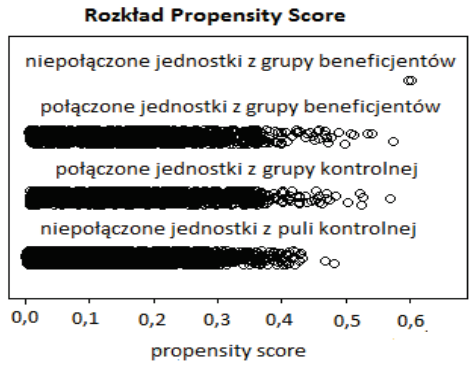
Źródło: opracowanie własne w *R* w pakiecie *MatchIt.*

Badanie rozpoczęto od estymacji modelu regresji logistycznej, w którym zmienną zależną jest uczestnictwo w stażu. Podjęto liczne próby w celu otrzymania jak najlepszego zbalansowania zmiennych, które polegały na modyfikacjach modelu regresji przez wprowadzanie interakcji, zmiennych podniesionych do kwadratu oraz sprawdzaniu różnych metod dopasowania grupy kontrolnej. Za każdym razem po dopasowaniu analizowano na wykresach rozkłady *propensity score* w grupie stażystów oraz grupie kontrolnej w celu sprawdzenia wspólnego obszaru określoności oraz badano zbalansowanie przed dopasowaniem i po dopasowaniu za pomocą: standaryzowanych różnic średnich, testów *t*-Studenta dla średnich, wykresów kwantyl-kwantyl, stosunku wariancji w grupie stażystów i w grupie kontrolnej.

Najlepsze zbalansowanie zmiennych uzyskano w przypadku modelu logistycznego z włączonymi interakcjami oraz zmienną wiek podniesioną do kwadratu (tab. 1). Zastosowana w badaniu Metoda Najbliższego Sąsiada (1:1, ze zwracaniem, z limitem

[17] Postępowanie analogiczne do postępowania przedstawionego u Trzcińskiego [2009]. Jednak ze względu na specyficzny kształt rozkładu czasu trwania staży zdecydowano o nieuwzględnianiu odchylenia standardowego w procesie wyznaczania dat zakończenia aktywizacji.

(*caliper* = 0,25)) spowodowała usunięcie dwóch stażystów, dla których nie istniało wystarczająco dobre dopasowanie w puli kontrolnej (rys. 3-4).

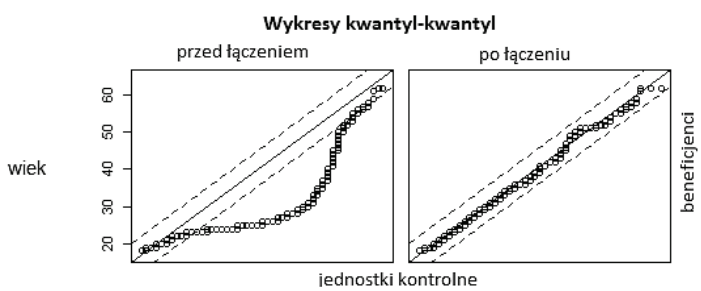


Rys. 4. Rozkłady PS w grupach przed dopasowaniem i po dopasowaniu za pomocą MNS (1:1, *caliper* = 0,25)

Źródło: opracowanie własne w *R* w pakiecie *MatchIt.*

W tabeli 1 przedstawiono średnie, standaryzowane różnice średnich oraz prawdopodobieństwa testowe otrzymane w wyniku zastosowania testu *t*-Studenta (dla średnich), przed dopasowaniem i po dopasowaniu dla wszystkich zmiennych, interakcji oraz zmiennej *wiek* podniesionej do kwadratu. Wszystkie standaryzowane różnice po dopasowaniu zmniejszyły się i na moduł nie przekraczają 2%, zaś testy *t*-Studenta nie wykazały istotnych różnic między średnimi. Stosunek wariancji w grupie stażystów oraz w grupie kontrolnej dla wszystkich zmiennych, interakcji oraz zmiennej $wiek^2$ należał do przedziału (0,83; 1,3).

