Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych

Rafał Trzciński

Wykorzystanie techniki propensity score matching w badaniach ewaluacyjnych

Warszawa 2009

Recenzenci:
prof. dr hab. Marek Góra
dr hab. Jarosław Górniak, prof. UJ

Grafika na okładce:
Małgorzata Wojciechowska

Wydawca:
Polska Agencja Rozwoju Przedsiębiorczości
ul. Pańska 81/83
00-834 Warszawa

ISBN: 978-83-7633-088-4
Wydanie pierwsze
Nakład: 600

Opracowanie wydawnicze:
Andrzej Kirsz, Joanna Fundowicz

Przygotowanie do druku, druk i oprawa:

Wydawnictwo Naukowe Instytutu Technologii Eksploatacji – PIB
26-600 Radom, ul. K. Pułaskiego 6/10, tel. centr. (48) 364–42-41, fax (48) 364-47-65
e-mail: instytut @itee.radom.pl, http://www.itee.radom.pl

SPIS TREŚCI

Wstęp

Celem programów realizowanych przez instytucje publiczne jest rozwiązywanie konkretnych problemów społeczno-gospodarczych. Skuteczność tych działań często jest ograniczana przez złożone otoczenie realizacji interwencji czy błędy na etapie projektowania bądź wdrażania. Bywa, że niektórych założonych celów nie udaje się osiągnąć, co rodzi uzasadnione pytania o przyczyny zaistniałego stanu. Z drugiej strony, czasami uzyskane efekty znacznie przewyższają zakładany plan – co najczęściej utożsamiane jest z sukcesem. Jednak zaistnienie jednej z powyższych sytuacji – przekroczenie lub nieosiągnięcie przyjętych celów – niekoniecznie musi świadczyć o powodzeniu lub porażce danego programu.

Podejmowane interwencje publiczne nie są realizowane w przysłowiowej próżni, lecz w przestrzeni złożonej, pełnej wzajemnych powiązań, gdzie na jeden problem wywiera wpływ wiele zjawisk. Z tego powodu zmiany, obserwowane w rzeczywistości społeczno-gospodarczej, są przeważnie wypadkową oddziaływania rozmaitych czynników, a nie konkretnej, pojedynczej interwencji. Nigdy nie ma całkowitej pewności, czy obserwowane po realizacji programu efekty są jego faktyczną konsekwencją, czy też zostałyby osiągnięte także w przypadku braku jakichkolwiek działań.

To, w jakim stopniu dana interwencja przekłada się na oczekiwane skutki, ma fundamentalne znaczenie z punktu widzenia racjonalności wydatkowania środków publicznych. Jeśli program nie przynosi pożądanych wyników, należy go zmodyfikować lub w skrajnym przypadku zrezygnować z jego realizacji. Tylko w jaki sposób ustalić, że dana interwencja jest nieskuteczna, skoro obserwowane zmiany zależą od tak wielu czynników?

Patrząc z perspektywy badawczej, jest to problem bezpośrednio związany z identyfikacją zależności przyczynowo-skutkowych. Zadanie ich ustalania leży w gestii badań ewaluacyjnych, które definiowane są jako przedsięwzięcia *zmierzające do określenia w oparciu o właściwie zgromadzone i przetworzone informacje, w jakim stopniu dane rozwiązanie (np. interwencja publiczna: polityka, program lub projekt) spełnia ustalone kryteria, w tym, w szczególności, w jakim stopniu osiągnęło cele, dla realizacji których zostało podjęte oraz jakie są relacje pomiędzy nakładami, działaniami i wynikami tego rozwiązania* (Górniak, 2007: 11). Ewaluacja ma za zadanie dostarczać wiedzy na temat skuteczności realizowanych interwencji oraz formułować wnioski, które przydatne będą podczas wdrażania kolejnych działań. Postuluje się, aby ewaluacja, spełniając postawione przed nią zadania, wpisywała się w cykl polityk publicznych, mając wpływ na podejmowane w ramach ich realizacji decyzje (Górniak, 2007: 11). Aby wiedza gromadzona w wyniku prowadzonych ewaluacji była użyteczna, muszą się one opierać na twardych podstawach, wyznaczanych przez rygory prowadzenia badań społecznych. Ewaluacja, która ma na celu ocenę skuteczności działań, nie może abstrahować od problematyki przyczynowości ani rezygnować z prób identyfikacji związków przyczynowych.

Wyznaczone dla ewaluacji zadania stanowią duże wyzwanie dla badaczy społecznych, bowiem uchwycenie relacji przyczynowych jest z natury rzeczy niezwykle trudne. Do dyspozycji jest jednak coraz większy arsenał metod i technik badawczych, które w założeniu mają pomagać w identyfikacji tych zależności. Poniższa praca prezentuje jedną z takich technik.

* * *

Zasadniczym celem niniejszego opracowania jest przedstawienie techniki statystycznej, która dotychczas funkcjonuje w Polsce pod anglojęzyczną nazwą *propensity score matching* (PSM). Technika ta należy do większej grupy metod wykorzystywanych w tzw. badaniach obserwacyjnych[1], stanowiących alternatywę dla badań eksperymentalnych. Eksperymenty, zwłaszcza te, w których wykorzystuje się randomizację, klasycznie już uznawane są za schemat badawczy o największej wartości dowodowej, jeśli chodzi o identyfikację relacji przyczynowych. Liczne ograniczenia, uniemożliwiają powszechne wykorzystanie eksperymentów. Owocem tego jest intensywny rozwój innych podejść badawczych, które również mają na celu wyjaśnienie związków przyczynowo-skutkowych (Rosenbaum, 2002: 1).

W badaniach obserwacyjnych dokonuje się porównania dwóch grup, z których jedna – podobnie jak w badaniach eksperymentalnych – określana jest mianem grupy eksperymentalnej, a druga – kontrolnej[2]. Grupa eksperymentalna objęta jest oddziaływaniem pewnego bodźca, zwanego czynnikiem eksperymentalnym (ang. *treatment*). Grupa kontrolna pozbawiona jest tego oddziaływania i jako taka ma służyć dla grupy eksperymentalnej za grupę porównawczą. Czynnikiem eksperymentalnym może być np. realizacja programu społecznego, podanie leku, emisja reklamy itp. Celem badań obserwacyjnych jest ustalenie efektu, będącego wynikiem oddziaływania czynnika eksperymentalnego, czy inaczej określenie związku przyczynowo-skutkowego pomiędzy bodźcem, jakiemu poddana została grupa eksperymentalna, a przewidywanym efektem – obserwowanym lub nie.

Zasadnicza różnica pomiędzy badaniami obserwacyjnymi i eksperymentalnymi polega na tym, że w tych pierwszych badacz nie ma kontroli nad procesem doboru jednostek do grupy eksperymentalnej. Zwykle jest ona utworzona, zanim badacz się pojawi. W badaniach eksperymentalnych to badacz ma bezpośredni wpływ na proces selekcji jednostek do grupy eksperymentalnej i grupy kontrolnej. Dobór jednostek do obu grup odbywa się zazwyczaj z wykorzystaniem mechanizmu losowego, nad którym czuwa badacz. W wyniku wykorzystania randomizacji utworzone grupy są bezpośrednio porównywalne. Jedyne, co je różni, to fakt występowania bodźca. Taka sytuacja zwykle nie ma miejsca w przypadku badań obserwacyjnych, w których grupa eksperymentalna najczęściej różni się od populacji jednostek kontrolnych. Tym samym wnioskowanie na temat związków przyczynowo-skutkowych na podstawie bezpośrednich porównań obu grup jest nieuprawnione.

Technika PSM jest narzędziem, które – w dużym uproszczeniu – pozwala dobrać grupę kontrolną tak, by była ona możliwie jak najbardziej podobna do grupy eksperymentalnej. Ma więc za zadanie naśladować działanie mechanizmu losowego stosowanego w eksperymentach.

W niniejszej pracy przedstawiono teoretyczne ramy techniki PSM, jej założenia metodologiczne oraz przykłady praktycznego zastosowania. Tłem dla prezentacji techniki jest ogólna charakterystyka metody eksperymentalnej oraz przegląd innych metod, wykorzystywanych w badaniach obserwacyjnych do pomiaru oddziaływania interwencji.

Rozdział pierwszy stanowi teoretyczne wprowadzenie do problematyki wnioskowania przyczynowego w badaniach obserwacyjnych. Zaprezentowano tu podejścia metodologiczne stosowane

[1] Zwanych również nieeksperymentalnymi. W literaturze można spotkać się również z klasyfikacją, w której technikę PSM lokuje się w obszarze tzw. badań quasi-eksperymantalnych.

[2] Przyjęło się, pisząc o badaniach obserwacyjnych, wykorzystywać terminologię stosowaną w badaniach eksperymentalnych, tak będzie również w niniejszej pracy.

w obszarze identyfikacji i pomiarze oddziaływania przyczynowego. Szerzej omówiono podejście eksperymentalne. Przedstawiono charakterystykę technik opartych na schemacie prób dopasowanych wg cech oraz samą technikę PSM wraz ze wskazaniem kluczowych dla niej założeń.

Rozdział drugi zawiera formalny opis techniki PSM. Zdefiniowana została procedura praktycznego jej stosowania. Scharakteryzowane zostały kluczowe momenty i decyzje, które należy podjąć, wykorzystując technikę PSM. Przedmiotem zainteresowania na tym etapie są również wymagania wobec danych niezbędnych do realizacji analiz z wykorzystaniem techniki PSM. Przybliżono też sposoby oceny doboru grupy kontrolnej. Na koniec poruszona jest kwestia ograniczeń techniki PSM.

W rozdziale trzecim przytoczono przykłady praktycznego wykorzystania techniki PSM. Opisany został przykład pierwszego polskiego zastosowania techniki w badaniu oddziaływania programów rynku pracy.

Publikacja została podzielona i opracowana w taki sposób, aby poszczególne rozdziały wzajemnie się uzupełniały. Większość kluczowych dla przedstawianej problematyki pojęć została wprowadzona i objaśniona w części pierwszej opracowania.

* * *

Niniejsza publikacja powstała na bazie doświadczeń i wiedzy, które autor zdobył podczas badań ewaluacyjnych realizowanych na zlecenie Polskiej Agencji Rozwoju Przedsiębiorczości (PARP). W latach 2006–2007 kierował kilkoma badaniami ewaluacyjnymi prowadzonymi przez PARP, w których wykorzystana została technika PSM. Badania te poświęcone były realizacji programów współfinansowanych ze środków przedakcesyjnego funduszu Phare.

Początek wykorzystania techniki PSM w PARP wiąże się z poszukiwaniem metod, czy też podejść badawczych, które pomogłyby lepiej przyjrzeć się skutkom realizowanych interwencji. Z nieocenioną pomocą przyszedł wtedy PARP dr hab. Jarosław Górniak, profesor Uniwersytetu Jagiellońskiego, który przybliżył istotę badania związków przyczynowych i zaproponował konkretne narzędzie badawcze, właśnie w postaci techniki PSM. Bardzo cenna okazała się również pomoc i wiedza profesora Górniaka w zakresie statystycznej analizy danych już podczas implementacji techniki, tj. w momencie planowania i realizacji badań. Należy podkreślić, że jego wsparcie było warunkiem koniecznym dla wykorzystania techniki PSM w praktyce badawczej PARP.

Pierwsze badanie, w którym zastosowano technikę PSM zrealizowane zostało w PARP w roku 2006 r. Była to ewaluacja projektów realizowanych w ramach dwóch edycji Phare SSG (2001, 2002) – komponent Rozwój Zasobów Ludzkich. Wykonawcą badania był Instytut PBS DGA. Ewaluacja miała za zadanie dokonać oceny efektu przyczynowego zrealizowanych na dużą skalę programów rynku pracy. Bardzo ważnym elementem realizacji badania było nawiązanie współpracy PARP z Ministerstwem Pracy i Polityki Społecznej, co zaowocowało uzyskaniem i wykorzystaniem w analizach danych pochodzących z rejestrów osób bezrobotnych. Niemal w tym samym czasie, co ewaluacja Phare SSG RZL, prowadzona była ewaluacja działań realizowanych w ramach Phare SSG 2002, tyle że w komponencie MSP. Celem badania była ocena oddziaływania programów przedakcesyjnych na sektor polskich przedsiębiorstw. Wykonawcą ewaluacji była firma Pentor. Badanie to, w odróżnieniu od ewaluacji Phare SSG RZL, bazowało na danych pierwotnych, uzyskanych w drodze badań terenowych. W roku 2007 PARP zleciło

kolejne dwa badania ewaluacyjne, w których wykorzystano technikę PSM. Podobnie jak wcześniej, analizami objęto programy finansowane ze środków Phare (zarówno w komponencie RZL, jak i MSP).

Obecnie realizowana jest ewaluacja działań finansowanych ze środków funduszy strukturalnych, w ramach Sektorowego Programu Operacyjnego Wzrost Konkurencyjności Przedsiębiorstw (SPO WKP). Ewaluacja prowadzona jest przy współpracy z Głównym Urzędem Statystycznym, która zaowocowała wykorzystaniem w ewaluacji danych sprawozdawczych.

Opracowania polskojęzyczne na temat techniki PSM są na tę chwilę jeszcze nieliczne. Wśród tych, które się pojawiły, wskazać należy dwie pozycje wydane przez PARP. W pierwszej z nich (*Ewaluacja ex-post. Teoria i praktyka badawcza*[3], 2007), dr Roman Konarski oraz Michał Kotnarowski przybliżają problematykę wykorzystania techniki PSM na przykładzie badania ewaluacyjnego zrealizowanego na zlecenie PARP w roku 2006. W drugiej z pozycji (*Środowisko i warsztat ewaluacji*[4], 2008) dr Paweł Strawiński przekrojowo prezentuje technikę PSM, wykorzystując przy tym sztucznie wygenerowane dane. Niniejsze opracowanie było w pierwotnym kształcie przedmiotem pracy magisterskiej, napisanej i obronionej w Instytucie Socjologii Uniwersytetu Warszawskiego. Znaczący wpływ na jej zawartość miał promotor pracy – dr hab. Krzysztof Koseła, prof. UW oraz jej recenzent, którym był dr hab. Grzegorz Lissowski, prof. UW.

[3] http://www.parp.gov.pl/index/more/2046
[4] http://www.parp.gov.pl/index/more/9658

Problematyka ustalania efektu przyczynowego

Realizując dowolne działanie – nieważne, czy jest to interwencja w postaci programu społecznego, kampania reklamowa produktu czy też zwykłe zażycie leku – racjonalnie postępujący podmiot zmierza zazwyczaj do osiągnięcia określonego efektu. Może nim być odpowiednio wzrost zatrudnienia w grupie objętej danym programem, zwiększenie sprzedaży reklamowanego produktu lub uśmierzenie bólu. W zależności od wielu czynników związanych z samym działaniem – jego adekwatnością, konstrukcją, sposobem realizacji – jak również w zależności od szeregu innych czynników zewnętrznych, zamierzony efekt może zostać osiągnięty lub nie. Jeśli tylko ów efekt jest mierzalny, można w oparciu o zwykłą obserwację lub – w przypadku bardziej złożonych działań – realizując badanie empiryczne, wykazać stopień, w jakim został osiągnięty.

Badanie efektów czy też skutków różnego rodzaju działań, interwencji, zdarzeń (ang. *treatments*) jest uniwersalnym problemem, którym od dawna zajmują się rozmaite gałęzie nauki. Nieodłączną częścią tej problematyki jest zagadnienie przyczynowości, a dokładniej kwestia ustalania związków o naturze przyczynowo-skutkowej. W praktyce chodzi o to, aby oprócz informacji na temat poziomu osiągniętego efektu, wiedzieć na ile jest on rzeczywistym wynikiem podjętego wcześniej działania. Z zapotrzebowania na tę wiedzę wynika dążenie do ustalenia tzw. efektu przyczynowego, czyli efektu, który ma zdawać sprawę z faktycznego, tj. uwzględniającego zależności przyczynowe, oddziaływania jednych zdarzeń na drugie. Dlaczego podejmuje się próby ustalania zależności przyczynowych? Przede wszystkim nigdy nie można wykluczyć sytuacji w pewnym sensie skrajnych, w których np. zamierzony efekt wystąpiłby niezależnie od podjętego działania. Biorąc za przykład program społeczny, możliwe jest, że zaobserwowany po nim spadek poziomu bezrobocia w grupie uczestników (wyznaczony cel interwencji), może nie mieć związku z realizacją tego programu, będąc wynikiem tylko i wyłącznie ogólnej poprawy sytuacji na rynku pracy. Wystąpienie takiej okoliczności pociąga za sobą wiele konsekwencji, w tym tę najważniejszą, polegającą na zakwestionowaniu zasadności podjętego działania. W końcu, po co realizować wybraną interwencję, angażować określone zasoby, skoro ten sam efekt można osiągnąć nie robiąc nic. Czasem może nawet zdarzyć się tak, że podjęte działanie wywoła skutek wprost przeciwny do zamierzonego. Przykładem może być tu przyjęcie leku, który zamiast eliminować chorobę, prowadzi do dodatkowych powikłań. Z drugiej strony, wykazanie, że pożądany i osiągnięty efekt jest rzeczywistym wynikiem wcześniej podjętego działania, może być argumentem na rzecz jego dalszej aplikacji. Aby jednak móc wyciągać takie czy inne wnioski, trzeba najpierw dysponować wiedzą na temat tego, jakie związki łączą badane działanie i obserwowany (lub nie) efekt. W tym celu dokonuje się tzw. oceny wpływu (ang. *impact assessment*), która ma za zadanie ustalić, czy dane działania rzeczywiście produkują zamierzone skutki. Ocena ta, jak zostało już nadmienione, nierozerwalnie połączona jest z problematyką przyczynowości[5].

[5] Oczywiście nie należy postrzegać problematyki przyczynowości tylko i wyłącznie w kontekście weryfikacji skuteczności działań. Przyczynowość jest z natury rzeczy znacznie pojemniejszym zagadnieniem. Związki przyczynowe mogą istnieć w szczególności tam, gdzie nikt ich się nie spodziewa. Ich identyfikacja w takim przypadku jest zwykle dużym wyzwaniem dla badaczy. Poszukiwanie regularności, potrzeba odkrywania stałych zależności, praw, jest odwiecznym motorem napędowym nauki. W niniejszej pracy główna uwaga zwrócona jest jednak na weryfikację istnienia relacji przyczynowych tam, gdzie oczekuje się ich obecności. Przykładem będą tu ewaluacje interwencji publicznych realizowanych dla osiągnięcia pewnego celu – zmniejszenia bezrobocia, wykluczenia społecznego, podniesienia konkurencyjności przedsiębiorstw itp. Również prezentowana tu technika – *propensity score matching* – zakłada, że badacz ją wykorzystujący wie, co może być potencjalnym skutkiem ocenianego

Poczynając od rozważań Arystotelesa, pojęcie przyczynowości i związku przyczynowego ma długą historię w filozofii i nauce, niemniej do dziś przysparza wielu kłopotów teoretycznych, a co za tym idzie również praktycznych. Przyjmuje się, że uwarunkowanie przyczynowe jest szczególnym rodzajem związku między zdarzeniami. Nie ma jednak ogólnej zgody między teoretykami tego zagadnienia, czym tak naprawdę jest owa przyczyna i jakiego rodzaju związki zasługują na miano związków przyczynowych (Karpiński, 1985: 7). Jak pisze Antoni Sułek: *Pojęcie związku przyczynowego jest jednym z bardziej niejasnych w słowniku nauki. Niegasnącą żywotność wykazują też spory o status zasady przyczynowości* (Sułek, 1979: 23). Powody takiego stanu rzeczy są rozmaite. Głównym jest jednak zapewne ten, że związek przyczynowy jest pojęciem czysto teoretycznym – nie jest bezpośrednio obserwowany. W praktyce, badacz-empiryk może najczęściej zaobserwować jedynie współwystępowanie pewnych zdarzeń lub co najwyżej ich następstwo. Badacz może np. zarejestrować, że po realizacji programu społecznego część z jego uczestników znalazła pracę. Jednak nie można wykluczyć takiej sytuacji, w której obu zdarzeń (realizacji programu i posiadania pracy przez jego uczestników) nie łączy żaden związek – obserwowany efekt może być bowiem autonomiczny w stosunku do podjętego działania. O występowaniu związku przyczynowego nie można wnioskować ani z samego faktu współwystępowania zdarzeń (występowanie korelacji nie oznacza zaistnienia związku przyczynowego[6]), ani (na tej samej zasadzie) z ich następstwa. Warto jednak zauważyć, że z oczywistych względów są to warunki konieczne dla zaistnienia związku przyczynowego. Uwzględniając ten fakt, najogólniej o związku przyczynowym można powiedzieć, że jest on *takim następstwem zdarzeń, w którym jedno zdarzenie wywołuje, produkuje czy generuje drugie* (Sułek, 1979: 25).

Kontrfaktyczna definicja przyczynowości

Na szczególną uwagę, jeśli chodzi o problematykę przyczynowości, zasługują prace dwudziestowiecznych statystyków i ekonometryków, takich jak R. Fisher, J. Neyman, W. Cochran, D. Cox, J. Heckman, A.D. Roy, R. Quandt, czy D. Rubin (Winship i in., 1999: 662). Niezależnie prowadzone na tych dwóch polach nauki rozważania zaowocowały wypracowaniem wspólnego podejścia do problematyki przyczynowości. Opiera się ono na pochodzącej z logiki koncepcji tzw. stanów kontrfaktycznych (ang. *counterfactual framework*). Koncepcję tę najprościej można scharakteryzować poprzez wskazanie podstawowego pytania, na które szuka się przy jej pomocy odpowiedzi. Jest to pytanie o to: *co by się stało z pewnym X, gdyby zamiast zdarzenia Y zaszło zdarzenie Z*? Swoje filozoficzne korzenie koncepcja stanów kontrfaktycznych ma natomiast w pracach Davida Hume'a, który w rozważaniach na temat przyczyny pisał: *można zdefiniować przyczynę jako przedmiot, po którym następuje przedmiot inny, przy czym, gdyby*

działania, a przynajmniej ma na ten temat hipotezę. Technika PSM pozwoli ją zweryfikować, nie wskaże jednak na istnienie związków przyczynowych, których nie spodziewano się a priori. Dla ilustracji weźmy klasyczny przykład wpływu palenia papierosów na prawdopodobieństwo zachorowania na raka płuc. Jeszcze kilkadziesiąt lat temu nie było wcale oczywiste, że istnieje zależność między tymi dwoma zdarzeniami. Jedno nie implikowało w sposób bezsprzeczny drugiego. Z czasem pojawiły się hipotezy występowania takiego związku. Ich weryfikacje umożliwia m.in. technika PSM. Powyższe, dosyć ogólne zarysowanie dwóch podejść do problematyki przyczynowości znajduje wyraz w pojawiającym się w literaturze przedmiotu subtelnym rozróżnieniu anglojęzycznych zwrotów: *effects of causes* i *causes of effects* (Holland, 1986: 945). Tłumacząc wprost, w pierwszym przypadku chodzi o badanie efektów działań (można powiedzieć, że jest to węższe podejście w problematyce przyczynowości), w drugim o szukanie przyczyn obserwowanych efektów (szersze podejście). Niniejsza praca w całości poświęcona jest pierwszemu z podejść.

[6] Wnioskowanie takie jest wykluczone, ze względu na możliwość występowania związków pozornych między zdarzeniami (Blalock, 1977: 376).

nie było przedmiotu pierwszego, drugi nie mógłby był istnieć (Greenland, 2004: 4). Współczesne prace statystyków i ekonometryków zaowocowały operacjonalizacją efektu przyczynowego właśnie w oparciu o koncepcję stanów kontrfaktycznych[7]. Tok rozumowania przyjmowany dla tego podejścia[8] zostanie przybliżony poniżej.

Dana jest populacja jednostek *I*[9]. W danym czasie, każda jednostka *i*, pochodząca z populacji *I*, może znaleźć się w jednej z dwóch[10] sytuacji, które wyraża zmienna $D \in \{0, 1\}$. I tak, dana jednostka może zostać objęta oddziaływaniem wybranego zdarzenia/działania/bodźca – wtedy $D_i = 1$ – lub może znaleźć się w grupie wykluczonej z obszaru oddziaływania tego zdarzenia/działania/bodźca – wtedy $D_i = 0$ (w języku eksperymentu jednostka może należeć odpowiednio do grupy eksperymentalnej lub grupy kontrolnej, zwanej czasem grupą odniesienia). Każdej sytuacji, w jakiej może znaleźć się jednostka, odpowiada potencjalny skutek/wynik, wyrażany przez zmienną *Y*. W zależności od tego, w jakiej grupie znalazła się jednostka, *Y* może przyjąć dla niej jedną z dwóch wartości: Y_{i1} lub Y_{i0}, gdzie Y_{i1} jest wartością, która zostałaby zaobserwowana, gdyby jednostka znalazła się w grupie eksperymentalnej – wystawionej na działanie bodźca – zaś Y_{i0} odpowiada wartości, która zostałaby zaobserwowana, gdyby jednostka znalazła się w grupie kontrolnej. W zależności od tego, do której grupy należy jednostka, jeden z wyników Y_{i1} lub Y_{i0} jest wynikiem hipotetycznym – nieobserwowanym w rzeczywistości. Przyjmuje się jednak, że jednostki mają „przypisane" potencjalne wyniki dla każdego z dwóch stanów: dla tego, w którym się rzeczywiście znalazły oraz dla tego, w którym mogłyby się znaleźć. Innymi słowy, każda jednostka z grupy eksperymentalnej i z grupy kontrolnej ma „przypisany" obserwowalny efekt, ale także nieobserwowalny efekt kontrfaktyczny[11]. Mówiąc ogólniej, zakłada się, że osoby mają przyporządkowane wyniki do wszystkich potencjalnych stanów, w jakich mogą się znaleźć. Przypuśćmy np., że chcemy poznać wpływ posiadania wyższego wykształcenia na wysokość zarobków. Przyjmijmy zatem, że $D_i = 1$ odpowiada sytuacji ukończenia przez osobę *i* studiów wyższych, zaś $D_i = 0$, sytuacji ich nie ukończenia. *Y* to wysokość zarobków.