Ocenę dopasowania uzupełniono o analizę wykresów kwantyl-kwantyl (przykładowe wykresy na rys. 5) oraz wykresów pudełkowych. Analiza podobieństwa rozkładów po dopasowaniu wypadła satysfakcjonująco.



Rys. 5. Wykresy kwantyl-kwantyl dla zmiennej *wiek* przed dopasowaniem i po dopasowaniu

Źródło: opracowanie własne w *R* w pakiecie *MatchIt.*.

Tabela 1. Średnie[a], standaryzowane różnice średnich oraz prawdopodobieństwa testowe otrzymane w wyniku zastosowania testu *t*-Studenta (dla średnich), przed dopasowaniem i po dopasowaniu

Zmienne	Średnia „przed" stażyści	Średnia pula kontrol.	Średnia „po" gr. kontr.	PRZED		PO	
				$SDiff_{przed}$	*Test t p-value*	$SDiff_{po}$	*Test t p-value*
plec	0,2995	0,5103	0,3004	-46,029	< 2,22e-16	-0,1354	0,96382
wiek	26,547	34,008	26,583	-81,425	< 2,22e-16	-0,3465	0,9014
sc	0,2785	0,4071	0,2757	-28,689	< 2,22e-16	0,6952	0,82872
s_w	0,0431	0,0685	0,0415	-12,475	2,4334e-06	0,8137	0,81293
n_p	0,0394	0,0823	0,0411	-22,031	4,4409e-16	-0,8320	0,81018
w_podst	0,0314	0,1528	0,0319	-69,571	< 2,22e-16	-0,27	0,92981
w_gim	0,0136	0,0435	0,0139	-25,891	< 2,22e-16	-0,2989	0,92969
w_zaw	0,1134	0,3521	0,1129	-75,27	< 2,22e-16	0,1873	0,9464
w_sr	0,4079	0,2929	0,4109	23,387	< 2,22e-16	-0,5179	0,85695
w_pm	0,0696	0,0364	0,065	13,04	3,0974e-07	1,7708	0,60225
w_w	0,3629	0,1175	0,3636	51,022	< 2,22e-16	-0,3022	0,90471
gr00	0,2403	0,1830	0,2422	13,402	1,9708e-07	-0,3689	0,90004
gr1	0,0012	0,0030	0,0014	-4,9407	0,069672	-0,3424	0,92376
gr2	0,3524	0,1151	0,3561	49,674	< 2,22e-16	-0,9395	0,71242
gr3	0,1349	0,1119	0,1303	6,746	0,0087341	1,3917	0,67085
gr5	0,1152	0,1587	0,1127	-13,602	2,197e-07	0,8319	0,80509
gr6	0,0080	0,0134	0,0082	-6,0738	0,02206	-0,2214	0,94993
gr7	0,0801	0,2632	0,0801	-67,416	< 2,22e-16	0,0409	0,98937
gr8	0,0099	0,0448	0,0102	-35,36	< 2,22e-16	-0,2835	0,93367
gr9	0,0277	0,0760	0,0286	-29,408	< 2,22e-16	-0,5252	0,87743
staz_pr	2,2064	6,4575	2,2099	-73,3	< 2,22e-16	-0,0132	0,99657
dl_bzr	0,3857	0,4051	0,3797	-3,9852	0,12355	1,1961	0,71054
l_prop	0,2933	0,2503	0,2829	6,7571	0,0085176	1,1218	0,73046
pr_zas	0,0265	0,0680	0,026	-25,812	< 2,22e-16	0,3286	0,92325
w_a	0,2298	0,1517	0,2140	10,836	2,3325e-05	1,979	0,54382
szk	0,0271	0,0147	0,0286	7,6549	0,0026068	-1,3122	0,71283
pr_B	0,5052	0,3355	0,5151	33,942	< 2,22e-16	-1,9764	0,50568
angBG	0,3031	0,1091	0,3056	42,212	< 2,22e-16	-0,7286	0,78928
angSL	0,4128	0,2767	0,4111	27,631	< 2,22e-16	0,4552	0,88297
j_n	0,2963	0,1721	0,2942	27,195	< 2,22e-16	0,5564	0,85521
*w_a*j_n*	0,0548	0,0369	0,0575	5,3546	0,036419	-0,7863	0,82356
*w_pm*szk*	0,0024	0,0010	0,0025	2,8707	0,25595	-0,0371	0,99161
*gr00*lprop*	0,0443	0,0284	0,0471	6,5501	0,010272	-1,0878	0,76805
*plec*lprop*	0,1004	0,1427	0,1004	-9,6675	0,0002017	0,0302	0,99282
wiek^2	788,67	1327,6	790,9	-81,573	< 2.22e-16	-0,2942	0,91793
*dl_bzr*w_a*	0,1571	0,0765	0,1436	12,906	3,7069e-07	**1,997**	0,53867
*pr_B*w_a*	0,1238	0,0711	0,1136	10,106	7,6004e-05	1,8811	0,56928

[a] Średnie w grupie stażystów po dopasowaniu zostały pominięte w tabeli, ponieważ zmieniły się tylko nieznacznie w stosunku do średnich przed dopasowaniem. Ta niewielka różnica spowodowana była usunięciem dwóch stażystów, dla których nie było dobrego dopasowania w puli kontrolnej.

* Standaryzowane różnice średnich wyznaczone zostały ze wzorów (6).

Źródło: obliczenia własne w *R*.

Wszystkie zmienne, interakcje oraz zmienną *wiek*^{2} uznano za zbalansowane i przystąpiono do oszacowania efektu netto staży. Oszacowany efekt netto wyniósł 7,895% z błędem standardowym[18] wynoszącym 1,48% (p = 9,1492e-08). Jest to różnica pomiędzy procentem stażystów, którzy po trzech miesiącach mieli zatrudnienie, a procentem osób zatrudnionych z dopasowanej grupy kontrolnej. Z grupy 1621 stażystów[19] zatrudnienie znalazło 951 osób, czyli ok. 58,67%, natomiast w grupie kontrolnej osób zatrudnionych było o 7,895% mniej. Procent zatrudnionych w całej puli kontrolnej wyniósł 47,895%.

7. Zakończenie

Konarski i Kotnarowski [2007] podkreślają, że jakość oszacowanego efektu netto „zależy od jakości dopasowania grupy kontrolnej oraz od stopnia spełnienia założeń leżących u podłoża metody PSM". Dlatego w literaturze tematu zalecane jest poszukiwanie zarówno metody dopasowania grupy kontrolnej, jak i poszukiwanie[20] modelu logistycznego, dla których zbalansowanie wszystkich zmiennych oraz włączonych do modelu interakcji i zmiennych w wyższych potęgach będzie można uznać za satysfakcjonujące. Mimo iż metod umożliwiających ocenę zbalansowania w literaturze tematu jest wiele, w praktyce ewaluacyjnej zazwyczaj ocena ta zawężona jest do metod stanowiących podstawowy kanon, do których należy ocena standaryzowanych różnic przed dopasowaniem i po dopasowaniu, czasami uzupełniona testami *t*-Studenta dla średnich. Istotą metody PSM jest zbalansowanie wszystkich cech obserwowanych (ich interakcji oraz potęg) tak, jakby to miało miejsce w badaniach eksperymentalnych, dlatego etap diagnostyczny dotyczący zbalansowania zmiennych warto uzupełnić o proste metody zarówno numeryczne, jak i graficzne, umożliwiające szybką ocenę i dające pełniejszy obraz zbalansowania cech. Bowiem dopiero wtedy, gdy zbalansowanie wszystkich zmiennych, ich interakcji oraz zmiennych w wyższych potęgach uwzględnionych w modelu logistycznym jest satysfakcjonujące, można przystąpić do oszacowania przeciętnego efektu oddziaływania.