Dla wybranej osoby jednostkowy efekt przyczynowy to różnica[12] pomiędzy Y_{i1} a Y_{i0} (Holland, 1986: 947), czyli w powyższym przykładzie jest to różnica pomiędzy zarobkami osoby, która posiada wyższe wykształcenie, a zarobkami tej samej osoby, gdyby wyższego wykształcenia nie miała. W przypadku, gdy osoba rzeczywiście posiada wyższe wykształcenie, jest to porównanie efektu faktycznego, obserwowalnego – zaistniałego po danym zdarzeniu – z efektem hipotetycznym, nieobserwowalnym, kontrfaktycznym, który zaistniałby w sytuacji przeciwnej do tej, która faktycznie miała miejsce.

Warto w tym miejscu zauważyć, że przy tak zdefiniowanym efekcie przyczynowym, zakłada się, że udział innych osób w danej interwencji nie ma wpływu na indywidualne wyniki jednostki (Winship i in., 1999: 663). Innymi słowy obserwowany efekt dla osoby biorącej udział w interwencji zależy tylko i wyłącznie od tej osoby. Założenie to, znane pod nazwą *Stable Unit Treatment Assumption* (SUTVA), będzie niespełnione, np. gdy jednostki współzawodniczą między sobą o zasoby (Rubin, 1983: 41). Zało-

---

7 W literaturze można spotkać się z opinią, że jest to kontrfaktyczna definicja przyczynowości (Winship i in., 1999: 662).

8 Prezentowane tu podejście znane jest jako przyczynowy model Rubina (ang *Rubin Causal Model*) (Holland 1986: 946).

9 Zwyczajowo jednostkami mogą być osoby, instytucje, firmy, ale także i grupy jednostek, regiony itp.

10 Jest to oczywiście sytuacja najprostsza, w rzeczywistości może to być nieskończona liczba wykluczających się stanów.

11 Donald Rubin postuluje, aby posługiwać się pojęciem potencjalnych wyników (ang. *potential outcomes*). Pojęcie to ma swoje korzenie w pracach Jerzego Neymana (Rubin, 2005).

12 Dopuszczalne są inne porównania, np. stosunek wyników.

żenie to nie utrzyma się więc w sytuacji, gdy w grę wchodzić będzie oddziaływanie „makroefektów". Za przykład może tu posłużyć realizacja dużego programu szkoleniowego, który jest zainicjowany w mieście z konkurencyjnym rynkiem pracy. Wraz ze wzrostem podaży pracy, tj. wraz z ilością osób kończących program, płaca na rynku, jaką oferować będą pracodawcy uczestnikom programu, będzie najprawdopodobniej spadać. Z drugiej strony założenie powinno być utrzymane w sytuacji, gdy rozmiar interwencji jest relatywnie mały w stosunku do wielkości obszaru, na którym interwencja jest realizowana (Winship i in., 1999: 663).

Oczywiście w praktyce, zaobserwowanie w danym czasie, dla tej samej jednostki, skutków dwóch (lub więcej) wykluczających się zdarzeń – Y_{i1} i Y_{i0} – jest niemożliwe. Nie da się jednocześnie posiadać i nie posiadać wyższego wykształcenia. W literaturze przedmiotu, sytuacja ta nosi nazwę fundamentalnego problemu wnioskowania przyczynowego[13] i wskazuje, że wnioskowanie przyczynowe jest niemożliwe do przeprowadzenia wprost. Opisaną powyżej zależność można wyrazić następująco:

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0} \qquad D \in \{0, 1\} \tag{1.01}$$

Zmienna D może przyjąć w rzeczywistości jedną z dwóch wartości – zero w przypadku, gdy jednostka należy do grupy kontrolnej lub jeden w sytuacji, gdy jednostka należy do grupy eksperymentalnej. Dlatego też zaobserwowany może być tylko jeden z wyników znajdujących się po prawej stronie powyższego równania (Y_{i0} lub Y_{i1}).

Jak zatem można poradzić sobie z fundamentalnym problemem wnioskowania przyczynowego i jak oszacować efekt przyczynowy danego działania? Brak informacji na temat Y_{i0}, w warunkach, gdy znamy Y_{i1}, można po prostu potraktować jako problem braku danych (Rubin i in., 1983; Heckman i in. 1997: 608), sposobem na jego rozwiązanie może być zaś ich imputacja. Zakłada się, że można tę czynność dokonać, wykorzystując jednostki niepoddane oddziaływaniu danego bodźca, u których wynik braku interwencji (Y_{i0}) jest obserwowany.

Między innymi Holland (Holland, 1986: 947), za swymi poprzednikami przytacza to – tzw. statystyczne – rozwiązanie, przenosząc problem z poziomu jednostki na poziom populacji, z której dana jednostka pochodzi. Tok rozumowania jest tu następujący. Przyjmijmy, że Y_{ATE} będzie tzw. średnim efektem przyczynowym (ang. *Average Treatment Effect – ATE*), określonym dla populacji *I*. Skoro tak, to zgodnie z tym, co zostało przedstawione wcześniej w odniesieniu do jednostkowego efektu przyczynowego:

$$Y_{ATE} = \mathrm{E}(Y_1 - Y_0) \tag{1.02}$$

Powyższe równanie może być również zapisane jako:

$$Y_{ATE} = E(Y_1) - E(Y_0) \tag{1.03}$$

gdzie $E(Y_1)$ jest wartością średnią efektu działania w sytuacji, gdy wszystkie jednostki w populacji *I* zostały wystawione na jego oddziaływanie, zaś $E(Y_0)$ jest wartością średnią efektu w sytuacji, gdy wszystkie jednostki z populacji *I* znalazły się w grupie kontrolnej. W praktyce $E(Y_1)$ oraz $E(Y_0)$ nie są jednocześnie

[13] Ang. *Fundamental Problem of Causal Inference* (Holland, 1986: 947).

znane, bowiem tylko część jednostek jest zwykle uczestnikami wybranego działania i tym samym tylko część należy do grupy porównawczej. Powyższe równanie wskazuje jednak, że informacja o jednostkach znajdujących się w różnych stanach może być wykorzystana do oszacowania Y_{ATE}. Jeśli np. część jednostek została wystawiona na oddziaływanie bodźca, to mogą one być wykorzystane do oszacowania $E(Y_1)$ (ponieważ jest to wartość średnia Y_1 określona dla populacji *I*), z kolei jeśli inne jednostki nie zostały wystawione na oddziaływanie badanego zdarzenia, to mogą one zostać wykorzystane do oszacowania $E(Y_0)$. Tym samym efekt przyczynowy, na poziomie wybranej populacji i pod pewnymi warunkami, o których mowa będzie dalej, jest możliwy do oszacowania.

Od razu należy jednak zauważyć, że miara wyrażana przez Y_{ATE} ma istotne ograniczenie. Przedstawia ona efekt działania dla przeciętnej, losowo wybranej jednostki pochodzącej z populacji *I*, bez uwzględnienia, czy została ona objęta danym oddziaływaniem czy nie. W przypadku oceny konkretnych działań, przedmiotem zainteresowania badacza jest raczej efekt przyczynowy ograniczony tylko do jednostek, które uczestniczyły w ocenianej interwencji. Poszukiwany jest tym samym tzw. przeciętny efekt oddziaływania na jednostki poddane oddziaływaniu (ang. *treatment on treated effect – ATT*). Efekt ten można wyrazić następująco:

$$Y_{ATT} = E(Y_1 - Y_0 | D = 1) = E(Y_1 | D = 1) - E(Y_0 | D = 1) \tag{1.04}$$

gdzie $E(Y_1 | D = 1)$ to średni wynik obserwowany po interwencji w grupie jednostek poddanych oddziaływaniu, zaś $E(Y_0 | D = 1)$ to średni wynik braku interwencji dla grupy poddanej oddziaływaniu. $E(Y_0 | D = 1)$ jest oczywiście efektem nieobserwowalnym – kontrfaktycznym – niemniej, tak jak w przypadku $E(Y_1)$ lub $E(Y_0)$, może on zostać oszacowany. W rzeczywistości badaczowi dane jest[14] $E(Y_0 | D = 0)$, czyli średni efekt obserwowany po interwencji dla jednostek, które w niej nie uczestniczyły. Do oszacowania $E(Y_0 | D = 1)$ można więc wykorzystać $E(Y_0 | D = 0)$. Pytaniem zasadniczym jest, na ile ten pierwszy efekt odpowiada drugiemu?

W przypadku, gdyby wszystkie jednostki w zbiorowości *I* były identyczne, zachodziłaby następująca równość: $E(Y_0 | D = 0) = E(Y_0 | D = 1)$. W takiej też sytuacji efekt przyczynowy można by było zdefiniować po prostu jako różnicę pomiędzy wynikiem uczestnika danego zdarzenia i wynikiem jakiejkolwiek osoby z puli kontrolnej. Innymi słowy, nie występowałby fundamentalny problem wnioskowania przyczynowego – samo wnioskowanie można by było przeprowadzać na poziomie poszczególnych jednostek obserwacji. W praktyce jednak – szczególnie w materii społecznej – taki przypadek zwykle nie występuje. Jednostki z grupy będącej pod wpływem wybranego działania, podobnie jak jednostki nim nie objęte, mają rozmaite cechy – zarówno obserwowane (takie jak płeć, wiek, wykształcenie itp.), jak i nieobserwowane (np. poziom determinacji, motywacji itp.). Cechy te z kolei mogą pozostawać zarówno w związku z prawdopodobieństwem uczestnictwa podmiotu w danej interwencji, jak i z później obserwowanym wynikiem, a więc z obserwowanym efektem, wyrażanym przez zmienną *Y*. Inaczej mówiąc, efekt działania może silnie zależeć od tego, jaka jest dystrybucja cech wpływających na zmienną *D* i jednocześnie zmienną *Y*, w grupie osób poddanych i niepoddanych oddziaływaniu. Jeśli dystrybucja tych zmiennych jest różna, proste porównanie efektów obserwowanych w tych dwóch grupach będzie nieuprawnione. W takich sytuacjach $E(Y_0 | D = 0)$ będzie więc niewłaściwym oszacowaniem $E(Y_0 | D = 1)$. Źródłem wskazanego problemu są mechanizmy selekcji, odpowiadające za to, które jed-

[14] W rozumieniu: możliwe do pozyskania.

nostki zostaną przypisane do grupy interwencji. Bezpośrednią konsekwencją działania mechanizmów selekcji jest występowanie różnic pomiędzy tymi jednostkami, które znalazły się w grupie eksperymentalnej, a tymi, które znalazły się w poza nią. W rezultacie może to prowadzić do uzyskania oszacowań wartości efektu przyczynowego, które będą obciążone[15]. Dla zobrazowania problemu warto posłużyć się przykładem programów rynku pracy. Praktyka pokazuje, że w czasie realizacji programów, których celem jest np. ograniczenie bezrobocia[16], rzadko kiedy grupa objęta pomocą jest porównywalna z grupą osób, która wsparciem jest nieobjęta. Może wystąpić szereg okoliczności, które mają wpływ na tę sytuację. W przypadku programów woluntarystycznych, a więc takich, w których udział jest dobrowolny, często spotykane jest tzw. zjawisko samoselekcji. Polega ono na tym, że jednostki bardziej aktywne, zdeterminowane i zmotywowane do podjęcia pracy, częściej stają się odbiorcami pomocy oferowanej w ramach tego typu programów. Z tego względu osoby te reprezentują tę część populacji, która w porównaniu z osobą przeciętną może mieć wyższe prawdopodobieństwo osiągnięcia „sukcesu" po otrzymaniu wsparcia. Podobne skutki ma zjawisko znane pod pojęciem efektu „spijania śmietanki" (ang. *creaming effect*). Efekt ten jest konsekwencją wpływu osób trzecich, które w związku z rolą, jaką pełnią, mogą decydować o tym, kto będzie objęty danym działaniem. Na przykład osoby odpowiedzialne za rekrutację osób do programów pomocowych (pracownicy publicznych służb zatrudnienia – powiatowych urzędów pracy; przedstawiciele firm realizujących szkolenia itp.), mogą do niego przyjmować takie osoby bezrobotne, które „dobrze rokują", tzn. posiadają cechy, które umożliwią im szybsze znalezienie zatrudnienia. I tak, do programu mogą być przyjmowane osoby z wyższym wykształceniem, większym doświadczeniem zawodowym, bardziej zmotywowane itp. Dzięki temu cele ilościowe programu (najczęściej – liczba osób, która ma znaleźć po programie pracę) mogą zostać łatwiej osiągnięte, z czego zarządzający danym programem mogą się potem pozytywnie rozliczyć. Z drugiej strony znane jest również zjawisko przeciwne do efektu „spijania śmietanki", jest nim efekt „kwaszenia" (ang. *souring*), pojawiające się również pod terminem *triaging*[17]. Zjawisko to polega na tym, że do grupy objętej pomocą rekrutuje się „trudne" jednostki. W przypadku programów rynku pracy będą to np. osoby z poważnymi problemami – niepełnosprawni, długotrwale bezrobotni, cudzoziemcy, bezdomni, byli więźniowie itp. Przykładem takich interwencji mogą być projekty realizowane w ramach Inicjatywy Wspólnotowej *Equal*[18]. Obserwowane po realizacji tego typu programów efekty mogą być niższe niż przeciętnie oczekiwane efekty dla całej populacji osób bezrobotnych. W takich przypadkach trudno jest powiedzieć, czy to program nie zadział tak, jak powinien, czy wręcz przeciwnie – obserwowane efekty znacząco przewyższają stan, jaki miałby miejsce, gdyby nie podjąć się realizacji danej interwencji.

Problematyka mechanizmów selekcji stanowi istotny element w procesie oceny oddziaływania wybranych programów. Nieuwzględnienie mechanizmów selekcji podczas szacowania efektów działań, a więc wykorzystywanie wyniku $E(Y_0 \mid D = 0)$ do oszacowania $E(Y_0 \mid D = 1)$ bez wprowadzenia żadnych korekt, może prowadzić do uzyskania błędnych oszacowań efektu przyczynowego. Zobrazować to można, rozwijając wykorzystywaną do szacowania *ATT* różnicę $E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0)$ w następujący sposób:

---

15 Jest to tzw. obciążenie selekcyjne (ang. *selecion bias*).

16 Np. poprzez podniesienie kwalifikacji zawodowych osób poszukujących pracy.

17 Selekcja i świadczenie usługi opiera się na wielkości potrzeb klienta (Guo, 2006: 9). Termin ten pojawia się również w medycynie, w kontekście stosowania procesu „porządkowania" pacjentów typowanych do leczenia w kolejności ustalanej na podstawie pilności udzielenia pomocy.

18 *www.equal.gov.pl* stan na 16.03.2009 r.

$$E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0) =$$
$$\{E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)\} + \{E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)\} \quad (1.05)$$

Wartość pierwszego dużego nawiasu jest faktycznym przedmiotem zainteresowania badacza. Nie da się jej bezpośrednio zmierzyć, bowiem $E(Y_0 \mid D = 1)$ jest efektem nieobserwowanym. Z kolei drugi duży nawias to wartość potencjalnego obciążenia selekcyjnego. Im bardziej $E(Y_0 \mid D = 0)$ różni się od $E(Y_0 \mid D = 1)$, tym większe będzie to obciążenie. A więc im bardziej grupa interwencji różni się od grupy pozostającej poza tą interwencją, tym gorsze będzie oszacowanie efektu przyczynowego przy wykorzystaniu do oszacowania stanu kontrfaktycznego, obserwowanego: $E(Y_0 \mid D = 0)$. Jak zostało nadmienione, w rzeczywistości społecznej spełnienie równości $E(Y_0 \mid D = 0) = E(Y_0 \mid D = 1)$ jest bardzo trudne, stąd też rzadko kiedy wartość obciążenia selekcyjnego równa się zero (Heckman i in., 1995: 88).

Metoda eksperymentalna

Podejściem badawczym, które – przynajmniej teoretycznie – rozwiązuje problem obciążenia wynikającego z działania mechanizmów selekcji, jest metoda eksperymentalna. Zgodnie z ogólną definicją eksperyment to: *zabieg polegający na planowanej zmianie przez badacza jednych czynników w badanej sytuacji, przy równoczesnej kontroli innych czynników, podjęty w celu uzyskania w drodze obserwacji odpowiedzi na pytanie o skutki tej zmiany* (Sułek, 1979: 15). Bezpośrednim celem eksperymentu jest więc poznanie skutków zmiany, czyli inaczej poznanie efektu przyczynowego. Zgodnie z powyższą definicją można wyróżnić dwie istotne składowe eksperymentu. Są nimi manipulacja oraz kontrola. Manipulacja polega na *wprowadzeniu zmian, których następstwa chce się ocenić, czyli bodźców* (Sułek, 1979: 15) i jako taka stanowi osobliwość metody eksperymentalnej. W założeniu badacz zarządza całością badanego procesu – jest on jego inicjatorem i bez jego udziału eksperyment nie zostałyby zrealizowany. Przykładem manipulacji może być uruchomienie programu eksperymentalnego (np. społecznego) skierowanego do wybranej grupy osób. Z kolei kontrola polega na *tworzeniu układów porównawczych w stosunku do tych układów, które poddaje się działaniu bodźców* (Sułek, 1979: 15). Sednem eksperymentu jest znalezienie odpowiedniego układu odniesienia dla badanego zjawiska. W praktyce chodzi o wskazanie lub utworzenie grupy kontrolnej. W założeniu ma ona przedstawiać hipotetyczny stan, jaki miałby miejsce w przypadku, gdyby grupa eksperymentalna nie została objęta badanym działaniem. Grupa kontrolna ma więc za zadanie odzwierciedlać stan kontrfaktyczny. Może ona być dobierana na różne sposoby, istotą jest jednak to, aby w jak najmniejszym stopniu różniła się od grupy eksperymentalnej.

Szczególną postacią kontroli wykorzystywanej w metodzie eksperymentalnej jest randomizacja. Istota tego podejścia polega na wykorzystaniu mechanizmu losowego podczas podziału jednostek na te, które znajdą się w grupie eksperymentalnej i na te, które znajdą się w grupie kontrolnej. Poprzez zastosowanie randomizacji gwarantuje się już na wstępie brak występowania systematycznych różnic między obiema grupami. Ponieważ wszystkie – zarówno obserwowane, jak i nieobserwowane – cechy/czynniki są kontrolowane przez mechanizm losowy, a zmianie podlega tylko czynnik, którego wpływ jest oceniany (zgodnie z klauzulą *ceteris paribus*), to ewentualna różnica w wartościach zmiennej będącej przedmiotem badania[19] może być postrzegana jako bezpośredni skutek wprowadzonej w warunkach eksperymentu zmiany/bodźca.

[19] Abstrahując od błędów pomiaru, artefaktów badawczych itd.

Mamy więc sytuację, w której teoretycznie możliwe jest wykorzystanie wprost $E(Y_0 \mid D = 0)$ do oszacowania $E(Y_0 \mid D = 1)$[20]. Jest to ogromna zaleta, która w obszarze badań nad efektem przyczynowym działań przysporzyła metodzie eksperymentalnej – opartej na planie losowym – tytułu „złotego standardu" (Rossi i in. 1999: 279). Z tego też względu część badaczy rezerwuje miano „prawdziwych eksperymentów" tylko dla tych, w których wykorzystywana jest randomizacja (Sułek, 1979: 18).

Typowy schemat tworzenia grupy eksperymentalnej i grupy kontrolnej sprowadza się do kilku etapów. Zwykle w pierwszej kolejności następuje poinformowanie wybranej populacji o uruchomieniu interwencji – programu eksperymentalnego. Osobom tłumaczone są na tym etapie zasady udziału w programie, w tym występowanie mechanizmu losowego, który determinuje, kto będzie ostatecznie uczestnikiem, a kto znajdzie się w grupie kontrolnej. Następnie przyjmowane są zgłoszenia od osób, które deklarują chęć udział w przedsięwzięciu. Kandydaci wypełniają karty uczestnictwa, a następnie są weryfikowani pod względem spełniania kryteriów formalnych – tego, czy kwalifikują się do udziału w danym programie (często są to takie kryteria, jak wiek, status na rynku pracy itp.). Zespół badawczy na tej podstawie określa populację jednostek spełniających kryteria. Populacja ta dalej stanowi bazę do utworzenia grupy eksperymentalnej i grupy kontrolnej. W kolejnym kroku następuje losowy dobór jednostek do grupy kontrolnej i grupy eksperymentalnej, po czym wszystkie osoby otrzymują informację, w której grupie ostatecznie się znalazły (Orr, 1999: 149).

Pomimo unikalnych zalet metody eksperymentalnej, wiele problemów ogranicza jej wykorzystanie w praktyce badawczej. Część z tych problemów jest szczególnie uciążliwa w przypadku, gdy przedmiotem badań jest materia społeczna. Przyczyny ograniczające możliwość wykorzystania metody eksperymentalnej są dwojakiego rodzaju. Po pierwsze można mówić o czynnikach, które podważają zasadność prowadzenia eksperymentów jako takich. Wymienić tu należy przede wszystkim koszty realizacji eksperymentów, czas ich trwania oraz obecność problemów etycznych. Po drugie terenowa realizacja eksperymentów zagrożona jest całą gamą potencjalnych zakłóceń, których wystąpienie obciąża wyniki badań. Mowa tu przede wszystkim o zaburzeniach w grupie eksperymentalnej i kontrolnej, które mogą pojawić się na etapie implementacji eksperymentu. Wskazane – zwłaszcza ostatnie – problemy powodują, że bardziej uprawnione jest mówienie o eksperymencie, jako o pewnym typie idealnym, do którego badacz dąży w całym postępowaniu. Poniżej nieco szerzej omówione zostały ww. niedoskonałości metody eksperymentalnej.

Podstawową i najbardziej oczywistą przeszkodą na drodze do stosowania eksperymentów są ich wysokie koszty finansowe oraz długi czas realizacji. Wielkość kosztów jest w dużej mierze pochodną czynnika czasu, choć nie mniej istotne znaczenie ma też cały proces gromadzenia danych eksperymentalnych. Jak pisze Larry Orr koszt typowego badania eksperymentalnego to około 2–3 miliony dolarów, choć da się przeprowadzić eksperyment, który będzie znacznie tańszy, jak i znacznie droższy (Orr, 1999: 39). Możliwą skalę przedsięwzięcia niech zobrazuje przykład jednego z eksperymentów, który przeprowadzony został w 1968 roku w Stanach Zjednoczonych (*New Jersey – Pennsylvania Income Maintenance experiment*). Koszt całego badania wyniósł 34 miliony dolarów, a jego realizacja trwała aż 7 lat – licząc od momentu zaprojektowania do chwili publikacji wyników (Rossi i in., 1999: 303). Czynnik czasu, oprócz bezpośredniego przełożenia na kosztochłonność metody eksperymentalnej, ma również drugie ważne znaczenie. Eksperymenty prowadzone są w czasie rzeczywistym – biegną równocześnie z ocenianą

[20] Analogicznie możliwe jest wykorzystanie $E(Y_1 \mid D = 1)$ do oszacowania $E(Y_1 \mid D = 0)$, gdzie $E(Y_1 \mid D = 0)$ to nieobserwowany w rzeczywistości efekt oddziaływania na jednostki nieuczestniczące w programie eksperymentalnym – jest to tzw. *treatment on untreated* (ATU).

interwencją. W związku z tym od momentu decyzji o przeprowadzeniu badania do chwili otrzymania wyników informujących o oddziaływaniu interwencji może minąć nawet kilka lat. Rodzi to ryzyko przedawnienia i utraty pierwotnie zakładanej dla eksperymentów użyteczności. W czasie ich realizacji bowiem mogą zmienić się koncepcje, czy też paradygmaty, w jakich funkcjonują decydenci, mający wpływ na to, w jaki sposób powinno rozwiązywać się problemy społeczne, nie wspominając już o możliwości zmiany samych decydentów. Tym samym może zmienić się zapotrzebowanie na wiedzę, którą przynoszą eksperymenty. Badania eksperymentalne znajdują więc uzasadnienie, o ile tylko ich wyniki będą mogły być użyteczne po kilku latach od momentu powzięcia decyzji o ich przeprowadzeniu. W przeciwnym przypadku realizacja eksperymentów mija się z celem. Z tego też względu metoda eksperymentalna nie znajdzie zastosowania w sytuacjach, które wymagają podjęcia szybkiej decyzji na podstawie wiarygodnych informacji (Rossi i in., 1999: 303).

Inną istotną kwestią, którą podnoszą przeciwnicy eksperymentów są pojawiające się przy ich realizacji problemy etyczne. W przypadku programów pomocowych pozbawienie części jednostek możliwości otrzymania wsparcia płynącego z udziału w interwencji bywa postrzegane jako nieetyczne. Taka percepcja eksperymentów jest szczególnie prawdopodobna, gdy docelową populację stanowią osoby w jakimś sensie upośledzone i rzeczywiście potrzebujące pomocy (np. osoby niepełnosprawne, chore, długotrwale bezrobotne, bezdomni itp.). Wątpliwości natury etycznej pojawiają się również, gdy bodziec ma krzywdzący wpływ na członków grupy eksperymentalnej. Dla przykładu, trudno sobie wyobrazić przeprowadzenie eksperymentu, w którym bada się na losowo dobranej grupie jednostek negatywne skutki palenia tytoniu.

Zastosowanie eksperymentu może być również kłopotliwe w przypadku interwencji publicznych, w których wszelka dowolność w przyznawaniu pomocy zaburza konkurencję rynkową. Przykładem będzie tu udzielanie przedsiębiorcom dotacji bezpośrednich. Firmy, wraz z otrzymaniem wsparcia finansowego, uzyskują przewagę na rynku, kosztem podmiotów pozbawionych dofinansowania. Podobne wątpliwości i wynikające z nich ograniczenia metody eksperymentalnej pojawią się w sytuacjach, w których realizacja interwencji jest kontrolowana przez proces polityczny, tak jak to jest np. w przypadku podejmowanych działań w sferze makroekonomii lub obszarze polityki fiskalnej państwa (Rosenbaum, 2002: 2).

Podsumowując, realizacja eksperymentów może być w pewnych okolicznościach niemożliwa (np. w związku z brakiem funduszy), nieuzasadniona (np. w związku z długim czasem realizacji i tym samym niską użytecznością wyników), czy wręcz nieuprawniona (np. w przypadku wystąpienia problemów natury etycznej).

Drugi zestaw czynników ograniczających wykorzystanie eksperymentów dotyczy aspektu ich terenowej realizacji oraz tego, na ile przyjmowane w metodzie założenia dają się utrzymać w praktyce. Listę problemów pojawiających się przy implementacji eksperymentów otwierają różnego rodzaju zaburzenia mogące pojawić się w grupie kontrolnej (ang. *Control Group Contamination*) (Orr, 1999: 165). Z pozoru paradoksalnym, ale jednak ważkim, jest pytanie o to, czy grupa kontrolna przedstawia rzeczywisty obraz grupy kontrolnej. Innymi słowy, czy sytuacja grupy kontrolnej, obserwowana w obecności występowania programu eksperymentalnego, jest taka, jaka byłaby w przypadku, gdyby eksperyment nie był w ogóle zrealizowany. Jest to więc pytanie o sytuację grupy kontrolnej w stanie kontrfaktycznym. Uzyskanie odpowiedzi na powyższe pytanie jest oczywiście niemożliwe. Warto jednak zauważyć, że badacz wykorzystujący grupę kontrolną do oszacowania sytuacji kontrfaktycznej grupy eksperymentalnej przyjmuje milcząco założenie o tym, że grupa kontrolna nie jest w żaden sposób zaburzona przez

fakt realizacji eksperymentu. W rzeczywistości może się jednak pojawić kilka problemów, które naruszą tę idealną sytuację. Przykładem będzie tu przypadek, w którym personel obsługujący badanie[21] stwierdzi, że należy w jakiś sposób zrekompensować jednostkom kontrolnym brak udziału w programie eksperymentalnym. Taka sytuacja jest całkiem realna, gdy ów program stawia sobie za cel wsparcie osób w trudnej sytuacji życiowej, np. bezrobotnych, wykluczonych społecznie, matki wychowujące samotnie dzieci, osoby upośledzone itp. W literaturze odnotowywane są sytuacje, w których osoby obsługujące eksperyment, z samej chęci bycia pomocnym lub z chęci przekazania czegoś w rodzaju „nagrody pocieszenia", wspomagają jednostki kontrolne, wskazując im np. alternatywy wobec działania eksperymentalnego (Orr, 1999: 165). W takim przypadku jednostki z grupy kontrolnej mogą otrzymać wsparcie, którego nie zyskałyby w przypadku braku eksperymentu. Gdy tak się stanie, jednostki te obciążają szacunki efektu przyczynowego, szacowanego z wykorzystaniem eksperymentu. Możliwość wystąpienia powyższego zaburzenia wynika niewątpliwie z braku znajomości lub zrozumienia założeń metody eksperymentalnej przez osoby uczestniczące w procesie badawczym. Oczywiście można próbować minimalizować występowanie tego typu problemów, np. poprzez zwiększanie świadomości osób obsługujących eksperyment, jednak jest to stosunkowo trudne. Co więcej ewentualny, choć nie gwarantowany, sukces w tej materii nie wykluczy innych możliwych czynników zaburzających obraz grupy kontrolnej. Kolejnym przykładem naruszenia założeń metody eksperymentalnej jest sytuacja, w której jednostki kontrolne, widząc, iż tracą pewne korzyści nie uczestnicząc w programie eksperymentalnym, na własną rękę poszukują alternatywnych środków poprawy własnej sytuacji. Robią to mimo, że bez realizacji eksperymentu nie podjęłyby takich działań[22]. Może wystąpić też sytuacja przeciwna – jednostki, będąc wykluczone z grupy eksperymentalnej mogą zniechęcić się i zrezygnować z działań, które podjęłyby w przypadku braku eksperymentu. Powyższe sytuacje wskazują, że ewentualne obciążenie wyniku grupy kontrolnej może działać w różne strony. Badacz może znaleźć się w tej trudnej sytuacji, że nie będzie wiedział o występowaniu potencjalnego obciążenia lub nie będzie znał jego kierunku.

Równie problematyczne dla poprawności wnioskowania na podstawie danych eksperymentalnych jest to, aby grupa eksperymentalna rzeczywiście uczestniczyła w programie eksperymentalnym, a więc, aby otrzymała bodziec, do którego została przypisana. Tymczasem częstym problemem występującym podczas realizacji eksperymentów jest „wypadanie" dobranych jednostek eksperymentalnych z grupy eksperymentalnej. Chodzi tu o sytuacje, w których jednostki już po znalezieniu się w grupie eksperymentalnej rezygnują z uczestnictwa w programie. Za jeden z częstych powodów takiego stanu rzeczy Larry Orr podaje m.in. długi czas, jaki dzieli moment doboru do grupy eksperymentalnej i faktyczny udział w danym programie eksperymentalnym – czasem może być to nawet kilka miesięcy (Orr, 1999: 167).

Oprócz powyższych problemów mogą wystąpić inne zakłócenia obciążające wyniki eksperymentu. Klasycznie podawanym przykładem jest tzw. efekt Hawthorne'a, zgodnie z którym już sam fakt pojawienia się bodźca, któremu towarzyszy większe zainteresowanie grupą eksperymentalną, przekłada się na osiągane efekty (Sułek, 1979: 36).

Wszystkie wymienione powyżej problemy negatywnie wpływają na moc przyjmowanych nietestowalnych założeń w modelu eksperymentalnym i w rezultacie czynią wątpliwym możliwość prostego

[21] Np. ankieterzy zbierający dane, osoby monitorujące przebieg eksperymentu (niekoniecznie badacze).

[22] W literaturze przedmiotu czasem określa się ten problem mianem obciążenia zastępowania (ang. *substitution bias)* (Heckman i in., 1995: 105).

wykorzystania grupy kontrolnej ($E(Y_0 | D = 0)$) do oszacowania kontrfaktycznego wyniku dla grupy eksperymentalnej ($E(Y_0 | D = 1)$).

Na koniec trzeba również napisać, że nierzadkie są sytuacje, w których pytania o przyczynowy efekt podjętych interwencji/programów stawia się dopiero po ich realizacji. Dopiero wtedy też myśli się na temat tego, w jaki sposób zmierzyć wielkość oddziaływania ocenianego przedsięwzięcia. W takich okolicznościach eksperyment ze swej definicji nie może być przeprowadzony – jest już na niego za późno. Będąc w takiej sytuacji, nieuniknione jest wykorzystanie innych podejść, aniżeli metoda eksperymentalna. Tu, jak i w przypadku wcześniej opisanych problemów, z pomocą przychodzą rozwiązania wypracowane na gruncie badań obserwacyjnych.

Badania obserwacyjne

W sytuacji, gdy zastosowanie eksperymentu zrandomizowanego jest niemożliwe lub z jakichś względów nieuzasadnione, badacze chcący ustalić efekt przyczynowy działań muszą opierać swoje wnioski na danych pochodzących z tzw. badań obserwacyjnych (nieeksperymentalnych). Są to dane z różnego rodzaju badań sondażowych, administracyjnych, spisów ludności itp. W badaniach tych podobnie jak w eksperymencie występuje grupa jednostek, które otrzymały pewien bodziec, np. uczestniczyły w programie społecznym. Istnieje też pula osób, które pozostają poza oddziaływaniem bodźca. Przez analogię do eksperymentu, grupy te określane są odpowiednio mianem eksperymentalnej i kontrolnej.

Do analizy danych pochodzących z badań obserwacyjnych – w celu ustalenia efektu danej interwencji – teoretycznie można by wykorzystać klasyczną analizę regresji. Jej użycie może okazać się jednak w praktyce problematyczne, z uwagi na wskazany problem mechanizmów selekcji. W przypadku danych pochodzących z badań nieeksperymentalnych, podział jednostek na te, które zostały poddane oddziaływaniu ocenianego bodźca oraz na te, które pozostały poza jego oddziaływaniem, nie ma już charakteru losowego (Rubin, 2005: 7). Podział ten najczęściej pozostaje również poza wpływem badacza, co więcej rzadko kiedy mechanizm doboru jednostek do danego zdarzenia jest w ogóle znany. Brak jest więc informacji na temat wszystkich czynników wpływających na to, jakie jednostki zostały objęte oddziaływaniem danego zdarzenia, a jakie nie. Może mieć to istotny wpływ na analizę danych pochodzących z badań obserwacyjnych. Weźmy bowiem za przykład klasyczny model regresji wyrażony równaniem:

$$Y_i = \alpha + \tau D_i + X_i \beta + \varepsilon_i, \tag{1.06}$$

gdzie D_i jest zmienną dychotomiczną, wskazującą czy dana jednostka wzięła udział w ocenianym zdarzeniu, zaś X_i to wektor zmiennych niezależnych dla jednostki i. Ponieważ w badaniach obserwacyjnych badacz nie ma wpływu na to, kto zostanie objęty ocenianym działaniem, bardzo prawdopodobne jest skorelowanie zmiennej D ze zmienną wyniku Y. Statystyczna kontrola ma tu za zadanie wydzielić wpływ zmiennych wyjaśniających (wektor niezależnych zmiennych X) na zmienną wyniku Y, aby móc przewidywać τ – oszacowanie efektu działania. Jednym z założeń metody najmniejszych kwadratów, wykorzystywanej w analizie regresji, jest jednak to, że składnik resztowy ε nie może być skorelowany z żadnym z predyktorów (Lissowski i in. 2008: 376). Założenie to nie będzie spełnione w przypadku, gdy w modelu regresji nie zostaną zawarte wszystkie ważne predyktory – a więc np. informacja o tym, jaki był klucz doboru jednostek do udziału w danej interwencji (tj. jaki był mechanizm selekcji). Gdy tak się stanie,

zmienne wyjaśniające nie mogą usunąć lub wyjaśnić całej systematycznej zmienności w Y i dlatego składnik resztowy ε może być silnie skorelowany z D. W rezultacie możliwe jest uzyskanie oszacowania efektu oddziaływania programu (τ), które jest obciążone i nie jest zgodne (zwiększanie liczebności próby nie zmniejsza obciążenia) (Guo i in., 2005: 360). W rzeczywistości, zagrożenie niespełnienia założeń modelu regresji jest zaś całkiem realne, bowiem często nie jesteśmy w stanie uwzględnić wszystkich istotnych zmiennych uwikłanych w mechanizm selekcji.

Metody nieeksperymentalne

W obszarze badań obserwacyjnych wypracowano wiele metod minimalizacji obciążania selekcyjnego. Metody te, czy też techniki, przyjęło się nazywać nieeksperymentalnymi lub quasi-eksperymentalnymi. Podobnie jak w metodzie eksperymentalnej dąży się w nich do ustalenia efektu przyczynowego w drodze porównania grupy eksperymentalnej z grupą kontrolną. Ponieważ jednak podział na grupy nie ma charakteru losowego, grupa kontrolna musi zostać stworzona sztucznie. Bazą do jej utworzenia jest większa grupa jednostek nieobjętych oddziaływaniem ocenianego bodźca. Ta ostatnia populacja w dalszej części pracy określana jest mianem puli kontrolnej.

Zadaniem metod nieeksperymentalnych jest eliminacja potencjalnego obciążenia selekcyjnego, czyli minimalizacja hipotetycznej wartości różnicy $E(Y_0 | D = 1) - E(Y_0 | D = 0)$. Jak zostało już powiedziane, w eksperymentach zrandomizowanych tworzone grupy – kontrolna i eksperymentalna – są tworzone z wykorzystaniem mechanizmu losowego. Tym samym są porównywalne przed ocenianą interwencją, bez konieczności rozumienia kontekstu, w jakim dana interwencja jest realizowana. W przypadku metod nieeksperymentalnych, z uwagi na występowanie mechanizmów selekcji, zrozumienie tego kontekstu staje się bardzo istotne (Rosenbaum, 2005: 4). Pojawienie się obciążenia selekcyjnego jest niezależne od procesu badawczego. Dlatego też poznanie czynników, które stały za tym, że część jednostek została objęta daną interwencją, a część nie, ma w przypadku technik nieeksperymentalnych znaczenie fundamentalne. W badaniach obserwacyjnych przyjmuje się, że drogą do minimalizacji obciążenia selekcyjnego jest maksymalne wykorzystanie dostępnych badaczowi informacji. W tym miejscu należy zauważyć, że obciążenia selekcyjne mogą być dwojakiego rodzaju. Do pierwszych należy zaliczyć te, które zostały zmierzone (tzw. *overt biases*), do drugich te, które nie zostały zmierzone, ale podejrzewa się ich istnienie (tzw. *hidden biases*). Jak zauważa Rosenbaum, usuwanie pierwszych i kontrolowanie obszaru niepewności w odniesieniu do drugich to centralne problemy badań obserwacyjnych (Rosenbaum, 2005: 1).

Techniki oparte na schemacie prób dopasowanych według cech

Ważnymi metodami redukcji obciążeń, wynikających z występowania mechanizmów selekcji, są techniki oparte na schemacie prób dopasowanych według cech (ang. *matched samples*). Zgodnie z ogólną koncepcją, metody te wychodzą z założenia, że wszystkie istotne różnice występujące między grupą, która wzięła udział w danej interwencji i grupą będącą poza interwencją, można w całości wytłumaczyć w kategoriach obserwowalnych charakterystyk (Bryson i in., 2002: 10). Na tej podstawie obciążenie selekcyjne można minimalizować poprzez zrównywanie jednostek z obu grup na wektorze pewnego zbioru cech X[23]. Inaczej mówiąc, minimalizacja obciążenia może zostać uzyskana poprzez

[23] Jest to tzw. łączenie według wartości współzmiennych – ang. *covariates matching*.

zapewnienie podobieństwa porównywanych grup (eksperymentalnej i kontrolnej) w zakresie określonych, obserwowalnych charakterystyk.

W praktyce powyższe podejście realizowane jest poprzez znalezienie dla każdej jednostki z grupy uczestniczącej w danym działaniu, co najmniej jednej jednostki z puli kontrolnej. Łączenie, czy też parowanie, odbywa się na podstawie wartości zmiennych, opisujących osoby z obu grup. W pary dobierane są jednostki o takich samych wartościach wszystkich zmiennych, składających się na wektor zmiennych *X*. Dokonane następnie porównanie uśrednionych wyników osób z grupy objętej działaniem z wynikami osób z grupy kontrolnej ma przedstawiać nieobciążony efekt interwencji.

Dla zademonstrowania tego podejścia rozważmy prosty przykład oceny skuteczności reklamy, zachęcającej do kupna pewnego szamponu do farbowania włosów. Załóżmy, że istnieje pewna populacja licząca 15 osób ($N = 15$). 5 jednostek pochodzących z tej populacji obejrzało reklamę szamponu ($N(D = 1) = 5$). Pozostałe 10 osób nie widziało tej reklamy ($N(D=0)=10$). Załóżmy też, że osoby scharakteryzować można za pomocą dwóch zmiennych binarnych $X_1 \in (0,1)$ oraz $X_2 \in (0,1)$, gdzie X_1 wskazuje płeć (mężczyzna, kobieta), a X_2 to informacja o tym, czy osoba kiedykolwiek wcześniej farbowała włosy (nie, tak). Przyjmijmy, że Y to zmienna wynikowa – również binarna $Y \in (0,1)$ – informująca o tym, czy osoba kupiła reklamowany szampon (nie, tak), po emisji reklamy. Przykładowe dane zawiera tabela 1.

Tabela 1. Macierz danych zawierająca 15 jednostek obserwacji

$I = i$	D	X_1	X_2	Y	K
1	0	Mężczyzna	Nie	Nie	Nie
2	0	Kobieta	Tak	Tak	Tak
3	0	Mężczyzna	Tak	Tak	Tak
4	0	Kobieta	Tak	Tak	Tak
5	0	Mężczyzna	Nie	Nie	Nie
6	0	Mężczyzna	Nie	Nie	Nie
7	0	Kobieta	Tak	Tak	Tak
8	0	Kobieta	Nie	Nie	Tak
9	0	Mężczyzna	Nie	Nie	Nie
10	0	Mężczyzna	Nie	Nie	Nie
11	1	Mężczyzna	Tak	Tak	Nie dotyczy
12	1	Kobieta	Tak	Tak	Nie dotyczy
13	1	Kobieta	Tak	Nie	Nie dotyczy
14	1	Kobieta	Tak	Tak	Nie dotyczy
15	1	Kobieta	Nie	Nie	Nie dotyczy

Źródło: opracowanie własne

Jak można zauważyć trzy z pięciu osób, które obejrzały reklamę zakupiło szampon ($E(Y|D=1)=0,6$). W przypadku puli kontrolnej szampon kupiły cztery osoby z dziesięciu ($E(Y|D=0)=0,4$). Pobieżna analiza wskazuje więc, że reklama mogła mieć pozytywny wpływ na decyzję jednostek odnośnie do zakupu szamponu. Należy jednak zauważyć, że osoby w obu grupach znacząco różnią się miedzy sobą pod

względem dwóch wyróżnionych zmiennych X_1 i X_2. W obu grupach inny jest zarówno udział mężczyzn, jak i osób, które widziały reklamę.

Metoda łączenia wg cech polega na wybraniu do grupy kontrolnej – porównawczej – tylko takich osób, które są identyczne, jak te w grupie eksperymentalnej, a więc w powyższym przykładzie identyczne ze względu na wartości obu zmiennych. W tabeli 1 ostatnia kolumna ze zmienną *K* wskazuje, które jednostki z puli kontrolnej są identyczne jak te w grupie eksperymentalnej. I tak, na podstawie podobieństwa, określonego przez wartości zmiennych X_1 i X_2, do grupy kontrolnej dobrane zostały jednostki $i = 2, 3, 4, 7, 8$. Można zauważyć, że w powstałej w ten sposób grupie kontrolnej aż cztery osoby z pięciu zakupiły szampon ($E(Y|X, D=0) = 0{,}8$). Jest to zgoła odmienna sytuacja niż w przypadku analizy zmiennej *Y* w całej puli kontrolnej. Na podstawie otrzymanych danych można powiedzieć, że wpływ reklamy szamponu do farbowania włosów na decyzję o jego zakupie jest raczej negatywny, a wartość oszacowanego efektu przyczynowego ma wartość ujemną $E(Y|X, D=1) - E(Y|X, D=0) = 0{,}6 - 0{,}8 = -0{,}2$.

Oszacowana powyżej różnica będzie oddawała faktyczną wartość efektu przyczynowego tylko, gdy spełnione będą dwa założenia. Po pierwsze należy założyć, że warunkowo względem zmiennych obserwowalnych *X* opisujących jednostki, (Y_0, Y_1) i *D* są od siebie niezależne:

$$(Y_0, Y_1) \perp D \mid X \tag{1.07}$$

gdzie Y_0 to wynik braku udziału w interwencji (przynależności do grupy kontrolnej), Y_1 to wynik udziału w interwencji (przynależności do grupy eksperymentalnej), „$\perp$" oznacza niezależność, *D* to zmienna wskazująca, do której grupy należy jednostka, zaś *X* to wektor zmiennych opisujący jednostki. Oczywiście w przykładzie powyżej Y_0 obserwujemy tylko dla jednostek z puli kontrolnej, a Y_1 tylko dla jednostek w grupie eksperymentalnej (zgodnie z przytoczonym wcześniej równaniem: $Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}$).

Aby pełniej zobrazować założenie 1.07, rozważmy kolejny przykład, tym razem programu rynku pracy. Niech $D = 0$ odpowiada jednostkom, które nie wzięły udziału w programie (grupa kontrolna), zaś $D = 1$ jednostkom, które uczestniczyły w programie (grupa eksperymentalna). Y_0 to wynik braku udziału w programie. Dla uproszczenia przyjmijmy, że Y_0 może mieć tylko dwie wartości $Y_0 \in (0,1) -$ (np. nie pracuje, pracuje). *X* to wektor zmiennych mierzalnych opisujących wszystkie jednostki, zaś x_1, $x_{2...}$ x_n to wybrane podzbiory (wektory) X, takie że $x_1 \neq x_{2...} \neq x_n$. W kolejnej tabeli przedstawiono teoretyczne rozkłady zmiennej wynikowej Y_0 względem zmiennej wskazującej grupę *D* oraz względem *X*. Dla przykładu w grupie osób nieuczestniczących w programie, którym odpowiada wektor zmiennych x_1, średni udział osób pracujących stanowi 30% ($E(Y_0|X = x_1, D=0) = 0{,}3$). Z punktu widzenia pomiaru efektu przyczynowego programu, kluczowym jest jednak pytanie o to, jak kształtowałby się rozkład zmiennej Y_0 w grupie jego uczestników ($E(Y_0|X = x_1, D=1) = ?$). Pytaniem jest więc, jaki byłby udział osób pracujących w grupie uczestników, o charakterystykach wyznaczonych przez wektor x_1, gdyby osoby te – wbrew temu, co faktycznie miało miejsce – nie wzięły udziału w programie eksperymentalnym. Rozkład ten jest oczywiście nie obserwowany w rzeczywistości – jest kontrfaktyczny. Jednak na mocy założenia 1.07 przyjmuje się, że w grupie osób o cechach opisanych przez wektor x_1 rozkład zmiennej Y_0 w grupie interwencji jest identyczny, jak ten obserwowany w grupie kontrolnej (Heckman i in. 1997: 610). W tabeli poniżej kontrfaktyczny rozkład zmiennej Y_0 dla uczestników programu wyróżniony został kursywą, a konkretne imputowane częstości – opatrzone zostały gwiazdką.

Tabela 2. Warunkowe rozkłady Y_0 ze względu na *D*, pod warunkiem *X*

X	$X = x_1$		$X = x_2$		…	$X = x_n$	
D / Y_0	D = 0	*D = 1*	D = 0	*D = 1*	…	D = 0	*D = 1*
$Y_0 = 0$	0,7	*0,7**	0,2	*0,2**	…	0,1	*0,1**
$Y_0 = 1$	0,3	*0,3**	0,8	*0,8**	…	0,9	*0,9**

*wskazuje wartość przypisaną na podstawie obserwowanej w danej warstwie X, wartości Y_0 dla grupy kontrolnej.
Źródło: opracowanie własne.

W konsekwencji, jeśli kontrolowane są obserwowalne różnice w charakterystykach osób z grupy interwencji i grupy kontrolnej, wynik braku udziału w programie dla grupy osób w nim uczestniczących, jest taki sam jak obserwowany wynik braku udziału w programie dla osób z grupy kontrolnej:

$$E(Y_0 \mid X, D = 1) = E(Y_0 \mid X, D = 0) = E(Y_0 \mid X) \tag{1.08}$$

Tym samym, warunkowo względem zmiennych *X*, wynik obserwowany w grupie osób nieuczestniczących w interwencji przedstawia sytuację kontrfaktyczną dla beneficjentów danego działania (Heckman i in., 1997: 610).