[18] Zob.: [Abadie, Imbens 2006].

[19] Analiza obejmowała 1621 stażystów (dwie osoby z grupy stażystów znalazły się poza obszarem wspólnej dziedziny i nie były brane pod uwagę w badaniach).

[20] Poprzez włączanie interakcji oraz zmiennych w wyższych potęgach.

Literatura

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

Augurzky B., Schmidt C.M., 2001, *The Propensity Score: a Means to an End*, IZA Discussion Paper Series, No. 271.

Austin P.C., 2011, *An introduction to propensity score methods for reducing the effects of confounding in observational studies*, Multivariate Behavioral Research, 46 (3), s. 399-424. www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483.

Austin P.C., Mamdani M.M., 2006, *A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use*, Statistics in Medicine, 25, 2084-2106.

Caliendo M., Kopeinig S., 2008*, Some practical guidance for the implementation of propensity score matching*, Journal of Economic Surveys, 22(1), 31-72.

Deheija R., Wahba S., 1999, *Causal effects in nonexperimental studies: reevaluating the evaluation of training program*, Journal of American Statistical Association, vol. 94, no 448.

Deheija R., Wahba S., 2002, *Propensity score-matching methods for nonexperimental causal studies,* Review of Economics and statistics, 84(1), 151-161.

Holland P.W., 1986, *Statistics and causal inference*, J. Amer.Statist. Assoc*.,* 81, 945-960.

Imai K., King G., Stuart E.A., 2008, *Misunderstandings between experimentalists and observationalists about causal inference,* Journal of the Royal Statistical Society, Series A, 171, 481-501.

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.

Rosenbaum P.R., Rubin D.B., 1983, *The central role of propensity score in observational studies for casual effects*, Biometrika, 70(1), 41-55.

Rubin D.B., 1978, *Bayesian inference for causal effects: the role of randomization*, Annals of Statistics 6 (1), 34-58.

Rubin D.B., 2001, *Using propensity scores to help design observational studies: Application to the tobacco litigation*, Health Services & Outcomes Research Methodology, 2, 169-188.

Rubin D.B., Thomas N., 1996, *Matching using estimated propensity scores, relating theory to practice*, Biometrics 52, 249-264.

Sekhon J.S., 2008, *The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods*, [w:] *The Oxford Handbook of Political Methodology,* J.M. Box-Steffensmeier, H.E. Brady, D. Collier (eds.), 271-299, sekhon.berkeley.edu.

Sianesi B., 2004, *An Evaluation of the Swedish System of Active Labor Market Programms in the 1990s*, The Review of Economics and Statistics, vol. 86/1.

Smith J., Todd P., 2005, *Does matching overcome lalonde's critique of nonexperimental estimators?*, Journal of Econometrics, 125(1-2), 305-353.

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

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

Trzciński R., 2009, *Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych*, PARP, Warszawa, http://www.parp.gov.pl/index/more/13335 (05.05.2014).

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 r.

SELECTED METHODS OF ASSESSING THE QUALITY OF MATCHING IN PROPENSITY SCORE MATCHING

Summary: Counterfactual methods are more and more frequently used in the evaluation of projects and programmes financed by the European Union. One of them is Propensity Score Matching, which allows for the reduction of the selection bias while assessing the average treatment effect on treated. The key stage of Propensity Score Matching is the assessment of matching a control group to a treated group, since the quality of this matching influences the quality of the effects of impact evaluation. The article aims at emphasising the importance of this vital stage of the PSM procedure and supplements it with graphical methods enabling fast diagnosis and yielding a fuller picture of variable balance. The empirical example illustrates the use of the PSM method to evaluate the net effect of internships organised in 2013 by District Employment Office in Tarnów. The calculations were performed in *R* with *Matching* and *MatchIt* packages.

Keywords: propensity score, Propensity Score Matching, counterfactual methods.