Analogicznie, na mocy założenia 1.07, przyjmuje się, że warunkowo względem *X*, wynik Y_1 – obserwowany w grupie eksperymentalnej ($D = 1$) – odpowiada kontrfaktycznemu wynikowi uczestnictwa w programie dla grupy kontrolnej ($D = 0$). Innymi słowy założenie 1.07 pozwala odpowiedzieć na pytanie o to, jaki byłby los osób nie uczestniczących w programie (jak kształtowałby się rozkład zmiennej Y_1), gdyby, przeciwnie niż w rzeczywistości, jednostki kontrolne wzięły udział w interwencji. Przykładowe rozkłady zmiennej Y_1 prezentuje poniższa tabela. Podobnie jak w tabeli 2 kursywą wyróżnione zostały wartości przypisane osobom z grupy kontrolnej, w oparciu o obserwowany rozkład zmiennej Y_1 dla uczestników programu.

Tabela 3. Warunkowe rozkłady Y_1 ze względu na *D*, pod warunkiem *X*

X	$X = x_1$		$X = x_2$		…	$X = x_n$	
D / Y_1	*D = 0*	D = 1	*D = 0*	D = 1	…	*D = 0*	D = 1
$Y_1 = 0$	*0,15**	0,15	*0,1**	0,1	…	*0,05**	0,05
$Y_1 = 1$	*0,85**	0,85	*0,9**	0,9	…	*0,95**	0,95

* wskazuje wartość przypisaną na podstawie obserwowanej w danej warstwie X, wartości Y_1 dla grupy eksperymentalnej.
Źródło: opracowanie własne.

Należy zauważyć, że do oszacowania *ATT*, tj. przeciętnego efektu oddziaływania interwencji na jednostki poddane oddziaływaniu, wystarczy słabsza wersja założenia 1.07. Mianowicie wystarczy, aby spełnione było następujące założenie (Heckman i in., 1997: 611):

$$Y_0 \perp D \mid X \tag{1.09}$$

Wystarczy więc, że prawdziwa jest sytuacja przedstawiona w tabeli 2. Pomimo osłabienia 1.07, występowanie warunkowej niezależności pomiędzy Y_0 i D pozostaje mocnym założeniem. Co więcej jest to założenie nietestowalne w rzeczywistości – badacz nigdy nie będzie miał pewności, na ile jest ono faktycznie spełnione[24].

Podstawowe znaczenie mają tu dane, które składają się na wektor *X*. W założeniu 1.07, jak i w jego słabszej wersji 1.09, przyjmuje się, że wszelkie ewentualne różnice pomiędzy grupą eksperymentalną i grupą kontrolną, występujące na zmiennych nieobserwowanych (nie zmierzonych lub nie dających się zmierzyć, a przez to niewłączonych do wektora *X*), są nieistotne. Krytycznym wymogiem dla spełnienia tego założenia jest posiadanie bogatego zbioru danych, który zawiera wszystkie zmienne odpowiedzialne za uczestnictwo jednostek w działaniu i później obserwowany efekt. Gdy dane nie zawierają wszystkich istotnych zmiennych odpowiedzialnych z jednej strony za uczestnictwo, z drugiej za wynik działania, założenie 1.09, a więc i 1.07, nie będzie spełnione. Efekt programu będzie bowiem zależał od informacji, która jest niedostępna badaczowi. Przykładem takiej informacji, w przypadku programów rynku pracy, będą powody, dla których jednostki nie uczestniczą w danym programie (np. informacja o otrzymaniu propozycji pracy). Jeśli jednak opisane założenie jest spełnione, dopasowanie jednostek w oparciu o wartość zmiennych *X* pełni analogiczną rolę, jak mechanizm randomizacji w metodzie eksperymentalnej, tj., warunkowo względem zmiennych obserwowanych, proces selekcji obserwacji do grupy interwencji jest losowy z punktu widzenia zmiennej wynikowej *Y* (Bryson i in., 2002: 10).

Drugim warunkiem, przyjmowanym w metodzie łączenia wg cech, jest założenie o tym, że każda jednostka ma szansę należeć zarówno do grupy eksperymentalnej, jak i kontrolnej (Rubin i in., 1983: 43)[25]:

$$0 < \Pr(D = 1 \mid X) < 1, \ \forall X \tag{1.10}$$

Założenia 1.07 oraz 1.10 łącznie, noszą miano założenia warunkowej niezależności (*Conditional Independence Asumption – CIA*)[26]. W kontekście obu założeń należy zasygnalizować pewien problem. Dobrane zmienne składające się na wektor *X* muszą pozwalać na spełnienie warunkowej niezależności, co najmniej Y_0 i *D* (zgodnie z 1.09). W związku z tym w wektorze *X* muszą zostać uwzględnione wszystkie istotne zmienne wpływające jednocześnie na Y_0 i *D*. Z drugiej strony, zgodnie z 1.10, nie mogą być to zmienne, które determinują udział jednostek w interwencji, tj. w wektorze X nie może być takich zmiennych, których wybrane wartości występują tylko w grupie uczestników ($D = 1$). W takim przypad-

[24] Wyjątek stanowią sytuację, w których badacz dysponuje danymi porównawczymi, pochodzącymi z badania eksperymentalnego.

[25] Naturalnie 1.10 implikuje również następującą nierówność: $0 < \Pr(D = 0 \mid X) < 1, \ \forall X$.

[26] Założenie to występuje w literaturze pod różnymi nazwami: *ignorable treatment assignment* (Rubin i in., 1983), *selection on observables* (Barnow i in., 1980), *conditional independence* (Lechner, 1999), *exogeneity* (Imbens, 2004). Nazwy te stosowane są zwykle zamiennie (Shenyang i in., 2005: 362).

ku bowiem niemożliwe byłoby znalezienie dla wszystkich jednostek z grupy eksperymentalnej odpowiadających im jednostek kontrolnych. Może tu również wystąpić problem pośredni. Mianowicie, mogą wystąpić trudności z utworzeniem grupy kontrolnej, jeśli jakieś cechy szczególnie sprzyjają przynależności do grupy interwencji. Szczególnie tzn. w taki sposób, że liczba jednostek o określonych charakterystykach w grupie eksperymentalnej jest większa niż liczba jednostek o takich samych cechach w puli kontrolnej[27]. Problemy te, jak również kwestia wymogów, jakie spełnić muszą dane, aby uprawdopodobnić spełnienie założenia 1.07 (i jednocześnie 1.09), zostały szerzej omówione w kolejnym rozdziale.

Tak jak zostało to pokazane w przykładzie dotyczącym skuteczności reklamy szamponu, ogólna procedura doboru próby kontrolnej w metodach opartych na dopasowaniu wg cech polega na wyborze z większej puli obserwacji takich jednostek, które pod względem pewnego zestawu zmiennych mają swoich identycznych odpowiedników w grupie interwencji. I tak, gdy mówimy o osobach, będą one łączone według takich cech, jak np. płeć, wiek, wykształcenie, wysokość zarobków itp. Intuicyjnie optymalnym jest, aby dobierane do siebie jednostki były w zakreślonym zestawie charakterystyk identyczne. Kłopot, jaki się tu pojawia, to możliwość praktycznej realizacji doboru grupy kontrolnej w takim układzie. Z punktu widzenia spełnienia 1.07 lub 1.09 celowe jest, aby kontrolowany zestaw cech był możliwie jak najbogatszy. Jednak wraz ze wzrostem liczby zmiennych, które chciałoby się kontrolować, wzrasta również trudność znalezienia odpowiadających sobie jednostek. W praktyce realizacja dopasowania grupy kontrolnej w takim podejściu wymaga posiadania olbrzymich zbiorów danych. Dla przykładu, dokładne dopasowanie grupy kontrolnej, w oparciu o 20 dwuwartościowych cech (20 zmiennych binarnych), generowałoby teoretycznie konieczność posiadania ponaddwumilionowego zbioru potencjalnych jednostek kontrolnych (istnieje 2^{20} różnych możliwych kombinacji tych cech)[28]. Sprawa dodatkowo komplikuje się, jeśli zmienne, wg których dobierane są jednostki, mają więcej niż dwie wartości lub są zmiennymi ciągłymi. Dysponując więc nawet stosunkowo dużymi – kilkudziestotysięcznymi – zbiorami, często może się okazać, że utworzenie grupy kontrolnej przy takim podejściu będzie bardzo trudne lub niemożliwe do osiągnięcia.

Technika *propensity score matching*

Alternatywą dla łączenia jednostek w oparciu o takie same wartości na wektorze warunkowych zmiennych X jest zbalansowanie X, tj. utworzenie grupy kontrolnej, która będzie miała taki sam rozkład zmiennych w X, jak grupa interwencji (Rosenbaum, 2004: 18). Powyższe rozwiązanie przedstawili w swoim artykule pomysłodawcy techniki PSM – Paul Rosenbaum i Donald Rubin. Zgodnie z ich propozycją, zbalansowanie zmiennych może być uzyskane nie tylko poprzez łączenie wg X, ale również wg pewnej funkcji X, która posiada tzw. właściwości balansujące (ang. *balancing score)*. Funkcja ta musi spełniać następujący warunek (Rubin i in. 1983: 42):

$$X \perp D \mid \mathrm{b}(X) \tag{1.11}$$

to znaczy warunkowy rozkład X, względem wartości funkcji $\mathrm{b}(X)$ jest taki sam dla jednostek w grupie eksperymentalnej ($D = 1$), jak dla jednostek w grupie kontrolnej ($D = 0$). Innymi słowy zmienne składające się na wektor X mogą silnie przewidywać, które jednostki znajdą się w grupie eksperymen-

[27] Jest to problem, który występuje pod nazwą *common support problem*.
[28] Jest to oczywiście przypadek skrajny, zakładający, że w zbiorze nie ma dwóch identycznych jednostek.

talnej, a które w kontrolnej, jednak 1.11 zapewnia, że dla jednostek o takiej samej wartości funkcji b(X), zmienne w X tracą swoje właściwości predykcyjne (Rosenbaum, 2004: 18). Jak dowodzą Rosenbaum i Rubin jedną z takich funkcji jest tzw. *propensity score* (Rubin i in., 1983: 42), która zdefiniowana jest, jako warunkowe prawdopodobieństwo doboru obserwacji do zdarzenia ($D = 1$), szacowane względem wektora zmiennych X:

$$P(X) = \Pr(D = 1 \mid X) \tag{1.12}$$

Rosenbaum i Rubin wykazali, że jeśli zachodzi: $(Y_0,Y_1) \perp D \mid X$ oraz $0 < \Pr(D=1 \mid X) < 1$ dla wszystkich X (a więc gdy spełnione jest *CIA*), to zachodzi również:

$$(Y_0,Y_1) \perp D \mid P(X) \tag{1.13}$$

oraz

$$0 < \Pr(D=1 \mid P(X)) < 1 \qquad \forall P(X) \tag{1.14}$$

Technika PSM, podobnie jak technika łączenia oparta o dokładne dopasowanie wg cech, ma na celu utworzenie grupy kontrolnej, składającej się z jednostek w jak największym stopniu podobnych do tych, które znalazły się w grupie eksperymentalnej. Różnicę stanowi to, że dopasowanie jednostek odbywa się w oparciu o wartość tylko jednej zmiennej – *propensity score*. Technika ta jest więc sposobem na redukcję ilości cech/wymiarów, za pomocą których możemy opisać obserwacje w zbiorze danych. Wymiary te zostają sprowadzone do jednego syntetycznego wskaźnika, definiowanego czasem jako skłonność do partycypacji w warunku interwencji (Konarski i in., 2007: 187).

Patrząc od strony technicznej, wykorzystanie techniki PSM jest procesem złożonym z trzech etapów (Guo, 2005: 362). Pierwszy to wyliczenie wartości *propensity score*, które w praktyce jest nieznane i należy je oszacować[29]. Można w tym celu wykorzystać np. model regresji logistycznej, w którym zmienną zależną jest fakt bycia w grupie objętej oddziaływaniem bodźca. Zmiennymi niezależnymi są cechy, które w założeniu mają wpływać z jednej strony na wynik (Y) z drugiej na uczestnictwo (D) w danym działaniu. Drugim etapem jest dokonanie doboru jednostek do grupy kontrolnej w oparciu o wyliczone *propensity score*. Dobór jednostek do grupy kontrolnej może odbywać się na wiele sposobów. Do najprostszych należy tzw. metoda najbliższego sąsiada[30], a więc dopasowanie jednostek najbardziej podobnych, tj. o najbliższej wartości *propensity score*. Efektem łączenia jest otrzymanie grupy kontrolnej, która zgodnie z założeniem będzie miała zbalansowane wszystkie zmienne obserwowalne wykorzystane w modelu prawdopodobieństwa. Grupa kontrolna będzie więc w zakresie wybranego zestawu cech podobna do istniejącej grupy interwencji. Trzecim etapem jest analiza efektów w oparciu o porównanie grupy interwencji z utworzoną grupą kontrolną. Cały proces zostanie szczegółowo opisany w kolejnym rozdziale.

[29] Warto zauważyć, że wartość *propensity score* znana jest, czy też może być z łatwością wyliczona, w przypadku eksperymentów zrandomizowanych.

[30] Ang. *nearest neighbour*.

Geneza techniki PSM

Technika PSM jest wynikiem prac nad doskonaleniem metod opartych na schemacie prób dopasowanych według cech. Początki rozwoju tej problematyki sięgają lat 70. i kojarzone są m.in. z osobą Donalda Rubina – statystyka, ucznia Williama Cochrana. Sama technika PSM po raz pierwszy zaprezentowana została w 1983 roku, w artykule Rosenbauma i Rubina: *The Central Role of the Propensity Score in Observational Studies for Causal Effects*. Równocześnie na polu ekonomii James Heckman rozwijał pracę w obszarze obciążenia selekcyjnego (ang. *selection bias*). Dotyczyła ona w pewnym zakresie tych samych kwestii co prace Rubina, tj. szacowania efektów działań, gdy przypisanie do warunku interwencji nie ma charakteru losowego. Obie szkoły miały znaczący wpływ na kierunek rozwoju problematyki ustalania efektu przyczynowego, choć co ciekawe, rozwój teorii odbywał się w nich w dużym stopniu niezależnie. Doprowadziło to m.in. do wypracowania różnej terminologii w obszarze tych samych pojęć (Winship i in., 1999: 678). Przyjęło się jednak, że to *propensity score matching*, będące autorstwa Rosenbauma i Rubina, jest obecnie stosowane jako bardziej ogólny termin dla grupy powiązanych technik wykorzystywanych do korekcji obciążenia selekcyjnego w badaniach nieeksperymentalnych (Guo, 2005: 361). Poważny wzrost zainteresowania techniką PSM przypada na lata 90. Stało się to m.in. za sprawą prac Dehija i Wahba (1998, 1999), zawierających porównanie efektów interwencji oszacowanych w drodze eksperymentu i z wykorzystaniem techniki PSM. Jak się okazało, porównania te przedstawiły technikę PSM w bardzo korzystnym świetle, bowiem oszacowane z jej zastosowaniem efekty niewiele odbiegały od efektów oszacowanych na podstawie zrealizowanego wcześniej eksperymentu. Był to swoisty test technik nieeksperymentalnych, które we wcześniejszych porównaniach z techniką eksperymentalną wypadały stosunkowo słabo. Na zwiększenie popularności techniki PSM duży wpływ miały też opisane wcześniej ograniczenia badań opartych na planie eksperymentalnym i krytyka z tym związana (Heckman i in., 1995).

Procedura wykorzystania techniki PSM

Zastosowanie techniki PSM jest w ogólnym zarysie stosunkowo proste i intuicyjne. Dla każdej jednostki z grupy uczestniczącej w badanym zdarzeniu – np. programie społecznym – należy znaleźć co najmniej jedną, jak najbardziej podobną jednostkę z grupy osób nieuczestniczących w nim. Podobieństwo wyrażane jest w kategoriach prawdopodobieństwa udziału w zdarzeniu, które szacowane jest na podstawie obserwowalnych charakterystyk poszczególnych osób. Wybrane osoby składają się dalej na grupę kontrolną, której wyniki można porównać z wynikami obserwowanymi w grupie osób uczestniczących w interwencji.

Mimo względnej prostoty podejście to zawiera wiele punktów krytycznych. Wykorzystując technikę PSM, badacz musi podjąć wiele decyzji, od których finalnie zależy oszacowany efekt ocenianej interwencji. Poniższy rozdział zawiera prezentację poszczególnych elementów składowych oraz kluczowych etapów związanych z zastosowaniem techniki PSM.

Dane w technice PSM

Punktem wyjścia do zastosowania techniki PSM jest dostępność odpowiednich danych. Problem danych w technice PSM jest złożony. Po pierwsze w szacowanym modelu prawdopodobieństwa musi znaleźć się taki zestaw zmiennych niezależnych *X*, który uczyni realnym – przedstawione w rozdziale pierwszym – założenie warunkowej niezależności (*CIA*). Po drugie muszą być spełnione pewne minimalne wymagania, jeśli chodzi o liczebność grupy interwencji oraz puli kontrolnej, wykorzystanej w procesie tworzenia grupy kontrolnej (Bryson i in., 2002: 14). Po trzecie wykorzystywane dane muszą być zebrane w odpowiednim czasie i wreszcie po czwarte dane te muszą być gromadzone w wystandaryzowany sposób. Wszystkie te kwestie są poniżej omówione.

Założenie *CIA,* o którym mowa była we wcześniejszym rozdziale, wymaga, aby warunkowo na wartościach *propensity score*, zmienne wynikowe (Y_0, Y_1) były niezależne od faktu przypisania (*D*) jednostek do badanego działania: $(Y_0,Y_1) \perp D \mid P(X)$. Bezpośrednio z założenia wynika więc, że muszą być to zmienne, które wpływają jednocześnie na decyzję uczestnictwa *D* oraz na zmienną wynikową *Y*. Nie ma sensu uwzględniać w modelu zmiennych, które wpływają tylko na partycypację (*D*) lub tylko na wyniki (*Y*). Jeśli istnieje jakiś czynnik, który ma wpływ tylko na uczestnictwo osób w danym programie, będzie on w modelu bezużyteczny, jako że i tak nie zaburza on zmiennej wynikowej[31]. Inaczej mówiąc włączenie do modelu zmiennej, oddziałującej tylko na prawdopodobieństwo udziału w interwencji, nie będzie miało wpływu na warunkowy względem *X*, rozkład zmiennej Y_0 w grupie eksperymentalnej i w grupie kontrolnej (por. przykład dotyczący skuteczności reklamy z rozdziału 1). Podobnie jeśli istnieje czynnik, który oddziałuje na wynik, lecz nie ma wpływu na uczestnictwo, jego kontrola jest również bezcelowa, bowiem jest on w tym samym stopniu obecny (występuje tak samo często) w grupie uczestników, co w puli kontrolnej (jeśli byłoby inaczej, rozkład takiej zmiennej w obu grupach byłby różny).

Zasadnym w takim razie jest pytanie o to, jakie zmienne wpływają jednocześnie na uczestnictwo osób w danej interwencji i późniejszy ich wynik? W zasadzie nie ma tu prostej i uniwersalnej odpowiedzi (Bryson i in., 2002: 13). W dużej mierze zależy to od tego, co jest przedmiotem badania. Trzeba jednak pamiętać, że *CIA* jest założeniem nietestowalnym, a więc badacz nigdy nie będzie miał pewności,

[31] Co więcej może okazać się szkodliwy podczas szacowania modelu prawdopodobieństwa.

czy zostało ono w pełni spełnione[32], tj. czy wszystkie istotne i właściwie zmienne zostały przez niego ujęte w modelu. Dlatego też dobór zmiennych musi być dokonany niezwykle uważnie. Powinien on opierać się na uznanych teoriach społecznych, ekonomicznych, doświadczeniach oraz wiedzy wyciągniętej z wcześniejszych badań przeprowadzonych w obszarze danego typu interwencji. Przydatne mogą okazać się także informacje pozyskane od samych potencjalnych odbiorców działania, na temat ich indywidualnych motywów uczestnictwa lub jego braku. Informacje od administratorów danej interwencji również mogą w tym względzie być pomocne (Bryson i in. 2002: 14). To, czego trzeba być świadomym to to, że pominięcie istotnych zmiennych może poważnie zwiększyć obciążenie szacunków efektu interwencji (Heckman i in., 1997: 637). Wskazuje się, że w przypadku szacowania partycypacji w programach rynku pracy szczególnie istotne wydaje się uwzględnienie zmiennych, które odnoszą się do postaw i indywidualnych motywacji jednostek. Zmienne tego typu często są jednak trudne do mierzenia i najczęściej nie są zbierane. Dla przykładu Heckman, Ishimura i Tod analizują w tym kontekście wpływ różnych modeli uczestnictwa na szacowany efekt jednego z programów rynku pracy (Heckman i in., 1997: 634). Wykluczając z modelu kolejno konkretne zmienne, obserwują, jak zmienia się obciążenie szacowanego efektu interwencji. Główne wnioski z ich badań wskazują, że nieuwzględnienie kluczowych zmiennych, takich jak np. historia zatrudnienia i zarobki znacznie obciąża szacunki efektu interwencji[33].

Oprócz teoretycznej wiedzy uzasadniającej włączenie określonych zmiennych do modelu szacującego prawdopodobieństwo uczestnictwa, można w tym celu wykorzystać również pewne metody statystyczne. Jedna z takich metod wskazuje na optymalny zestaw zmiennych, jaki powinien znaleźć się w modelu. Metoda ta (ang. *Hit or Miss Method*), odwołuje się do współczynnika poprawnej predykcji uczestników działania (Heckman i in., 1997: 617). Zmienne są wybierane w taki sposób, aby maksymalizować wewnątrzgrupowy współczynnik predykcji faktycznie uczestniczących w danym działaniu jednostek. Metoda klasyfikuje obserwacje jako 1, jeśli oszacowane dla niej *propensity score* jest większe niż odsetek jednostek, które faktycznie uczestniczyły w zdarzeniu, czyli jeśli $P(X) > P$, gdzie $P = E(D = 1)$. Jeśli $P(X) \leq P$, obserwacja jest klasyfikowana z wartością 0. Metoda ma za zadanie maksymalizację ogólnego współczynnika klasyfikacji, zakładając, że koszt złego zaklasyfikowania jest jednakowy dla obu grup. Dla przykładu, jeśli w pewnym działaniu uczestniczyło 0,03 populacji potencjalnych uczestników, wartość jeden zostanie przypisana tym jednostkom, których oszacowane *propensity score* jest większe od 0,03, wartość zero zostanie przypisana w przeciwnym przypadku. Na tej podstawie należy zbadać odsetek poprawnie zaklasyfikowanych uczestników i osób nieuczestniczących w działaniu, a więc należy zobaczyć, jaki jest odsetek jedynek w grupie uczestników i jaki odsetek zer w puli kontrolnej. Należy wybrać taki zestaw zmiennych, który maksymalizuje obie częstości.

Trzeba jednak zaznaczyć, że niepożądane jest również umieszczenie w modelu zmiennych „zbyt dobrze" przewidujących uczestnictwo osób w danym działaniu. Jeśli dla pewnych X, $P(X) = 0$ lub $P(X) = 1$, a więc gdy można dokładnie zaklasyfikować uczestników lub pozostałe osoby na podstawie oszacowanego *propensity score*, to wtedy niemożliwe jest dokonanie łączenia warunkowo względem X, aby oszacować efekt przyczynowy. Dla pewnych X niespełniony byłby więc warunek: $0 < Pr(D = 1 \mid P(X)) < 1$. Zmienne muszą być zatem wystarczająco dobre, aby możliwe było uzyskanie warunkowej niezależności

[32] Wyjątkiem jest sytuacja posiadania porównawczych danych pochodzących z eksperymentu zrandomizowanego – por. Raeejev H. Dehejia, Sadek Wahba.

[33] Autorzy mogli przeprowadzić takie porównanie/ocenę obciążenia, bowiem oprócz danych nieeksperymentalnych posiadali dane, które pochodziły z przeprowadzonego na dużą skalę eksperymentu.

pomiędzy *Y* i *D*, jednak nie mogą być za dobre, tzn. nie mogą dokładnie przewidywać *D*. Przykładem takiej zmiennej może być obszar realizacji pewnego projektu. Przypuśćmy, że badamy pilotażowy projekt, którego realizacja odbywa się tylko w wybranym obszarze oraz że wszystkie jednostki w puli kontrolnej pochodzą spoza tego obszaru. W takiej sytuacji zmienna obszar będzie idealnie klasyfikować uczestników działania, a utworzenie grupy kontrolnej będzie niemożliwe.

Drugą istotną kwestią jest liczebność porównywalnych grup. Powszechnie wskazuje się, że badacz, chcący zastosować technikę PSM, musi dysponować stosunkowo licznym zbiorem jednostek stanowiących potencjalną grupę kontrolną. W przypadku małej liczebności puli kontrolnej rośnie zagrożenie wystąpienia tzw. *common support problem*. Problem ten jest bezpośrednio związany z omówioną powyżej kwestią umieszczania w modelu zmiennych dokładnie przewidujących fakt partycypacji w działaniu eksperymentalnym. Oprócz tego skrajnego przypadku w praktyce może wystąpić również sytuacja pośrednia. Mianowicie jednostki eksperymentalne mogą na tyle różnić się od reszty populacji kontrolnej, że trudno będzie o dobranie podobnych do siebie jednostek (o takiej samej lub bliskiej wartości *propensity score*). Może się więc okazać, że nie wszystkie osoby w grupie interwencji znajdą swojego odpowiednika w puli kontrolnej. Jako że najczęściej osoby nie uczestniczące w interwencjach znacząco różnią się od uczestników, uzyskanie bliskich dopasowań wymagać będzie puli kontrolnej o znacznej liczbie potencjalnych jednostek porównawczych. Brak odpowiednich jednostek porównawczych będzie natomiast skutkował bądź to zawężeniem dokonywanych uogólnień (do populacji jednostek interwencji, dla których udało znaleźć się odpowiednio podobne jednostki kontrolne), bądź podważy w ogóle zasadność oszacowania efektu działania z wykorzystaniem techniki PSM (w przypadku dużych rozbieżności między porównywanymi grupami). Zagadnienie *common support problem* zostanie jeszcze nieco szerzej omówione dalej.

Istotny jest również czas pomiaru zmiennych. Dane włączane do modelu muszą być aktualne na moment przystępowania jednostek do udziału w danej interwencji. Ich badanie musi więc poprzedzać otrzymanie bodźca (Rubin i in., 1983: 42). Dane nie powinny być zbierane np. po uczestnictwie/braku uczestnictwa osób w programie. Mogą bowiem być już zaburzone interwencją – w przypadku uczestników – lub oddziaływaniem innych zdarzeń – w przypadku osób w danym działaniu nieuczestniczących. Mowa tu oczywiście o zmiennych, które mogą podlegać zmianie w czasie. Takimi zmiennymi dla programów rynku pracy będą np. kwalifikacje zawodowe, wykształcenie, motywacje i inne.

Ważną kwestią, na którą wskazuje m.in. Heckman, jest również sprawa standaryzacji zbieranych danych (Heckman i in., 1997: 606). Dane z grupy interwencji i puli porównawczej powinny mieć to samo źródło, tj. powinny być zbierane za pomocą tej samej metody – np. takiego samego kwestionariusza. Ma to oczywiście na celu eliminację dodatkowych błędów systematycznych, związanych z różnicami w narzędziu do zbierania danych. W tym kontekście zwraca się również uwagę na to, że osoby z grupy interwencji oraz z grupy kontrolnej powinny pochodzić z tego samego środowiska ekonomicznego. Wskazuje się na przykład, że w przypadku programów wspierających zatrudnienie, duże znaczenie dla jakości szacowanych estymatorów ma to, aby jednostki z grupy eksperymentalnej i z grupy kontrolnej pochodziły z tego samego, lokalnego rynku pracy (Heckman i in., 1997: 612).

Model uczestnictwa – szacowanie *propensity scores*

Mając wstępnie[34] zdefiniowany katalog zmiennych, które zostaną zawarte w modelu partycypacji, kolejnym krokiem implementacji techniki PSM jest oszacowanie wartości *propensity score*. Na tym etapie należy podjąć decyzję, jaki wykorzystać model estymacji. Istnieją różnorodne metody szacowania $P(X_i)$ jednak najczęściej w literaturze wskazuje się na model logitowy lub probitowy – z preferencją dla tego pierwszego (Konarski i in., 2007: 191). Marco Caliendo oraz Sabine Kopeinig zwracają uwagę, że w przypadku, gdy zmienna zależna ma charakter dychotomiczny (uczestnictwo lub brak uczestnictwa), oba modele przynoszą podobne wyniki (Caliendo, i in. 2005: 5). Wybór metody estymacji *propensity score* może mieć jednak bardziej krytyczne znaczenie, gdy przewidywane zdarzenie ma charakter wielowartościowy (ang. *multiple treatment case*), tj., gdy jednostka może wybierać pomiędzy więcej niż dwoma możliwościami (uczestniczyć lub nie). Mowa tu np. o różnych ścieżkach specjalizacyjnych realizowanych w ramach danego programu. W takim przypadku należałoby wykorzystać tzw. wielomianowy model logitowy (ang. *mulinominal logit*) lub wielomianowy model probitowy (ang. *multinominal probit*). Ten pierwszy wymaga jednak mocniejszych założeń, stąd czasem wskazuje się na wykorzystanie modelu probitowego.

Trzecim, pośrednim rozwiązaniem, jest zastosowanie wielu modeli regresji logistycznej. W takim układzie tworzone są kolejno modele regresji, w których uwzględnia się wszystkie możliwości, przed którymi stoi jednostka.

Wskazuje się na dwie wady takiego podejścia. Po pierwsze, wraz ze wzrostem liczby możliwych opcji, spomiędzy których może wybierać jednostka, liczba modeli do oszacowania rośnie nieproporcjonalnie (dla k opcji należy przygotować $0{,}5(k(k - 1)$ modeli). I po drugie, w każdym z modeli rozpatrywane są równocześnie tylko dwie opcje, a więc szacuje się prawdopodobieństwo udziału w jednej z dwóch wybranych grup, mimo że łącznie wszystkich grup jest więcej. Brak jest więc spojrzenia całościowego na daną interwencję.

W ewaluacjach interwencji publicznych najczęściej spotykaną sytuacją jest jednak ocena dychotomicznego zdarzenia. Zwykle problem badawczy dotyczy oceny udziału w danym programie. Najpowszechniej w takim przypadku wykorzystywana jest regresja logistyczna.

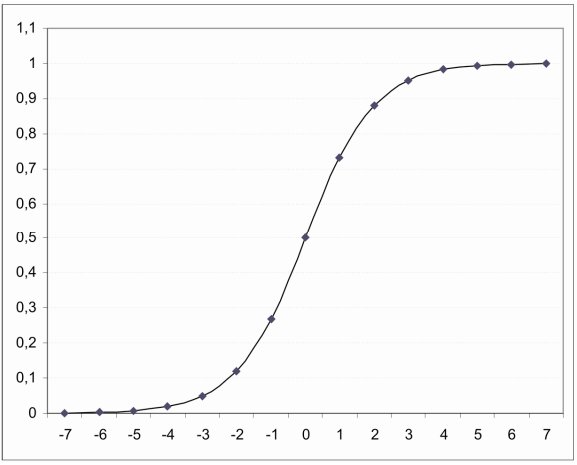
Regresja logistyczna

Regresja logistyczna, jest matematycznym modelem, który można wykorzystać do opisania wpływu kliku zmiennych niezależnych $x_1, x_2,\ldots,x_k$ na dychotomiczną (przyjmującą wartość 0 lub 1) zmienną Y (Stanisz, 2007: 218). Przykładami takiej zmiennej zależnej będzie np. uczestnictwo w programie społecznym (1 – osoba uczestniczyła, 0 – osoba nie uczestniczyła). Model regresji logistycznej opiera się na funkcji logistycznej, która jest następującej postaci:

$$f(z)=\frac{e^z}{1+e^z} \tag{2.01}$$

[34] Zmienne mogą jeszcze podlegać modyfikacjom na etapie analiz.

Wykres tej funkcji przedstawiony jest na rysunku 1.



Rys. 1. Postać funkcji logistycznej

Jak widać na rysunku, funkcja logistyczna przyjmuje wartości z przedziału (0; 1). Wartość funkcji zmierza do zera, gdy x dąży do minus nieskończoności. Dla x dążących do plus nieskończoności funkcja logistyczna dąży do jedności. Charakterystyczny kształt funkcji podobny jest do litery *s*. Ponieważ funkcja przyjmuje wartości między 0 a 1, może ona zostać wykorzystana do opisywania wartości prawdopodobieństwa. W szczególności może być to prawdopodobieństwo wzięcia udziału w pewnym zdarzeniu, np. programie społecznym. Model regresji logistycznej wyraża się następującym równaniem:

$$P\left(D=1\middle| x_1, x_2, ..., x_k\right)=\frac{e^{\beta_0+\sum_{i=1}^{k}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{k}\beta_i x_i}} \tag{2.02}$$

gdzie β_0 to stała regresji, β_i, $i = 1,\ldots, k$ to współczynniki regresji, zaś x_1, $x_2,\ldots, x_k$ to zmienne niezależne, które mogą być zarówno ilościowe, jak i jakościowe. Aby oszacować β, wykorzystywana jest metoda największej wiarygodności (ang. *maximum likelihood*). Ogólna idea tej metody polega na szacowaniu wartości nieznanych parametrów w taki sposób, aby te maksymalizowały prawdopodobieństwo uzyskania zaobserwowanych wartości zmiennej zależnej (Hosmer i in., 2000: 8).

W skład typowej diagnostyki utworzonego modelu wchodzi m.in. weryfikacja, czy powstały model – zawierający zmienne niezależne – mówi więcej na temat zmiennej wynikowej niż model bez zmiennych objaśniających. Istotną kwestią w tym miejscu, jest pytanie o to, co robić ze zmiennymi, które okażą się być nieistotne statystycznie[35]. Możliwe są tu dwa przeciwstawne podejścia: usunięcie nieistotnych zmiennych z modelu lub pozostawienie ich. Pierwsze z rozwiązań jest zgodne z ogólną

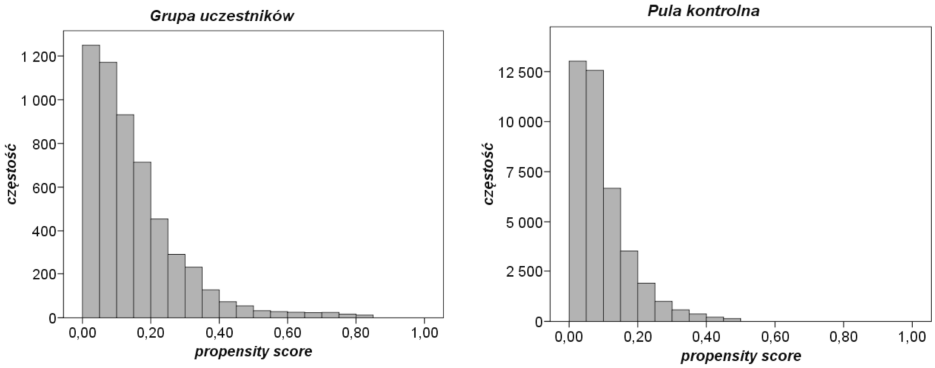
[35] Problem ten jest bezpośrednio związany zagadnieniami poruszanymi w części dotyczącej danych powyżej. Weryfikacja istotności zmiennych może bowiem być jednym ze sposobów na zdefiniowanie ostatecznego kształtu szacowanego modelu prawdopodobieństwa.

logiką prowadzenia analiz z wykorzystaniem metody regresji logistycznej. Przyjmuje się bowiem, że celem przeprowadzania analizy regresji jest znalezienie najlepiej dopasowanego i jednocześnie najprostszego modelu opisującego związek pomiędzy zmienną zależną (wynikową) i zestawem zmiennych niezależnych (predyktorów) (Hosmer i in., 2000: 1). W tym kontekście, w modelu powinny pozostać tylko zmienne istotne statystycznie. Za takim rozwiązaniem opowiadają się m.in. Bryson, Dorset i Purdon (Bryson i in. 2002: 25), którzy dodatkowo wskazują, że o ile umieszczenie zmiennych nieistotnych statystycznie nie obciąży wyników estymacji, może jednak zwiększyć wariację szacowanych zmiennych. Z drugiej strony Rubin i Thomas wskazują, że zmienne nieistotne statystycznie powinny pozostać w modelu, chyba że istnieje zgoda co do tego, że nieistotna zmienna nie oddziałuje na wynik lub jest niewłaściwym predyktorem (Rubin i in. 1996: 253). Jeśli są w tym względzie jakieś wątpliwości, autorzy radzą pozostawić zmienną w modelu. Należy tu pamiętać, że głównym celem techniki PSM nie jest predykcja warunku selekcji jednostek do działania – tak dobrze, jak to tylko możliwe – lecz zbalansowanie wszystkich istotnych predyktorów, tak aby ich rozkład był możliwie jak najbardziej zbliżony w grupie eksperymentalnej i w grupie kontrolnej (Caliendo, i in. 2005: 7). Pozostawienie nieistotnych statystycznie zmiennych przyczynia się do większej porównywalności obu grup, czyniąc jednocześnie założenie 1.01 bardziej realnym.

W tym miejscu warto poruszyć jeszcze kwestię szczególnie istotnych zmiennych w modelu. W praktyce taka sama wartość *propensity score* u dwóch różnych jednostek nie oznacza, że będą one miały takie same wartości na wszystkich zmiennych wykorzystanych w modelu. Możliwe jest, że będą się między sobą różnić wartościami wybranych zmiennych. Na przykład w jednej parze mogą znaleźć się mężczyzna i kobieta, osoba z wykształceniem podstawowym i wykształceniem wyższym itp. Z różnych względów może się wydawać zasadne, aby w przypadku wybranych zmiennych w parach znajdowały się osoby, które w zakresie pewnych charakterystyk są identyczne. Podejście takie znajduje szczególnie uzasadnienie w przypadku, gdy istnieją przypuszczenia co do tego, że oceniany efekt działania jest silnie zróżnicowany (np. wielkość efektu obserwowanego w grupie kobiet znacząco odbiega od efektu, który można zaobserwować w grupie mężczyzn). W przypadku programów rynku pracy mogą to być takie cechy jak wspomniana płeć, wykształcenie czy region zamieszkania. Podejście takie przypomina w tym przypadku metodę dokładnego łączenia wg cech, które opisane zostało w rozdziale pierwszym i jako takie stanowi pierwowzór PSM. O ile opisane powyżej rozwiązanie może znaleźć uzasadnienie, o tyle rodzi również pewne komplikacje praktyczne. Mianowicie, wydzielając dodatkowe zmienne z modelu, cała procedura – szacowanie *propensity score*, dobór grupy kontrolnej i inne – muszą zostać przeprowadzone osobno dla każdej z wydzielonych podgrup (Caliendo, i in. 2005: 8). W rozdziale kolejnym zaprezentowany został przykład, w którym zdecydowano się wydzielić z modelu zmienną region (a dokładnie województwo). W konsekwencji zamiast jednego, oszacowanych zostało 13 modeli prawdopodobieństwa.

Problem wspólnego przedziału określoności

Oszacowanie modelu uczestnictwa daje w rezultacie wyliczenie wartości *propensity score* dla każdej osoby w zbiorze. Wartości te stanowią klucz do utworzenia grupy kontrolnej spośród wszystkich jednostek obserwacji nieobjętych oddziaływaniem ocenianego działania (tzw. pula kontrolna). Przydatne przed przejściem do procedury łączenia jest wykonanie rozkładu oszacowanych *propensity score* w grupie ocenianej interwencji oraz w puli kontrolnej. Rozważmy następujący przypadek:



Rys. 2. Rozkład przykładowych *propensity score* w grupie ocenianej interwencji oraz w puli kontrolnej
Źródło: opracowanie własne.

Powyższe wykresy obrazują przykładowe rozkłady wartości *propensity score* w dwóch grupach – grupie uczestników oraz w puli kontrolnej. Jak można zauważyć, rozkłady różnią się między sobą. Wartości *propensity score* dla grupy uczestników zawierają się w przedziale [0,01; 0,83]. Dla puli kontrolnej przedział ten jest istotnie mniejszy [0,01; 0,49]. Poza wspólnym przedziałem znajduje się ok. 3% (162 z 5452) jednostek przynależących do grupy uczestników. Zobrazowana sytuacja prognozuje możliwość wystąpienia tzw. problemu wspólnego przedziału określoności (ang. *common support problem*) (Strawiński, 2008: 209). Problem ten – mówiąc najprościej – wyraża się w braku odpowiedników w puli kontrolnej, dla części jednostek z grupy działania. Problem stanowi empiryczne naruszenie założenia $0 < \Pr(D = 1 \mid P(X)) < 1$ (Heckman i in., 1997: 663). Jak już zostało wspomniane w części dotyczącej danych, problem wspólnego przedziału określoności może mieć w praktyce dwa źródła (Lechner, 2000: 14). Po pierwsze w modelu mogą znaleźć się idealne predyktory uczestnictwa. Zawarcie takich zmiennych grozi oszacowaniem prawdopodobieństwa, które dla części obserwacji równe będzie 1. Po drugie wyliczone wartości *propensity score* mogą mieć różne rozkłady w grupie interwencji i w puli kontrolnej – tak jak na rysunku powyżej. W rezultacie część uczestników może nie posiadać w puli kontrolnej swojego odpowiednika. Pierwsza kwestia została omówiona już wcześniej. Należy wystrzegać się umieszczania w modelu zmiennych „zbyt dobrze" przewidujących partycypację jednostek w ocenianym zdarzeniu. Druga sprawa dotyczy występujących różnic pomiędzy uczestnikami i osobami nieuczestniczącymi w interwencji. Jeśli jednostki z obu grup znacząco różnią się między sobą, może okazać się, że w przypadku pewnych wartości *propensity score* nie będzie możliwe przeprowadzenie wiarygodnego dopasowania. Najczęściej taka sytuacja może mieć miejsce w odniesieniu do wyższych wartości *propensity score*. W dużej mierze badacz niewiele może w takiej sytuacji poradzić. Jeśli dysponuje zbiorem danych, z którego w wyniku procedury łączenia nie jest w stanie wygenerować porównywalnej z grupą interwencji grupy kontrolnej, to musi przyjąć, że dla pewnych jednostek nie można po prostu oszacować efektu interwencji. W innym przypadku „jakość" dopasowanej grupy – oceniana przez pryzmat wielkości różnic w wartościach *propensity score* między jednostkami w grupie interwencji a jednostkami w grupie kontrolnej, może okazać się niska. W konsekwencji może to doprowadzić do uzyskania obciążonego oszacowania efektu interwencji. Pytanie, co w takiej sytuacji robić?

Proponowanym podejściem do zilustrowanego na rysunku 2 problemu jest ograniczenie zbiorów danych, tj. wykluczenie z nich jednostek, które nie posiadają swoich odpowiedników w grupie interwencji. Chodzi więc o zawężenie wartości *propensity score* do wspólnego przedziału (tzw. *common support region*). W powyższym przykładzie byłby to przedział [0,01; 0,49]. Oczywiście, w takim przypadku, dla jednostek wyłączonych z analiz efekt danej interwencji nie może zostać obliczony. Stąd, o ile nie da się założyć, że efekt ocenianego działania dla wszystkich jednostek jest taki sam, to oszacowany efekt będzie przedstawiał wynik jedynie dla pewnej ograniczonej subpopulacji, dla której akurat dostępne są jednostki w grupie porównawczej. Takie podejście może jednak generować problemy, zwłaszcza gdy z powodu braku odpowiednich jednostek w grupie porównawczej, konieczne jest wyłączenie z analiz znacznej części badanej populacji. W takiej sytuacji warto przeanalizować charakterystykę usuniętych obserwacji, te mogą bowiem istotnie wpływać na ostateczne wnioski w zakresie oszacowanego efektu interwencji.

Metody doboru grupy kontrolnej w technice PSM

Kolejnym krokiem w zastosowaniu techniki PSM jest wybór odpowiedniej techniki selekcji jednostek z puli kontrolnej do grupy kontrolnej. Możliwych jest tu co najmniej kilka podejść, które w praktyce wyrażane są przez różne algorytmy dopasowania jednostek. Oprócz tego każda z technik występuje w kilku wariantach. Mnogość podejść oddaje problem doboru grupy kontrolnej, który na tym etapie sprowadza się do optymalizacji wykorzystania danych znajdujących się w puli kontrolnej. W praktyce przed rozpoczęciem procedury łączenia należy dokonać trzech decyzji:

1) czy dokonywać dopasowania ze zwracaniem czy bez zwracania, tzn., czy raz wykorzystaną jednostkę z puli kontrolnej włączać na powrót do tej puli,
2) ile jednostek kontrolnych ma przypadać na jednego beneficjenta, i finalnie,
3) jaką metodę łączenia zastosować.

Czynność doboru jednostek kontrolnych można dokonywać bez zwracania lub ze zwracaniem. W pierwszym przypadku jednostka po wybraniu jej do grupy kontrolnej zostaje usunięta z dalszej procedury łączenia, w drugim wraca do puli. Wybór opcji w praktyce zależy od tego, jak liczna jest grupa porównawcza oraz jaki jest w niej rozkład *propensity score*. Dobór ze zwracaniem może okazać się szczególnie korzystny w sytuacji, gdy dana jednostka z puli kontrolnej będzie miała najbliższe *propensity score* dla więcej niż jednej jednostki z grupy eksperymentalnej. Z kolei w przypadku występowania znaczących różnic między grupą uczestników i pulą kontrolną istnieje ryzyko, polegające na dobieraniu w pary jednostek, które znacząco się od siebie różnią – jeśli chodzi o wielkość oszacowanego *propensity score* (Dehejia, 1999: 9). W takich sytuacjach zastosowanie wariantu ze zwracaniem podniesie znacząco jakość dopasowania i tym samym zmniejszy obciążenie (suma wartości różnic *propensity score* liczona dla całego zbioru będzie najmniejsza). Jednak konsekwencją takiego postępowania może być zwiększenie błędu standardowego szacowanego efektu, bowiem liczebność grupy kontrolnej będzie mniejsza niż w przypadku zastosowania algorytmu bez zwracania (Smith i in., 2005: 315). Z kolei zaletą łączenia ze zwracaniem jest to, że wyniki są potencjalnie mniej czułe na porządek, w jakim jednostki zostały

ustawione przed procedurą łączenia[36]. Dlatego w podejściu przeciwnym, tj. w przypadku wykorzystania wariantu bez zwracania zaleca się, aby przed rozpoczęciem procesu doboru dokonać losowego uporządkowania wszystkich obserwacji.

Z powyższą kwestią związane jest zagadnienie liczby jednostek, jaka ma być dobrana do jednostek z grupy uczestników. Najbardziej podstawowe podejście zakłada realizację dopasowań typu 1 do 1, tj. takich, w których do jednej jednostki obserwacji z grupy interwencji dobiera się dokładnie jedną jednostkę z puli kontrolnej. W konsekwencji grupa kontrolna ma taką samą lub zbliżoną liczebność jak grupa interwencji. Inne podejścia wspierają dopasowania typu 1 do n, tj. takie, w którym do jednej jednostki z grupy interwencji dobiera się kilka jednostek z puli kontrolnej. W tym wypadku grupa kontrolna jest oczywiście odpowiednio większa niż grupa interwencji, natomiast jednostki, które znalazły się w grupie kontrolnej powinny zostać przeważone. Za takim rozwiązaniem przemawia chęć maksymalizacji wykorzystania informacji zawartych w jednostkach należących do puli kontrolnej. Wariant ten zwiększa liczebność grupy kontrolnej, a tym samym zmniejsza błąd standardowy szacowanego efektu interwencji. Z drugiej strony może również zwiększyć obciążenie szacunku efektu, gdyż zwiększa prawdopodobieństwo gorszych dopasowań (z większymi różnicami w wartości *propensity score)*.

Ostateczny wybór techniki i jej wariantu w praktyce zależy najczęściej od jakości i charakteru dostępnych danych, tj. przede wszystkim od tego, jak liczna jest pula kontrolna i na ile możliwe są dokładne dopasowania. Poniżej przedstawione zostaną niektóre z najczęściej stosowanych algorytmów tworzenia grupy kontrolnej, wraz z komentarzem na temat możliwości aplikacji danej techniki w różnych sytuacjach.

Metoda najbliższego sąsiada

Metoda najbliższego sąsiada (ang. *nearest neighbour*)[37] jest najprostszą i zarazem najbardziej intuicyjną metodą doboru grupy kontrolnej. W podstawowej wersji tej metody dla każdej jednostki z grupy interwencji dobiera się jedną jednostkę z puli kontrolnej o najbliższym *propensity score*. Czynność ta może być dokonywana ze zwracaniem lub bez. Można wprowadzić również modyfikację, polegającą na przyporządkowaniu do jednej jednostki eksperymentalnej kilku jednostek kontrolnych. Dopasowanie może być więc typu 1 do 1 lub 1 do n.

Patrząc z punktu widzenia doboru grupy kontrolnej, metoda najbliższego sąsiada jest techniką najbardziej skuteczną. O ile tylko pula kontrolna jest co najmniej tak liczna, jak grupa eksperymentalna, realizacja algorytmu w każdym przypadku zakończy się sukcesem – tj. każda jednostka z grupy interwencji będzie posiadała dobraną jednostkę w grupie kontrolnej. Technika ta wymaga z tego względu sporej uwagi, bowiem na etapie łączenia różnice między jednostkami nie są kontrolowane przez żaden mechanizm. Mogą więc wystąpić słabe dopasowania – w utworzonych parach mogą znaleźć się jednostki o znacząco różnych wartościach *propensity score*. Z tego względu zaleca się sprawdzenie otrzymanego rozkład różnic na wartościach *propensity score* pomiędzy dobranymi jednostkami. Gdyby różni-

[36] Gdy np. w puli kontrolnej jest mniej obserwacji o określonych wartościach *propensity score* (w praktyce najczęściej dotyczy to wysokich, bliskich 1, wartości) niż w grupie eksperymentalnej, to w wariancie bez zwracania swoich sąsiadów znajdą tylko te jednostki, które były łączone w pierwszej kolejności.

[37] W przypadku programu SPSS dostępne są różne makra, które w oparciu o język wbudowany w SPSS język poleceń (command syntax), są w stanie wygenerować grupę kontrolną metodą najbliższego sąsiada. Jedno z takich makr dostępne jest pod adresem: http://sswnt7.sowo.unc.edu/VRC/Lectures z dnia 01.03.2009 r. W pakiecie statystycznym Stata dostępny jest osobny, dodatkowy moduł do prowadzenia analiz z wykorzystaniem techniki PSM – jest to moduł PSMATCH2.

ce okazały się znaczne, można zastosować wariant ze zwracaniem, pamiętając jednak, że decyzja ta jest kwestią pewnego kompromisu pomiędzy mniejszym obciążeniem a większym błędem standardowym szacowanego efektu. Metoda najbliższego sąsiada w wariancie bez zwracania jest również czuła – w pewnych okolicznościach – na kolejność, w jakiej ułożone są obserwacje w zbiorze. Jeśli pula kontrolna nie jest wystarczająco liczna, tj. jeśli w pewnych wartościach *propensity score* liczebność jednostek z grupy kontrolnej przewyższa liczebność jednostek z puli, to swoich odpowiedników znajdą tylko te jednostki interwencji, które będą dobierane w pary w pierwszej kolejności. Należy więc pamiętać, aby wykorzystując tę technikę, zapewnić wstępną losową kolejność obserwacji.

W swoim podstawowym wariancie opisywana metoda jest szczególnie efektywna, gdy pula kontrolna jest stosunkowo liczna i jednocześnie wszystkie jednostki w grupie interwencji mają szansę znaleźć swoich bliskich odpowiedników. W przypadku pokaźnego zbioru z jednostkami kontrolnymi korzystne może okazać się dopasowanie typu 1 do *n*. Może to przysłużyć się precyzji oszacowanego efektu, kosztem jednak zwiększenia jego obciążenia – wynikającego z tego, że mogą pojawić się gorsze dopasowania.

Przykładowe zestawienie pokazujące rezultat zastosowania metody najbliższego sąsiada – w jej podstawowej wersji – prezentuje tabela 4.

Tabela 4. Macierz danych z oszacowanymi wartościami *propensity score* i dobranymi jednostkami kontrolnymi

ID	*D*	X_1	X_2	X_3	…	X_n	*PS*1	*ID*2	*PS*2	*Delta*
278	1	1	51	1	…	1	0,040630	220	0,040637	0,000007
21	1	1	58	0	…	1	0,057154	110	0,057120	0,000034
170	1	1	53	0	…	1	0,058688	137	0,058601	0,000087
203	1	1	57	0	…	1	0,059505	173	0,059405	0,000100
219	1	2	55	1	…	1	0,081292	127	0,081596	0,000304
186	1	1	52	0	…	0	0,081745	248	0,082396	0,000650
32	1	2	57	1	…	1	0,089105	272	0,089100	0,000005
…	…	…	…	…	…	…	…	…	…	…
292	1	2	54	0	…	0	0,196158	136	0,202996	0,006837
6	0	2	51	0	…	0	0,217200	-	-	-
287	0	1	59	1	…	1	0,029192	-	-	-
279	0	1	59	1	…	1	0,029192	-	-	-
…	…	…	…	…	…	…	…	-	-	-
n	0	1	53	1	…	1	0,037421	-	-	-

Źródło: opracowanie własne.

ID to unikalny identyfikator jednostek w zbiorze. Zmienna *D* wskazuje, czy jednostka należy do grupy interwencji (1) czy do puli kontrolnej (0). $X_1 \ldots X_n$ to zmienne niezależne wykorzystane do oszacowania wartości *propensity score* (*PS*1) dla wszystkich jednostek w zbiorze. Kolumna *PS*2 to wartość *propensity score* dopasowanych jednostek z grupy kontrolnej. *Delta* to moduł różnicy między *PS*1 i *PS*2. Zbiór wyjściowy, dla badacza stosującego technikę PSM, składa się z zestawu zmiennych od *ID* do X_n. Z wykorzystaniem np. metody regresji logistycznej należy oszacować wartość *PS*1, gdzie zmienną zależną jest oczywiście *D* zaś $X_1 \ldots X_n$ to zmienne niezależne. Następnie dla jednostek z grupy interwencji należy wybrać z puli kontrolnej jednostki o najbliższym możliwym *PS*1. Jednostki te należy oznaczyć, np.

poprzez dopisanie do zbioru nowych zmiennych. W przykładzie powyżej są to zmienne *ID2* oraz *PS2* (*ID*2 – odpowiada wartości *ID* dopasowanej jednostki kontrolnej, a *PS*2 – odpowiada wartości *PS*1). Po znalezieniu jednostek kontrolnych dla wszystkich jednostek interwencji (a więc mając utworzone pary *ID-ID*2), można następnie zobaczyć jak różnią się one między sobą, badając moduł różnicy zmiennych *PS*1 i *PS*2 (*Delta*).

Metoda z limitem

Metoda z limitem (ang. *nearest neighbor with caliper*) jest modyfikacją metody najbliższego sąsiada. W sposób kontrolowany rozwiązuje problem możliwych „słabych" czy też „dalekich" dopasowań, poprzez zdefiniowanie maksymalnej akceptowalnej różnicy między wartościami *propensity score* dopasowywanych jednostek. Jeśli jakaś jednostka w grupie interwencji nie posiada swojego odpowiednika w puli kontrolnej, który różniłby się od niej o maksymalną wartość *c*, jednostka taka pozostaje bez pary. Problematyczne jest ustanowienie odpowiedniej wartości *c*. W literaturze przedmiotu prezentowane są różne podejścia, większość jednak opiera się na „regule kciuka". Rajeev H. Dehejia i Sadek Wahba zaprezentowali wyniki analiz, w których posługują się *c* przyjmującą wartości od 0,001 do 0,00001 (Dehejia i in. 2002: 158). Pokazali, że w ich przypadku różne wartości c w zbliżony sposób pozwalały zbalansować zmienne wykorzystane w modelu prawdopodobieństwa. Oczywiście im *c* jest mniejsze, tym grupa kontrolna będzie posiadała lepiej dopasowane jednostki. Z drugiej strony zwiększa się jednocześnie prawdopodobieństwo tego, że osoby z grupy interwencji pozostaną bez pary. W praktyce zwykle wariantowo stosuje się różne wartości *c*. Należy przy tym kontrolować skalę ewentualnego problem braku dopasowania jednostek kontrolnych dla części jednostek z grupy interwencji.

Metoda z promieniem

W opisanej wyżej metodzie z limitem, podobnie jak metodzie najbliższego sąsiada, można dokonywać doboru jednostek do grupy kontrolnej ze zwracaniem lub bez. Możliwe są też dopasowania 1 do 1 lub 1 do n. Szczególną postacią tych ostatnich jest metoda z promieniem (ang. *radius matching*). W wariancie tym do każdej jednostki z grupy interwencji dobierane są wszystkie obserwacje z puli kontrolnej, których *propensity score* nie różni się więcej niż przyjęta wartość *c*. Innymi słowy dobierane są wszystkie jednostki w obrębie określonego promienia. Podejście to eliminuje możliwość wystąpienia słabych dopasowań i pozwala jednocześnie zwiększyć precyzję estymacji efektu interwencji, zwiększa się bowiem liczebność grupy kontrolnej. Liczba jednostek kontrolnych dobranych dla każdego przypadku w grupie eksperymentalnej jest oczywiście zmienną losową.

Metoda z promieniem będzie szczególnie efektywna w przypadku, gdy grupa kontrolna jest znacząco większa od grupy eksperymentalnej. Kłopotliwe wydaje się jednak stosowanie tej metody, w wariancie bez zwracania. Metoda ta jest bardzo „chciwa" – każda jednostka eksperymentalna wybiera wszystkie obserwacje z grupy kontrolnej, znajdujące się w ustalonym promieniu. Jeśli jednostki eksperymentalne mają bliskie wartości *propensity score*, skutkować to będzie tym, że jakość dopasowania jednostek kontrolnych będzie wysoka tylko dla wybranych przypadków – konkretnie tych, które są łączone w pierwszej kolejności. Z tego też względu dużą wagę odgrywa kolejność, w jakiej uporządkowane są obserwacje, przed rozpoczęciem procedury łączenia (problem ten ma tu nawet większe znaczenie niż w przypadku stosowania techniki najbliższego sąsiada w wariancie bez zwracania). Stąd rekomendacja, aby dokonać losowego uporządkowania zbiorów przed rozpoczęciem doboru grupy

kontrolnej. Częściowym rozwiązaniem tego problemu może być również ustalenie odpowiednio niskiej wartości *c*. Czynność taka może jednak wywołać problemy ze znalezieniem wystarczająco podobnych jednostek kontrolnych. Dlatego też wydaje się zasadnym, aby metodę z promieniem stosować raczej w wariancie ze zwracaniem.

Metoda Kernel

W dotychczas opisanych technikach możliwe były łączenia obserwacji typu 1 do 1 lub 1 do *n*, przy czym w przypadku tych ostatnich owo „*n*" w praktyce powinno być raczej niewielką liczbą (w stosunku do całego zbioru danych)[38]. Metoda Kernel (ang. *Kernel matching*) podchodzi w inny sposób do wykorzystania informacji, którą „niosą" ze sobą jednostki z puli kontrolnej. Zgodnie z jej ogólną ideą do oszacowania kontrfaktycznego wyniku dla jednostki *i* wykorzystywane są wyniki wszystkich jednostek, które należą do puli kontrolnej, a więc jednostek nieobjętych ocenianą interwencją. Innymi słowy grupa kontrolna jest tożsama z pulą kontrolną – w przeciwieństwie do innych metod łączenia nie następuje tu wybór grupy kontrolnej. Przy czym znaczenie poszczególnych jednostek kontrolnych w kształtowaniu kontrfaktycznego wyniku zależy od ich bliskości do porównywanej jednostki interwencji. Bliskość wyrażana jest poprzez wartość różnicy *propensity score*, jaka dzieli obie porównywane jednostki. W związku z tym, jeśli osoba z puli kontrolnej znacząco różni się od jednostki z grupy interwencji, to znaczenie tej pierwszej jest znikome w szacowaniu efektu interwencji. Udział (waga) jednostek kontrolnych liczony jest zgodnie z następującym wzorem:

$$w_{ij} = \frac{K\left(\frac{P(X_i) - P(X_j)}{h}\right)}{\sum_{j \in \{d=0\}} K\left(\frac{P(X_i) - P(X_j)}{h}\right)}, \tag{2.03}$$

gdzie $P(X_i)$ i $P(X_j)$ to wartości *propensity score* odpowiednio dla jednostki z grupy interwencji oraz grupy kontrolnej, zaś h odpowiada za szerokość pasma (ang. *bandwidth*, jest to tzw. parametr wygładzający – ang. *smoothing parameter*). K to wybrana funkcja prawdopodobieństwa – najczęściej z rozkładu normalnego (Gaussa), wyrażona w takim przypadku następującym równaniem:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}u^2} \tag{2.04}$$

gdzie π oraz *e* to stałe.

W praktyce, wykorzystując metodę Kernel, badacz musi podjąć dwie decyzje. Po pierwsze musi wybrać postać funkcji Kernel oraz po drugie określić wielkość parametru h. Jak piszą Marco Caliendo i Sabine Kopeinig, w pierwszym przypadku wybór jest raczej mało istotny. Większe znaczenie ma określenie wielkości parametru wygładzającego (Caliendo i in., 2005: 11).

[38] N może być liczbą z góry wskazaną przez badacza lub zmienną losową w przypadku wykorzystania metody *radius matching*.

Wybór metody doboru grupy kontrolnej

Istotnym pozostaje pytanie o to, jaki konkretny algorytm doboru grupy kontrolnej zastosować w praktyce. Nie ma tu jednoznacznej odpowiedzi, wszystko bowiem zależy od posiadanego zbioru danych oraz od rozkładu *propensity score* w grupie interwencji i w grupie kontrolnej. Wyniki powinny być mniej czułe na wybór konkretnego algorytmu w przypadku posiadania bogatego zbioru danych w puli kontrolnej, gdy każda z jednostek z grupy interwencji ma szansę znalezienia swojego odpowiednika (Bryson i in., 2002: 27). Bardziej krytyczny staje się wybór metody, gdy zbiory mają małe liczebności i np. występuje zagrożenie wystąpienia problemu wspólnego przedziału określoności. Oczywiste jest np., że nie ma sensu stosować wariantu bez zwracania w sytuacji, gdy pula kontrolna w porównaniu z grupą interwencji jest nieliczna. Z kolei, jeśli pula kontrolna znacznie przewyższa liczebność grupy eksperymentalnej, warto zastanowić się nad zastosowaniem metody z wariantem łączenia typu 1 do n, takiej jak np. metoda z promieniem, czy metoda Kernel. W praktyce nic nie stoi na przeszkodzie – i takie podejście jest zalecane – aby zastosować symulacje doboru grupy kontrolnej z wykorzystaniem różnych metod łączenia. Należy przy tym monitorować jakość dobranej grupy kontrolnej – o czym mowa w kolejnej części – oraz stopień, w jakim wynik (szacowany efekt interwencji) jest podatny na wprowadzane zmiany.

Znamienne jest, że wybór metody łączenia przekłada się zwykle na dwa parametry. Są nimi wielkość potencjalnego obciążenia szacunku interwencji oraz precyzja estymacji, która wyraża się w wielkości wariancji. Na pierwszy z parametrów wpływ będzie miało to, na ile dobrane do grupy kontrolnej jednostki różnią się od jednostek w grupie interwencji. Wzrostu obciążenia szacowanego efektu interwencji można się więc spodziewać, gdy dopasowania będą słabe, tj. gdy jednostki będą się między sobą znacząco różnić pod względem wartości *propensity score*. Najbardziej podatna na ten problem jest metoda najbliższego sąsiada w wariancie bez zwracania. Pozostałe metody w różny sposób radzą sobie z tą kwestią – metoda z limitem, tak jak jej pochodna metoda z promieniem, wyznaczają maksymalną dopuszczalną różnicę między jednostkami, z kolei metoda Kernel pomija „odstające" obserwacje, nadając im odpowiednio mniejsze wagi.

Na drugi ze wskazanych parametrów, tj. na precyzję estymacji, wpływ będzie miała liczebność dobranej grupy kontrolnej. Wraz z jej wzrostem precyzja rośnie (maleje wariancja estymatora). Większej precyzji szacunków efektu interwencji sprzyjają więc wszystkie metody, w których wykorzystywane jest łączenie typu 1 do n.

Zależności pomiędzy ww. parametrami zostały przestawione na przykładzie zestawienia wybranych metod z najprostszym sposobem doboru grupy kontrolnej, tj. z metodą najbliższego sąsiada w wariancie bez zwracania oraz z doborem typu 1 do 1. Wyniki zestawienia zaprezentowane zostały w tabeli 5. Dla przykładu metoda najbliższego sąsiada w najprostszej postaci (NS-BZ-1/1), w porównaniu z metodą najbliższego sąsiada w wariancie ze zwracaniem oraz z możliwością wykorzystania *n* jednostek kontrolnych (NS-ZZ-1/n) będzie miała najprawdopodobniej większe obciążenie szacunków interwencji. Powodem tego jest możliwość wystąpienia dalekich dopasowań, które w drugiej z metod zminimalizowane zostaną poprzez możliwość wykorzystania tej samej jednostki kontrolnej więcej niż raz. Jednocześnie precyzja oszacowania z wykorzystaniem metody NS-ZZ-1/n powinna być większa z uwagi na większą liczebność grupy kontrolnej.

Tabela 5. Konsekwencje różnych metod i wariantów łączenia obserwacji

	Porównywana metoda	Obciążenie	Precyzja
Metoda najbliższego sąsiada – bez zwracania, z łączeniem typu 1 do 1 NS-BZ 1/1 vs.:	NS-BZ-1/n	–/+	–/+
	NS-ZZ-1/1	+/–	–/+
	NS-ZZ-1/n	+/–	–/+
	ML-BZ	+/–	+/–
	ML-ZZ	+/–	+/–
	MP-ZZ	+/–	–/+
	MK	+/–	–/+

Legenda: NS-BZ-1/n – metoda najbliższego sąsiada bez zwracania, z możliwością łączenia typu 1 do n;
NS-ZZ-1/1 – metoda najbliższego sąsiada ze zwracaniem, z możliwością łączenia typu 1 do 1;
NS-ZZ-1/n – metoda najbliższego sąsiada ze zwracaniem, z możliwością łączenia typu 1 do n;
ML-BZ – metoda z limitem bez zwracania; ML-ZZ – metoda z limitem ze zwracaniem;
MP-ZZ – metoda z promieniem ze zwracaniem; MK – metoda Kernel.

„+" oznacza względny, w ramach porównania, wzrost, zaś „–" oznacza względny, w ramach porównania, spadek.

Źródło: opracowanie własne na podstawie Marco Caliendo, Sabine Kopeinig *Some Practical Guidance for the Implementation of Propensity Score Matching* 2005.

Ocena jakości łączenia

Po dokonaniu łączenia jednostek należy sprawdzić, czy w wyniku zastosowanej procedury udało się uzyskać zbalansowane rozkłady zmiennych włączonych do modelu w grupie działania i w grupie kontrolnej. Ogólna idea polega tu na porównaniu sytuacji sprzed łączenia, z sytuacją uzyskaną w wyniku zastosowania wybranego algorytmu selekcji grupy kontrolnej. W pierwszym kroku porównywana jest więc grupa interwencji z całą pulą kontrolną (tj. całą dostępną grupą jednostek nieuczestniczących w ocenianej interwencji). W kroku drugim należy porównać grupę interwencji z dobraną grupą kontrolną. Głównym przedmiotem zainteresowania jest stopień, w jakim udało się zminimalizować pierwotne różnice pomiędzy jednostkami w puli kontrolnej i jednostkami w grupie interwencji. O ile różnice między obiema grupami pozostają znaczące, to jest to sygnał do tego, aby cofnąć się do poprzednich etapów zastosowania techniki PSM. Może być to powrót np. do fazy wyboru algorytmu łączenia, czy nawet do momentu szacowania *propensity score* (Caliendo i in., 2005: 16).

Jeden ze sposobów weryfikacji jakości doboru grupy kontrolnej przedstawiają P. Rosenbaum i D. Rubin. Autorzy proponują, aby dokonać oceny doboru grupy kontrolnej poprzez przeanalizowanie zmiany wielkości standaryzowanego obciążenia zmiennych (a więc stopnia, w jakim różnią się rozkłady poszczególnych zmiennych w grupie eksperymentalnej i grupie kontrolnej). Za miarę obciążenia przyjmują oni wystandaryzowaną procentową różnicę pomiędzy średnimi każdej zmiennej predykcyjnej dla grupy poddanej interwencji i osób z puli kontrolnej (Rubin i in., 2006: 212):

$$SD_{przed} = \frac{100\left(\overline{X}_1 - \overline{X}_{0P}\right)}{\sqrt{\frac{\left(s_1^2 + s_{0P}^2\right)}{2}}} \tag{2.05}$$

gdzie $\overline{X}_1$ i $\overline{X}_{0P}$ to średnie analizowanej zmiennej w grupie działania i w puli kontrolnej, zaś s_1^2 i s_{0P}^2 to odpowiednie oszacowania wariancji. Wyliczone na podstawie powyższego wzoru wartości odnoszą się więc do różnic na poszczególnych zmiennych, jakie dzielą grupę eksperymentalną i pulę kontrolną. Odpowiednio zmodyfikowaną różnicę należy oszacować już po dokonaniu doboru grupy kontrolnej:

$$SD_{po} = \frac{100\left(\overline{X}_1 - \overline{X}_{0K}\right)}{\sqrt{\frac{\left(s_1^2 + s_{0P}^2\right)}{2}}} \tag{2.06}$$

gdzie $\overline{X}_{0K}$ to średnia dla grupy kontrolnej[39]. Inne parametry zostały zdefiniowane wcześniej. Trzeba zauważyć, że mianownik w obu wyrażeniach pozostał bez zmian. W opisanym podejściu pozostaje problematyczne ustalenie akceptowalnej wartości obciążenia zmiennej. W literaturze przedmiotu spotyka się jednak opinie, że zredukowanie obciążenia poniżej 3%–5% jest satysfakcjonujące (Caliendo i in., 2005: 15).

Dla sprawdzenia, czy istnieją istotne różnice pomiędzy średnimi poszczególnych zmiennych włączonych do modelu, Rosenbaum i Rubin (Rubin i in, 2006: 213) proponują zastosować również test t. Przewiduje się, że test t wykaże występowanie statystycznie istotnych różnic pomiędzy średnimi w grupie interwencji i puli kontrolnej. Po dokonaniu doboru taka sytuacja nie jest pożądana w przypadku porównywania pierwszej z grup z grupą kontrolną[40].

Ograniczenia techniki PSM

Ograniczenia techniki PSM wypływają wprost z jej założeń. Po pierwsze należy mieć świadomość konsekwencji sposobu, w jaki został zdefiniowany efekt przyczynowy. Wprost wynika z niego założenie SUTVA przedstawione w rozdziale pierwszym. Oprócz tego w szacowanym efekcie przyczynowym nie uwzględnia się wpływu podjętego działania na równowagę ogólną – tj. na otoczenie pozostające poza interwencją.

Krytycznym elementem techniki PSM jest poprawność przyjmowanego założenia o warunkowej niezależności (Strawiński, 2008: 216). Jak zostało powiedziane, jest ono nietestowalne. Wyjątek stanowią sytuacje, w których wyniki analiz przeprowadzonych w badaniach obserwacyjnych możemy porównać z wynikami analiz wykorzystujących dane eksperymentalne (np. Heckman i in. 1997, s. 606; Dehejia i in. 2002, s. 151). Natomiast niespełnienie powyższego założenia grozi błędnym oszacowaniem efektu przyczynowego.

Technika PSM eliminuje obciążenie szacunków interwencji, wynikające z tego, że grupa działania różni się od grupy kontrolnej w zakresie obserwowanych charakterystyk. Problematyczne pozostaje ewentualne obciążenie, wynikające z występujących różnic na zmiennych nieobserwowalnych. W przeciwieństwie do metody eksperymentalnej, która zapewnia zrównanie rozkładu na wszystkich cechach, techniki polegające na łączeniu jednostek zapewniają zbalansowanie podobieństwa grup

[39] Problematyczne staje się wykorzystanie opisanego podejścia w przypadku modeli, w których wykorzystywane są zmienne jakościowe. Dla nich nie można policzyć średniej oraz opisywanego dalej testu t. Pośrednim rozwiązaniem problemu jest zastosowanie procedury dychotomizacji, tj. rozbicia zmiennych jakościowych na wiele zmiennych dychotomicznych. Należy jednak ostrożnie podchodzić do standaryzacji takich zmiennych z uwagi na współzależność średniej i wariancji.

[40] Przykład oceny obciążania z wykorzystaniem standaryzowanych różnic średnich oraz testu t pokazany został w kolejnym rozdziale, w tabeli 7 oraz w tabeli 14.

jedynie w zakresie zmiennych, które zostały uwzględnione w modelu prawdopodobieństwa (i do pewnego stopnia zmienne nieuwzględnione, o ile tylko pozostają one w związku z cechami włączonymi do procesu badawczego). Ma to fundamentalne znaczenie dla trafności wyników analiz realizowanych w badaniach obserwacyjnych.

Wykorzystanie techniki PSM jest również warunkowane dostępnością danych. W praktyce, aby dobrać grupę kontrolną, która nie różni się od grupy eksperymentalnej, niezbędne są duże zbiory danych. Tym większy zbiór jest potrzebny, im większa jest grupa eksperymentalna i im bardziej różni się ona od populacji, z której pochodzi. Ma to oczywiście przełożenie na koszty, zwłaszcza w przypadku, gdy analizy realizowane są na danych pierwotnych, które uzyskiwane są w drodze badań terenowych.

Przykład zastosowania techniki PSM

W poniższym rozdziale zaprezentowany został przykład wykorzystania techniki PSM w ewaluacji przeprowadzonej w roku 2007 na zlecenie Polskiej Agencji Rozwoju Przedsiębiorczości[41]. Prezentowana tu ewaluacja stanowi, jak dotąd, jeden z nielicznych przypadków badania, w którym podjęto się pomiaru oddziaływania przyczynowego interwencji publicznej, finansowanej ze środków Unii Europejskiej. Względna rzadkość takiego podejścia wynika z co najmniej dwóch, nie do końca rozłącznych, względów. Po pierwsze ewaluacja, mimo że bazuje na dorobku badań społecznych, jest w świadomości osób czy instytucji ją wykonujących/zlecających, wciąż nowym pojęciem. Wynika to z faktu, że badania ewaluacyjne – sensu stricte – zagościły na dobre w polskiej praktyce, wraz z otrzymaniem przez Polskę pierwszych funduszy unijnych. Obecnie mamy do czynienia z pewnego rodzaju transferem metodologii badań społecznych na grunt badań ewaluacyjnych. Po drugie pomiar nastawiony na uchwycenie przyczynowego oddziaływania zdarzeń, w swojej najbardziej podstawowej wersji, możliwy jest dopiero po faktycznym zrealizowaniu interwencji. Z uwagi na dosyć krótką historię naszego członkostwa w UE, liczba ewaluacji, w których możliwa jest empiryczna weryfikacja występowania relacji przyczynowych – między podejmowanymi działaniami a obserwowanymi efektami – jest jeszcze stosunkowo niewielka.

W rozdziale przedstawiony został również drugi przykład ewaluacji, w której zastosowano technikę PSM. Przykład ten ma pokazać, jak ważne jest poprawne zdefiniowanie modelu prawdopodobieństwa i co ewentualnie grozi w przypadku niespełnienia założenia o warunkowej niezależności (*CIA*).

Phare Spójność Społeczna i Gospodarcza – komponent RZL

Program Phare Spójność Społeczna i Gospodarcza – komponent Rozwój Zasobów Ludzkich (Phare SSG RZL) – realizowany był w Polsce w ramach czterech kolejnych edycji w latach 2002–2006[42]. Jego głównym celem było zmniejszenie opóźnień i nierównomierności w rozwoju regionów Polski, poprzez rozwiązywanie problemów społecznych związanych z rynkiem pracy. Ogólnie projekty realizowane w ramach Phare SSG RZL miały na celu wsparcie czterech filarów Europejskiej Strategii Zatrudnienia, którymi są: wzrost możliwości uzyskania zatrudnienia, rozwój przedsiębiorczości, wzrost zdolności adaptacyjnych (umiejętności dostosowania się do nowych warunków rynku pracy) oraz równość szans na rynku pracy (mężczyzn, kobiet, osób niepełnosprawnych)[43]. Ze wsparcia programu mogły skorzystać przede wszystkim osoby bezrobotne oraz zagrożone bezrobociem.

W grupie państw Europy Środkowej i Wschodniej Polska była największym beneficjentem pomocy finansowej Phare. Łącznie nasz kraj uzyskał w ramach czterech edycji samego Phare SSG RZL[44]

---

41 *www.parp.gov.pl* stan na 16.03.2009 r.

42 Numery kolejnych edycji to odpowiednio 2000, 2001, 2002 i 2003.

43 *http://ec.europa.eu/employment_social/employment_strategy/index_en.htm* stan na 15.03.2009 r.

44 Oprócz komponentu Rozwój Zasobów Ludzkich, w ramach Phare SSG realizowane były komponenty: Infrastruktura oraz Rozwój Małych i Średnich Przedsiębiorstw.

wsparcie w wysokości ok. 170 mln EUR. W ramach tej kwoty wsparciem objętych zostało ok. 168 tys. osób (PARP i in., 2006: 99). W obliczu tak dużej skali zaangażowanych zasobów niezwykle istotne było przeprowadzenie oceny rzeczywistego wpływu działań realizowanych w ramach Phare. Ocena, zgodnie z dyrektywami unijnymi, realizowana była w ramach przeprowadzanych ewaluacji programu.

W strukturze organizacyjnej za realizację kolejnych edycji Phare SSG RZL odpowiedzialna była Polska Agencja Rozwoju Przedsiębiorczości. Ona również przeprowadziła ewaluacje wdrażanych przez siebie programów. Badania miały charakter tzw. ewaluacji ex-post, tj. były realizowane po zakończeniu wdrażania kolejnych edycji programu (zwykle rok po ich zamknięciu). Począwszy od ewaluacji Phare 2001 SSG RZL do oceny wpływu działań zaczęto wykorzystywać technikę *propensity score matching*[45]. Jednocześnie były to pierwsze przykłady zastosowania tej techniki w badaniach ewaluacyjnych na gruncie polskim. Niniejszy rozdział zawiera prezentację wykorzystania metody PSM w ewaluacji jednego z projektów realizowanych w ramach ostatniej edycji Phare SSG RZL 2003 – projektu Alternatywa II.

Projekt Alternatywa II

Projekt Alternatywa II realizowany był od 2005 do 2006 roku w wybranych z całej Polski 59 powiatach. Głównym celem[46] projektu było zapobieganie bezrobociu wśród osób młodych. Cel ten miał być osiągnięty dzięki zwiększeniu potencjału uczestników projektu do bycia zatrudnionym. Przyjęto, że należy w tym celu pomóc młodym osobom w:

a) dostosowywaniu umiejętności i kwalifikacji zawodowych do zmieniających się potrzeb rynku pracy – dotyczyło to zwłaszcza umiejętności związanych z technologiami informatycznymi;
b) zdobyciu doświadczenia zawodowego.

W projekcie zaoferowano młodym, bezrobotnym osobom wiele kompleksowych instrumentów wsparcia. Wśród nich znalazło się przygotowanie tzw. Indywidualnych Planów Działania (IPD), na które składało się m.in. opracowanie kompletnej ścieżki rozwoju zawodowego dla każdego uczestnika projektu. IPD przygotowane były przy współudziale doradców zawodowych z powiatowych urzędów pracy i miały uwzględniać oczekiwania, plany zawodowe, predyspozycje, ewentualne doświadczenie zawodowe młodych osób.

Do określonych planów dobierane były niezbędne działania, które miały zapewnić realizację postawionych w IPD celów. W ramach projektu umożliwiono beneficjentom realizację działań przewidzianych w opracowanym planie. Do dyspozycji były różnego rodzaju szkolenia (zgodne z kierunkami wyznaczonymi w IPD), w tym szkolenia praktyczne (staże i kursy połączone ze stażami) w miejscu pracy[47], doradztwo biznesowe, pośrednictwo pracy.

Obowiązki wynikające z realizacji projektu były podzielone pomiędzy Wykonawcę (wybrana przez PARP firma komercyjna) oraz powiatowe urzędy pracy (PUP). Co istotne do projektu zaproszone były powiaty, które miały szczególnie trudną sytuację na rynku pracy, tj. takie, w których odsetek bezrobot-

[45] Wykonawcami badań była firma PBS DGA (w przypadku ewaluacji Phare SSG RZL 2001 oraz Phare SSG RZL 2002) oraz konsorcjum firm PAG Uniconsult i ARC Rynek i Opinia (w przypadku ewaluacji Phare SSG RZL 2003).
[46] *Fiszka projektowa: PL National Human Resource Development Project*.
[47] Szkolenia praktyczne miały odbywać się poza powiatem – miejscem zarejestrowania beneficjenta jako bezrobotnego – tak, aby dodatkowo stymulować mobilność zawodową osób poszukujących pracy.

nej młodzieży w ogólnej liczbie bezrobotnych przekraczał 18% i gdzie stopa bezrobocia wynosiła powyżej 20%.

W projekcie mogły wziąć udział osoby z powiatu, w którym był on realizowany. Zastosowano dodatkowo dwa formalne kryteria rekrutacyjne. Pierwszym z nich był obowiązek bycia zarejestrowanym w powiatowym urzędzie pracy jako osoba bezrobotna. Drugim było kryterium wieku – w projekcie mogły uczestniczyć osoby, które nie ukończyły 25 roku życia lub które nie ukończyły 27 roku życia i ukończyły studia wyższe w okresie 12 miesięcy przed dokonaniem rejestracji w rejestrze bezrobotnych.

W sumie w projekcie uczestniczyło 5657 młodych osób bezrobotnych, a całkowite wydatki projektu wyniosły 4 090 702,03 euro.

Założenia ewaluacji projektu Alternatywa II

Celem ewaluacji zrealizowanej po zakończeniu wdrażania programu, w 2007 roku, była ocena całego Phare 2003 SSZ RZL. Kompleksowa ocena została dokonana w rozbiciu na projekty realizowane w ramach programu. W szczególności ocenie podlegał projekt Alternatywa II. Z punktu widzenia niniejszej pracy istotny jest tylko jeden z poruszonych w ewaluacji tematów, a konkretnie ocena w wymiarze wpływu projektu Alternatywa II na zmianę sytuacji zawodowej jego beneficjentów. W tym kontekście w ewaluacji poszukiwano odpowiedzi na pytanie o to, jaka jest sytuacja zawodowa beneficjentów projektu po jego zakończeniu oraz na ile ewentualną zmianę (poprawę) sytuacji zawodowej można przypisać oddziaływaniu samego projektu Alternatywa II. Inaczej mówiąc, celem szczegółowym ewaluacji było oszacowanie efektu netto działań podjętych w ramach projektu. Przyjęte podejście wymagało identyfikacji relacji przyczynowej pomiędzy działaniem, jakim była realizacja projektu a obserwowanymi efektami w grupie jego uczestników. Zlecający ewaluację – PARP – zaleciła wykorzystanie w tym celu techniki PSM.

Problem selekcji

W rozdziale pierwszym zarysowany został problem selekcji, który w przypadku programów rynku pracy przyjmuje realne oblicze. Charakter projektu Alternatywa II wskazywał na to, iż różne czynniki mogły wpływać na to, że beneficjenci Alternatywy II nie byli losowymi reprezentantami populacji osób bezrobotnych. Przede wszystkim projekt realizowany był tylko w wybranych w powiatach, o szczególnie wysokiej stopie bezrobocia. Ponadto zastosowano dwa kryteria formalne dla uczestników projektu – obligatoryjna rejestracja w urzędzie pracy oraz wiek – nieukończone 27 lat. Narzucało to z góry zawężenie populacji, która mogłaby posłużyć jako grupa porównawcza dla beneficjentów projektu. O ile powyższe kryteria nie stanowiły problemu, gdyż były dane wprost, o tyle bardziej kłopotliwe pozostawały nieformalne kryteria selekcji, nigdzie nieujęte. Należy zauważyć, że uczestnictwo w projekcie Alternatywa II było działaniem dobrowolnym. Projekt miał charakter otwarty, w którym mogły wziąć udział wszystkie osoby bezrobotne spełniające opisane kryteria formalne. Z drugiej strony był to program realizowany przez komercyjną firmę szkoleniową, która na koniec była „rozliczana" z osiągniętych efektów (operacyjnym celem projektu było osiągnięcie poziomu zatrudnialności wśród jego beneficjentów na poziomie 40%). Co istotne, rekrutacja odbywała się za pośrednictwem powiatowych urzędów pracy, a więc szczególną szansę na uczestnictwo miały oso-

by pozostające w kontakcie z PUP (co potwierdzają prowadzone równolegle w ewaluacji badania ankietowe[48]). Oprócz powyższych zjawisk występować mogły różnego rodzaju zaburzenia po stronie beneficjentów pomocy, takie jak samoselekcja do programu osób bardziej aktywnych, o większej motywacji do podjęcia pracy, świadomych konieczności podnoszenia własnych kwalifikacji zawodowych itp. Chcąc oszacować rzeczywiste efekty osiągnięte przez projekt, należało te potencjalne czynniki uwzględnić i zminimalizować ich wpływ podczas szacowania efektu przyczynowego działania. Dlatego w przypadku ewaluacji projektu Alternatywa II zdecydowano się wykorzystać technikę PSM w celu wyeliminowania potencjalnego obciążenia selekcyjnego.

Dane wykorzystane w ewaluacji

Tak jak zostało to opisane w rozdziale drugim, jedną z krytycznych części składowych, która warunkuje możliwość wykorzystania techniki PSM, jest dostępność odpowiednich danych. Informacje gromadzone w ramach systemu wdrażania i monitorowania Projektu Alternatywa II były niewystarczające do przeprowadzenia analiz z wykorzystaniem techniki PSM[49]. Dysponowano co prawda bazą beneficjentów, którzy otrzymali wsparcie. Znana była też w ograniczonym zakresie charakterystyka tych osób, ze względu na szereg cech społeczno-demograficznych. Brakowało natomiast rzeczy podstawowej, tj. zbioru osób, który mógłby posłużyć do utworzenia grupy kontrolnej. Utworzenie takiej grupy ex post, np. w wyniku realizacji dodatkowych badań terenowych, byłoby oczywiście bardzo kosztowne, o ile w ogóle możliwe (ewaluacja realizowana była rok po zakończeniu wdrażania projektu, sam projekt również trwał rok, zatem dane, które miałyby zostać wykorzystane do analiz, powinny odzwierciedlać stan, który miał miejsce dwa lata przed realizacją badania ewaluacyjnego). Konieczne było więc raczej znalezienie odpowiedniego zbioru z danymi zastanymi. Zbiór taki, zgodnie z tym, o czym mowa była w rozdziale drugim, powinien spełniać następujące warunki:

- dane w nim zawarte powinny zawierać informację o osobach, które były uczestnikami Projektu Alternatywa II,
- zbiór danych powinien zawierać „odpowiednio" większy zbiór osób niebędących beneficjentami Phare, jednak pochodzący z tej samej populacji oraz z tego samego środowiska ekonomicznego co uczestnicy Phare,
- zbiór ten powinien zawierać „odpowiednio bogate" informacje charakteryzujące zgromadzone w nim osoby, w tym w szczególności powinien zawierać informacje na temat ich sytuacji zawodowej oraz zmienne, które powinny być odpowiedzialne jednocześnie za udział w projekcie, jak i późniejszy jego efekt – w tym przypadku zmianę sytuacji zawodowej,
- dane zawarte w zbiorze musiały być aktualne w momencie rozpoczęcia realizacji projektu (z wyłączeniem zmiennych odnoszących się do badanych efektów),
- dane w grupie beneficjentów i w potencjalnej grupie kontrolnej powinny być zbierane tą samą metodą.

[48] Blisko 90% respondentów, którymi byli beneficjenci projektu Alternatywa II, na pytanie o źródło informacji o projekcie wskazuje PUP (PARP, 2007: 113).

[49] Jest to szerszy problem związany z brakiem kompleksowego podejścia do programowania interwencji publicznych.

Jak się okazało, wymagania te mogły zostać spełnione przez funkcjonujący w Polsce System Informatyczny PULS[50]. SI PULS[51] jest systemem ewidencji osób bezrobotnych wykorzystywanym w około 90% powiatowych urzędów pracy w Polsce. W ramach Systemu, oddzielnie dla każdego PUP, gromadzone są dane dotyczące zarejestrowanych w nim osób bezrobotnych. W bazie danych gromadzone są informacje dotyczące historii zatrudnienia każdej osoby, tj. okresy, w których osoba była zarejestrowana jako bezrobotna, pracująca lub wykreślona z ewidencji bezrobotnych. Oprócz tego w Systemie zbierane są wybrane charakterystyki społeczno-demograficzne zarejestrowanych w nim osób, takie jak wykształcenie, kwalifikacje, umiejętności, doświadczenie zawodowe, sytuacja rodzinna i wiele innych.

Dane pochodzące z SI PULS pozwoliły na pozyskanie informacji na temat większości beneficjentów projektu Alternatywa II. Zebrane dane pochodziły z 55 na 59 powiatowych urzędów pracy biorących udział w projekcie. Pozostałe 4 powiatowe urzędy pracy[52] nie posiadały Systemu Informatycznego PULS. W ten sposób uzyskano w sumie informacje na temat 5065 uczestników projektu Alternatywa II, co stanowiło 90% wszystkich jego beneficjentów. Jednocześnie, w wyniku agregacji danych z poszczególnych powiatów, udało się uzyskać potencjalną grupę kontrolną składającą się z 126 633 osób. Były to osoby zarejestrowane jako bezrobotne, w czasie, gdy w poszczególnych powiatowych urzędach pracy realizowany był nabór do projektu[53]. Ponadto były to osoby, które miały nie więcej niż 27 lat (zgodnie z formalnymi kryteriami selekcji do projektu). Szczegółowe informacje na temat beneficjentów zidentyfikowanych w SI PULS, w rozbiciu na poszczególne województwa, prezentuje poniższa tabela:

Tabela 6. Regionalny rozkład beneficjentów Alternatywy II i jednostek w puli kontrolnej

Województwo	Beneficjenci zidentyfikowani w SI PULS	Pula kontrolna	Udział beneficjentów w puli kontrolnej
Dolnośląskie	249	11 373	2%
Kujawsko-pomorskie	957	23 082	4%
Lubelskie	159	4 698	3%
Lubuskie	270	6 472	4%
Łódzkie	448	7 029	6%
Mazowieckie	436	12 206	4%
Opolskie	202	2 742	7%
Podlaskie	169	2 625	6%
Pomorskie	527	13 332	4%
Śląskie	222	7 757	3%
Warmińsko-mazurskie	868	20 207	4%
Wielkopolskie	402	12 601	3%
Zachodniopomorskie	156	2 509	6%
Suma	5065	12 6633	4%

Źródło: opracowanie własne na podstawie raportu z badań: PAG Uniconsult, ARC Rynek i Opinia (2007), Ewaluacja ex-post Phare 2003 Spójność Społeczna i Gospodarcza komponent Rozwój Zasobów Ludzkich.

[50] Informacje o SI PULS dostępne są pod adresem: h*ttp://puls.computerland.pl/index.php?option=com_docman &task=cat_view&gid=38&Itemid=31* na dzień 15.12.2008 r.

[51] Dane zawarte w SI PULS zostały wykorzystane również podczas wcześniejszych dwóch ewaluacji, w których zastosowana była technika PSM – Phare 2001 SSG RZL oraz Phare 2002 SSG RZL.

[52] PUP w powiecie: gnieźnieńskim, ostródzkim, strzelecko-drezdeneckim i m. Jaworzno.

[53] Ewaluacja ex-post Phare 2003 Spójność Społeczna i Gospodarcza komponent Rozwój Zasobów Ludzkich.

Tabela 6 wskazuje liczbę beneficjentów projektu, jaką udało się zidentyfikować w SI PULS. Kolumna Pula kontrolna wskazuje, ile osób łącznie było zarejestrowanych w SI PULS w momencie rozpoczęcia rekrutacji do projektu Alternatywa II (wyłączając beneficjentów). Jak widać udział beneficjentów w ogólnej populacji jest niewielki. Ma to duże znaczenie dla praktycznej możliwości zastosowania techniki PSM. W dużym stopniu rzutuje również na jakość dopasowania grupy kontrolnej, o czym mowa będzie dalej.

Co istotne, dane zawarte w SI PULS umożliwiły ustalenie dla każdego beneficjenta jego sytuacji zawodowej w kolejnych miesiącach od zakończeniu udziału w programie (z dokładnością co do dnia). Osoby w puli kontrolnej naturalnie nie miały określonej ani daty rozpoczęcia, ani zakończenia udziału w projekcie. W związku z tym daty te zostały do zbioru imputowane[54]. Data rozpoczęcia udziału w projekcie była każdorazowo losowana z przedziału od najwcześniejszej daty rozpoczęcia udziału w projekcie odnotowanej u beneficjentów w powiecie, z którego pochodziła dana osoba z puli kontrolnej, do najpóźniejszej daty rozpoczęcia udziału w projekcie w danym powiecie. Z kolei datę zakończenia udziału przypisywano osobom z puli kontrolnej w sposób losowy, posługując się średnią długością udziału w Alternatywie II beneficjentów z danego powiatu, powiększoną lub pomniejszoną o wartość odchylenia standardowego długości trwania projektu w powiecie (wprowadziło to bardziej rzeczywiste zróżnicowanie teoretycznej długości udziału jednostek kontrolnych w projekcie).

Wybór zmiennych do modelu

System Informatyczny PULS zawiera bogaty zestaw informacji na temat zarejestrowanych w nim osób bezrobotnych. Jednak z punktu widzenia oszacowania efektu przyczynowego projektu Alternatywa II – zgodnie z wymaganiami techniki PSM – istotna była identyfikacja takich zmiennych, które mogły wpływać zarówno na udział jednostek w projekcie, jak i na obserwowane efekty. Za zmienną wynikową przyjęto wskaźnik wyrażający cel główny Phare SSG RZL, jakim był odsetek osób pracujących mierzony po zakończeniu trwania programu. W przypadku zmiennych predykcyjnych wiele cennych wskazówek przynoszą badania empiryczne oraz inne studia wykorzystujące technikę PSM do oceny oddziaływania programów rynku pracy. Jako podstawowe wskazuje się oczywiście zmienne społeczno-demograficzne takie, jak płeć czy wykształcenie, których związek z zatrudnieniem jest ewidentny. Mniej oczywisty wydaje się ich wpływ na prawdopodobieństwo udziału w programach rynku pracy, choć istnieją studia, z których wynika, że kobiety i osoby lepiej wykształcone są bardziej skłonne podnosić swoje kwalifikacje zawodowe. W literaturze przedmiotu znajdują się również głosy, które szczególnie duże znaczenie przypisują zmiennym odnoszącym się do historii bezrobocia (Heckman i in., 1997: 615) osób przystępujących do programów oraz zarobków poprzedzających udział w programie (Bryson i in., 2002: 13). Zmienne predykcyjne, ostatecznie wykorzystane w szacowanym modelu prawdopodobieństwa, można podzielić następująco:

1. Cechy społeczno-demograficzne:
   - płeć,
   - wiek,
   - stan cywilny,

[54] Pozwoliło to określić pulę jednostek kontrolnych kwalifikujących się do udziału w programie, a co ważniejsze przyjęte rozwiązanie umożliwiło ustalić w konkretnym punkcie czasu charakterystyki tych osób. Miało to duże znaczenie z uwagi na stosunkowo długi czas trwania projektu Alternatywa II.

- samotne wychowywanie dzieci,
- liczba dzieci,
- wykształcenie,
- powiat.

2. Cechy związane z zatrudnieniem, aktywnością zawodową i rodzajem tej aktywności, odnoszące się do sytuacji aktualnej w momencie przystępowania poszczególnych osób do programu:
   - zawód (klasyfikacja zgodna z rozporządzeniem z 8 grudnia 2004 r. w sprawie klasyfikacji zawodów i specjalności na potrzeby rynku pracy oraz zakresu jej stosowania – Dz. U. z 16 grudnia 2004 r.),
   - staż pracy,
   - sumaryczna liczba wykonywanych zawodów,
   - liczba dni na bezrobociu przed udziałem w Phare,
   - liczba dni na zasiłku przed udziałem w programie,
   - liczba propozycji pracy w ciągu roku przed Phare,
   - liczba dni przepracowanych w ramach staży absolwenckich, prac interwencyjnych, robót publicznych,
   - liczba dni bezrobocia ciągłego (w okresie 2 lat przed projektem),
   - praca lub staż +/- 7 dni od programu.
3. Cechy związane z wcześniejszym podnoszeniem kwalifikacji, aktywnością szkoleniową:
   - liczba szkoleń, w których osoba wzięła udział w ciągu roku przed udziałem w Phare,
   - liczba dni spędzonych na szkoleniach,
   - udział w stażu przed rozpoczęciem Phare.
4. Cechy odnoszące się do względnej motywacji osób do poszukiwania pracy:
   - odsetek stawień się na wezwanie z urzędu,
   - posiadanie prawa do zasiłku w momencie przystępowania do programu.
5. Cechy dotyczące posiadanych umiejętności:
   - posiadanie prawa jazdy kat. B[55].

Tabela 7 pokazuje charakterystyki beneficjentów Alternatywy II oraz osób z potencjalnej puli kontrolnej, liczącej 126 633 jednostek. Porównanie średnich zmiennych predykcyjnych obrazuje różnice pomiędzy tymi grupami jeszcze przed procedurą łączenia (por. rozdział II). Ostatnia kolumna tabeli to wystandaryzowana procentowa różnica pomiędzy średnimi danej zmiennej w grupie kontrolnej i grupy osób nieuczestniczących w projekcie. Po doborze próby kontrolnej pomoże ona w ocenie jakości zastosowanego algorytmu łączenia (Konarki, 2007: 195).

Jak widać różnice pomiędzy uczestnikami projektu Alternatywa II a innymi osobami zarejestrowanymi w PUP są znaczące. W puli wszystkich zarejestrowanych bezrobotnych mężczyźni stanowią ponad

[55] W SI PULS było jeszcze kilka innych zmiennych odnoszących się do umiejętności zarejestrowanych w PUP osób bezrobotnych – m.in. posiadane umiejętności zawodowe, zdobyte certyfikaty itp. Natomiast ze względu na specyficzny sposób zbierania danych zawartych w SI PULS – dane dotyczące umiejętności są zapisywane bez odnotowywania momentu ich wprowadzenia – trzeba było zrezygnować z nich, jako że mogły być one faktyczną konsekwencją uczestnictwa w Phare. Mogły być więc potencjalnym źródłem błędów modelu. Zmienna posiadanie prawa jazdy kat. B pozostała w zbiorze, bowiem Phare nie uniemożliwiało doskonalenia swoich możliwości w tym zakresie.

43%, podczas gdy w projekcie ich udział nieznacznie przekroczył jedną czwartą. Beneficjenci projektu, częściej niż pozostałe osoby zarejestrowane w PUP, pozostawali w stanie wolnym. Z kolei ci ostatni, średnio częściej – w momencie realizacji projektu – posiadali już dzieci. W projekcie wzięły udział osoby średnio o ponad rok młodsze od pozostałych bezrobotnych osób (w populacji bezrobotnych do 27 roku życia). Oprócz tego w projekcie wzięły udział osoby zdecydowanie lepiej wykształcone – blisko 50% osób z puli kontrolnej posiadało wykształcenie zasadnicze zawodowe lub niższe. W grupie beneficjentów Alternatywy II odsetek osób z takim wykształceniem nie przekroczył 12%, za to ponad 88% z nich posiadało wykształcenie co najmniej średnie. Grupy różniły się również średnim stażem pracy – w przypadku puli kontrolnej wynosił on około roku, w przypadku beneficjentów był to staż o połowę krótszy. Znamiennymi wydają się być zwłaszcza różnice dotyczące historii bezrobocia w porównywanych grupach. Średni łączny czas przebywania na bezrobociu w okresie poprzedzającym udział w programie, w przypadku beneficjentów wynosił 473 dni. Pozostałe osoby zarejestrowane w PUP mają historię bezrobocia dłuższą o ok. 73% (819 dni). Podobnie ma się sprawa w przypadku bezrobocia ciągłego. W okresie 2 lat poprzedzającym rozpoczęcie projektu jego beneficjenci nieprzerwanie bez pracy pozostawali średnio 233 dni, w przypadku osób z puli kontrolnej jest to średnio 100 dni więcej. Również czas spędzony na zasiłku jest średnio o 50% dłuższy w przypadku osób, które nie uczestniczyły w projekcie. Warto również zwrócić uwagę na różnicę w odsetku osób pozostających bez prawa do zasiłku w momencie przystępowania do projektu – było to odpowiednio 38% z grupy beneficjentów i 26% z grupy pozostałych osób zarejestrowanych w PUP.

Ponadto wykonany test t wykazał statystycznie istotne różnice między średnimi większości zmiennych.

Podsumowując, grupa beneficjentów Alternatywy II znacząco różniła się od osób, które mogły wziąć udział w projekcie. Bardziej ogólne spojrzenie (Dehejia i in., 2002: 11) na problem różnicy pomiędzy grupą beneficjentów i osobami w puli kontrolnej będzie możliwe po oszacowaniu i zestawieniu rozkładu *propensity score*, o czym dalej.

Tabela 7. Stopień zbalansowania zmiennych przed procedurą doboru grupy kontrolnej

Zmienna			Beneficjenci Phare – średnia	Pula kontrolna – średnia	Standaryzowana różnica w %[56]
v3	Płeć (mężczyzna)		0,259	0,433	-37,1
v5	Stan cywilny (wolny/ związany)		0,132	0,242	-28,6
v6	Wykształcenie	brak lub niepełne podstawowe	0,001	0,006	-8,4
		podstawowe	0,024	0,189	-55,5
		gimnazjalne	0,012	0,033	-13,9
		zasadnicze zawodowe	0,075	0,260	-50,9
		średnie	0,694	0,382	66,0
		pomaturalne/policealne	0,072	0,044	11,6
		wyższe (plus licencjat)	0,121	0,086	11,6
v7	Samotne wychowywanie dzieci do 7 lat (nie wychowuje/ wychowuje)		0,032	0,077	-20,1
v8	Liczba dzieci		0,087	0,272	-37,6
v9	Wiek		21,932	23,302	-66,1
v10	Staż pracy w momencie przystępowania do programu Phare		159,267	366,093	-42,9
v17	Zawód	siły zbrojne	0,005	0,007	-2,9
		przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy	0,001	0,001	-1,8
		specjaliści	0,080	0,060	7,7
		technicy i średni personel	0,251	0,145	26,8
		pracownicy biurowi	0,085	0,047	15,1
		pracownicy usług osobistych i sprzedawcy	0,171	0,207	-9,2
		rolnicy, ogrodnicy, leśnicy i rybacy	0,006	0,013	-7,9
		robotnicy przemysłowi i rzemieślnicy	0,065	0,192	-38,7
		operatorzy i monterzy maszyn i urządzeń	0,019	0,026	-4,4
		pracownicy przy pracach prostych	0,054	0,086	-12,5
		brak zawodu	0,248	0,189	14,4

[56] $SD_{przed} = 100(\bar{X}_1 - \bar{X}_{0P}) \Big/ \sqrt{(s_1^2 + s_{0P}^2)/2}$, gdzie $\bar{X}_1$ i $\bar{X}_{0P}$ to średnie analizowanej zmiennej w grupie beneficjentów i w puli kontrolnej, zaś s_1^2 i s_{0P}^2 to oszacowania wariancji dla grupy beneficjentów i w puli kontrolnej

		brak danych	0,017	0,027	-7,1
v18	Liczba wykonywanych zawodów		1,413	1,459	-4,3
v19	Liczba propozycji pracy otrzymanych w ciągu roku przed przystąpieniem do programu		0,300	0,283	2,4
v20	Odsetek stawień na wezwanie	0%	0,080	0,089	-3,6
		0,01%-50%	0,103	0,070	12,0
		50,01% - 99.99%	0,083	0,081	0,6
		100%	0,345	0,318	5,8
		nie dotyczy	0,389	0,442	-10,7
v21	Posiadanie prawa jazdy kat. B		0,152	0,139	3,6
v22	Liczba dni przepracowanych przed udziałem w Phare w ramach staży absolwenckich, prac interwencyjnych, robót publicznych		2,733	1,482	6,3
v25	Liczba szkoleń w ciągu roku przed udziałem w programie		0,032	0,023	5,6
v26	Liczba dni spędzona na szkoleniach, w ciągu roku przed rozpoczęciem programu		1,393	1,175	2,4
v44	Udział w stażu przed programem	brak udziału	0,681	0,760	-17,6
		udział później niż rok i w ostatnim roku przed programem	0,029	0,017	7,8
		udział później niż rok przed programem	0,155	0,100	16,4
		udział w ostatnim roku przed programem	0,135	0,122	3,8
v45	Liczba dni na bezrobociu przed udziałem w programie		473,867	819,623	-59,1
v46	Liczba dni na zasiłku przed udziałem w programie		299,207	447,247	-42,3
v47	Prawo do zasiłku	bez prawa do zasiłku	0,380	0,262	25,4
		prawo do zasiłku przez okres krótszy niż 1 miesiąc	0,062	0,072	-4,3
		prawo do zasiłku przez okres dłuższy niż 1 miesiąc i krótszy niż 3 miesiące	0,121	0,153	-9,3
		prawo do zasiłku przez okres dłuższy niż 3 miesiące	0,437	0,512	-15,0
v48	Praca lub staż +/- 7 dni od programu		0,051	0,009	24,5
v49	Liczba dni bezrobocia ciągłego (w okresie 2 lat przed projektem)		233,624	330,349	-39,6

Źródło: opracowanie własne.

Szacowanie *propensity scores*

Do oszacowania wartości *propensity score* posłużyła regresja logistyczna, której analiza została przeprowadzona z użyciem pakietu statystycznego SPSS. Analizy były prowadzone w każdym województwie osobno, w związku z czym finalnie utworzono 13 modeli prawdopodobieństwa (projekt nie był realizowany w 3 województwach). Zastosowane podejście podyktowane było chęcią uchwycenia wpływu regionu realizacji projektu (por. rozdział 2).

Zmienną zależną w modelu była dychotomiczna zmienna uczestnictwo w programie (wartość jeden odpowiadała beneficjentowi Alternatywy II, wartość zero odpowiadała osobie z puli kontrolnej). Zmiennymi niezależnymi był zbiór 21 zmiennych wskazanych powyżej. Do zbioru dołączono również zmienną powiat, tak aby ująć w modelu ewentualny wpływ miejsca, w którym realizowany był projekt. Oprócz tego zmienne niezależne o charakterze jakościowym na etapie analizy zostały rozbite na zmienne zero-jedynkowe (tzw. *dummy variables*). Dotyczyło to zmiennych wykształcenie, powiat, kod zawodu, udział w stażu przed programem, prawo do zasiłku. W praktyce więc w modelu znajdowało się nawet do kilkunastu zmiennych więcej. Ostateczna liczba zmiennych włączanych do modelu nie była jednak stała dla wszystkich województw, w których wykonano analizy. W celu osiągnięcia konwergencji algorytmu estymacji w niektórych województwach lista zmiennych analitycznych wprowadzonych do modelu była nieznacznie ograniczona. Z modelu usuwane były zmienne, które generowały problem w szacowaniu *propensity score*. Poniżej znajdują się przykładowe wyniki z przeprowadzonej analizy regresji logistycznej w województwie pomorskim.

Tabela 8 – Przebieg iteracji – zawiera wyniki estymacji modelu bez zmiennych objaśniających. Można zauważyć, jak zmieniało się oszacowanie stałej oraz funkcja wiarygodności modelu w kolejnych iteracjach (jej wartość pomnożona razy minus dwa). Obserwowana wartość funkcji wiarygodności modelu (4471,727) posłuży następnie jako punkt odniesienia dla ustalenia miar jakości modelu ze zmiennymi objaśniającymi.

Tabela 8. Przebieg iteracji (a, b, c)

Iteracja		-2 logarytm wiarygodności	Współczynniki
		Stała	Stała
Krok 0	1	5976,346	-1,847
	2	4649,901	-2,679
	3	4478,404	-3,110
	4	4471,743	-3,217
	5	4471,727	-3,223
	6	4471,727	-3,223

a – Stała została włączona do modelu.
b – Początkowa wartość -2 logarytm wiarygodności: 4471,727.
c – Estymacja została zakończona na iteracji o numerze 6, ponieważ oszacowania parametrów zmieniły się o mniej niż 0,001.

Tabela 9 przedstawia wyniki testów, w których sprawdza się, czy model zawierający zmienne predykcyjne jest istotnie różny od modelu tylko z wyrazem wolnym. Jak widać dodanie zmiennych poprawia możliwość przewidywania zmiennej zależnej[56].

Tabela 9. Test zbiorowy współczynników modelu

		Chi-kwadrat	df	Istotność
Krok 1	Krok	836,951	44	,000
	Blok	836,951	44	,000
	Model	836,951	44	,000

W tabeli „Podsumowanie dla modelu" widoczne są miary jakości dopasowania modelu, w tym logarytm wiarygodności (pomnożony razy minus dwa). Jego mniejsza wartość niż w przypadku modelu tylko z wyrazem wolnym wskazuje na polepszenie wynikające z wprowadzenia do modelu zmiennych niezależnych. Niepokój mogą budzić niskie wartości miar pseudo-R-kwadrat, które wskazują, że w modelu znalazły się stosunkowo słabe predyktory. Istnieje również ryzyko nieuwzględnienia innych istotnych predyktorów[57].

Tabela 10. Podsumowanie dla modelu

Krok	-2 logarytm wiarygodności	R kwadrat Coxa i Snella	R kwadrat Nagelkerke'a
1	3634,776(a)	,059	,213

a Estymacja została zakończona na iteracji o numerze 9, ponieważ oszacowania parametrów zmieniły się o mniej niż ,001

W tabeli 11 przedstawione są wyniki testu Hosmera-Lemeshowa, w którym weryfikowana jest hipoteza zerowa o tym, że model jest dobrze dopasowany do danych. Widoczny brak istotności wskazuje na to, że model jest dobrze dopasowany do danych.

Tabela 11. Test Hosmera i Lemeshowa

Krok	Chi-kwadrat	Df	Istotność
1	7,177	8	,518

W ostatniej tabeli znajdują się oszacowania szukanych parametrów. Wartości B to szacowane parametry modelu. Wartości Sig. wskazują na istotność poszczególnych zmiennych dla modelu. Exp(B) mierzy zmianę prawdopodobieństwa zajścia zdarzenia, jakim jest udział w projekcie Alternatywa II pod wpływem wzrostu wartości danego predyktora o 1. Wartości większe od 1 odpowiadają wpływowi

[56] Wszystkie trzy porównania są tożsame, (przy jednym bloku i metodzie Enter, jaką zastosowano, jest tylko jeden krok) i odnoszą się do modelu tylko ze stałą (Blok = 0).

[57] Problem ten ma swoje źródło w danych, które wykorzystane zostały do stworzonego modelu uczestnictwa. Celem funkcjonowania Systemu Informatycznego PULS nie jest gromadzenie danych na potrzeby badań ewaluacyjnych. Niejako przy okazji udało się go wykorzystać w badaniach realizowanych przez PARP. Z punktu widzenia trafności wniosków płynących z tego typu analiz idealna byłaby sytuacja, w której niezbędne dane gromadzone byłyby w ramach sprawnie funkcjonującego systemu monitoringu. Dodatkowo zakres danych powinien podlegać głębokiej badawczej refleksji, tak aby w możliwie najpełniejszym stopniu uchwycić czynniki odpowiadające za mechanizmy selekcji.

dodatniemu, mniejsze od 1 ujemnemu. Jak widać tylko kilka zmiennych jest istotnych statystycznie. Tak jak to zostało zaprezentowane w rozdziale drugim, podczas szacowania modelu udziału nie chodzi jednak o to, by model zawierał zmienne wyłącznie istotnie statystycznie. Przyjmuje się, że każda ze zmiennych niesie ze sobą pewną informację, którą należy wykorzystać do utworzenia grupy kontrolnej, w jak największym stopniu podobnej do grupy interwencji. Należy pamiętać, że głównym celem estymacji *propensity score* nie jest jak najlepsza predykcja selekcji obserwacji do działania, ale zbalansowanie wszystkich zmiennych włączonych do modelu (Caliendo i in., 2005: 10).

Tabela 12. Zmienne w modelu

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	v3	-,639	,126	25,560	1	,000	,528
	v5	-,391	,162	5,819	1	,016	,676
	v6			136,965	5	,000	
	v6(1)	-3,002	,397	57,207	1	,000	,050
	v6(2)	-4,857	1,033	22,109	1	,000	,008
	v6(3)	-2,153	,289	55,409	1	,000	,116
	v6(4)	-,513	,216	5,650	1	,017	,599
	v6(5)	-,469	,269	3,057	1	,080	,625
	v7	-,633	,293	4,656	1	,031	,531
	v8	,049	,161	,093	1	,760	1,050
	v9	-,244	,040	37,747	1	,000	,784
	v10	,000	,000	,917	1	,338	1,000
	v13			49,784	5	,000	
	v13(1)	,884	,156	31,958	1	,000	2,420
	v13(2)	,516	,182	8,084	1	,004	1,676
	v13(3)	,964	,195	24,505	1	,000	2,622
	v13(4)	,305	,162	3,529	1	,060	1,356
	v13(5)	,317	,264	1,444	1	,230	1,373
	v17_1	,037	1,100	,001	1	,973	1,038
	v17_2	-,909	,430	4,466	1	,035	,403
	v17_3	-,224	,335	,445	1	,505	,800
	v17_4	-,312	,381	,669	1	,413	,732
	v17_5	-,136	,343	,158	1	,691	,873
	v17_6	,663	,494	1,800	1	,180	1,941
	v17_7	-,094	,360	,069	1	,793	,910
	v17_8	,479	,426	1,265	1	,261	1,614
	v17_9	-,024	,372	,004	1	,948	,976
	v17_10	-,221	,329	,450	1	,502	,802
	v18	,020	,065	,098	1	,754	1,021
	v19	,021	,116	,031	1	,860	1,021

	v20_0	,743	,627	1,401	1	,237	2,101
	v20_1	-1,648	1,034	2,541	1	,111	,192
	v20_2	-,472	,501	,885	1	,347	,624
	v20_3	,296	,144	4,257	1	,039	1,345
	v21	-,085	,149	,325	1	,569	,919
	v22	,003	,002	1,152	1	,283	1,003
	v25	,774	,391	3,922	1	,048	2,168
	v26	-,011	,009	1,400	1	,237	,989
	v44			32,706	3	,000	
	v44(1)	1,161	,204	32,526	1	,000	3,194
	v44(2)	,638	,414	2,379	1	,123	1,893
	v44(3)	,993	,225	19,533	1	,000	2,699
	v45	-,001	,000	9,845	1	,002	,999
	v46	,000	,000	1,540	1	,215	1,000
	v47			20,967	3	,000	
	v47(1)	,454	,121	14,053	1	,000	1,574
	v47(2)	-,088	,213	,170	1	,680	,916
	v47(3)	-,197	,163	1,452	1	,228	,821
	Stała	2,300	,976	5,559	1	,018	9,975

a – Zmienne wprowadzone w kroku 1: v3, v5, v6, v7, v8, v9, v10, v13, v17_1, v17_2, v17_3, v17_4, v17_5, v17_6, v17_7, v17_8, v17_9, v17_10, v18, v19, v20_cat_0, v20_cat_1, v20_cat_2, v20_cat_3, v21, v22, v25, v26, v44, v45, v46, v47.

Efektem przeprowadzonej analizy regresji logistycznej było uzyskanie końcowego oszacowania wartości *propensity score* dla wszystkich osób w grupie uczestników oraz w puli kontrolnej. Ostateczne wartości *propensity score* wyliczone zostały[58], jako konsekwencja maksymalizacji funkcji wiarygodności, zgodnie z równaniem przytoczonym w rozdziale 2:

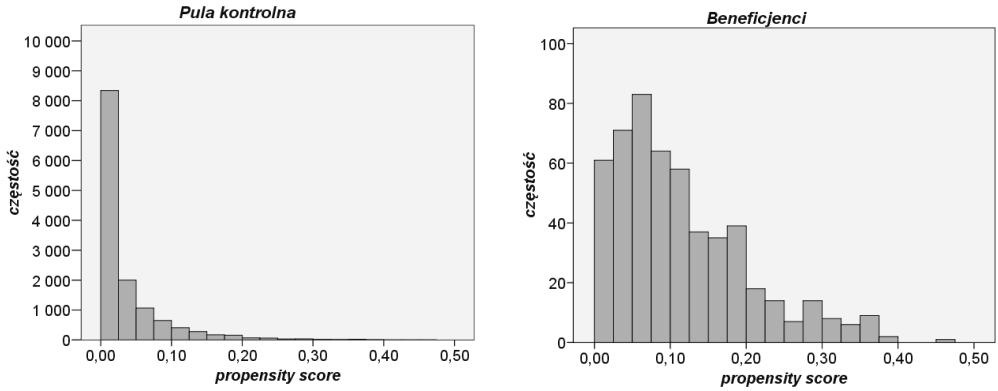
$$P(D=1|x_1,x_2,,...,x_k)=\frac{e^{\beta_0+\sum_{i=1}^{k}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{k}\beta_i x_i}}$$

Modelowanie analogiczne do powyższego miało miejsce dla każdego z województw.

[58] W przypadku pakietu statystycznego SPSS, chcąc aby oszacowane prawdopodobieństwa zostały dołączone do zbioru danych, należy w menu regresji logistycznej wskazać w opcji zapisz –> wartości przewidywane –> *prawdopodobieństwa*.

Dobór grupy kontrolnej

Po oszacowaniu wartości *propensity score* kolejnym punktem w zastosowaniu techniki PSM jest dobór próby kontrolnej z wykorzystaniem wybranej metody łączenia. Zgodnie z tym, co zostało napisane w rozdziałem drugim, na tym etapie należy podjąć trzy decyzje: czy dokonywać łączenia ze zwracaniem, ile jednostek kontrolnych będzie przypadać na jednego beneficjenta i finalnie, jaką metodę zastosować. Podjęcie decyzji warto poprzedzić analizą rozkładu wartości *propensity score* w grupie uczestników Alternatywy i osób w puli kontrolnej. Poniższy wykres obrazuje rozkład oszacowanych wartości *propensity score* dla osób w województwie pomorskim (527 beneficjentów i 13 332 osób w puli kontrolnej).



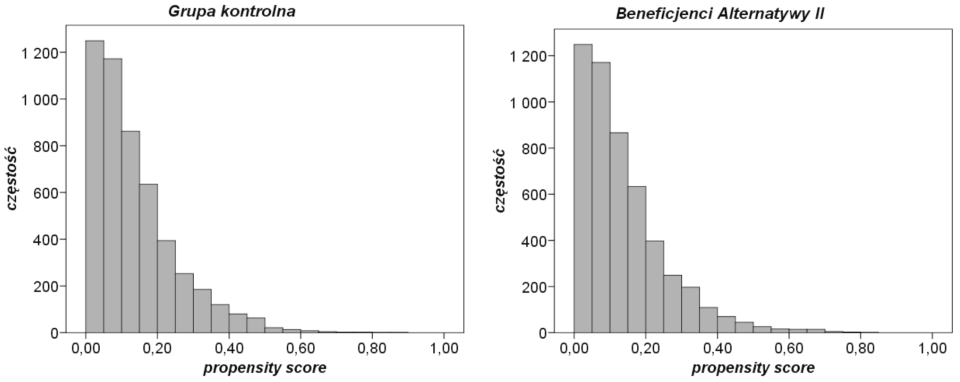
Rys. 3. Rozkład oszacowanych *propensity score* w puli kontrolnej oraz w grupie beneficjentów projektu w województwie pomorskim

Źródło: opracowanie własne.

Jak można zauważyć, beneficjenci stanowili tylko niewielką część w każdym z przedziałów prawdopodobieństwa. Istniała duża nadwyżka jednostek z puli kontrolnej – szczególnie dla niższych wartości *propensity score* (do 0,15) – przewyższająca zwykle kilkunastokrotnie liczebność grupy uczestników Alternatywy II. Rozkład prawdopodobieństwa w puli kontrolnej wskazywał więc, że efektywną metodą doboru grupy kontrolnej będzie dobór jednostek bez zwracania oraz w układzie 1 do 1. Z kolei liczebność w poszczególnych przedziałach prawdopodobieństwa wskazywała na to, że skuteczną może okazać się technika najbliższego sąsiada, nawet bez konieczności określania górnej wartości różnicy prawdopodobieństwa (*caliper*). Każdemu beneficjentowi projektu przyporządkowano jedną osobę spośród osób znajdujących się w puli kontrolnej, której wynikające z modelu prawdopodobieństwo wzięcia udziału w projekcie było najbliższe prawdopodobieństwu danego beneficjenta. Ponieważ liczebność grupy osób nieuczestniczących w projekcie była w przypadku każdego województwa wielokrotnie wyższa niż liczba uczestników, w pewnych przypadkach zachodziła konieczność wyboru osoby do grupy kontrolnej w sytuacji dopasowania 1 do n (gdy np. w grupie kontrolnej występowało kilka jednostek o takim samym *propensity score*). W takim przypadku wybierano jedną osobę w sposób losowy.

Skuteczność zastosowanej metody selekcji grupy kontrolnej pokazują dwa poniższe wykresy na których znajduje się rozkład *propensity score* – odpowiednio w grupie beneficjentów oraz w grupie

kontrolnej. Jak widać rozkłady te są bardzo do siebie zbliżone, co świadczy o dobrym dopasowaniu grupy kontrolnej.



Rys. 4. Rozkład oszacowanych *propensity score* w grupie kontrolnej oraz w grupie beneficjentów projektu
Źródło: opracowanie własne.

Wybrane statystyki zaprezentowane w tabeli 13 potwierdzają powyższy wniosek, wskazując na duże podobieństwo rozkładów prawdopodobieństwa w grupie beneficjentów i w dobranej grupie kontrolnej.

Tabela 13. Wybrane statystyki dla zmiennej *propensity score* w grupie beneficjentów projektu Alternatywa II i w grupie kontrolnej

	Beneficjenci	Grupa kontrolna
N	5065	5065
Średnia	0,13729	0,13660
Odchylenie std.	0,11828	0,11551
Minimum	0,00034	0,00034
Maksimum	0,82117	0,85111
Pierwszy kwartyl	0,05082	0,05082
Mediana	0,10405	0,10407
Trzeci kwartyl	0,18999	0,19003
Suma	695,3926	691,8659

Źródło: opracowanie własne.

Ocena stopnia zbalansowania zmiennych wykorzystanych w modelu

Przedstawione powyżej wnioski należy uzupełnić o ocenę dopasowania grupy kontrolnej z punktu widzenia stopnia zbalansowania zmiennych predykcyjnych. Ocena ta dokonana została z wykorzystaniem standaryzowanej różnicy średnich oraz za pomocą testu t. Jak widać, wszystkie znaczące różnice pomiędzy grupą beneficjentów i osób nieuczestniczących w Alternatywie, a dobranych do grupy kontrolnej, zostały zminimalizowane (por. tabela 7). Wartość standaryzowanej różnicy średnich rzadko

kiedy przekracza 2%, podczas gdy w zestawieniu beneficjentów z całą pulą kontrolną wielkości te najczęściej przewyższały 10%. Na niezadowalającym poziomie pozostaje jedynie minimalizacja obciążenia na zmiennych *Praca lub staż +/– 7 dni od programu* oraz *Liczba dni bezrobocia ciągłego*. Wykonany test t nie wykazał statystycznie istotnych różnic w średnich większości zmiennych.

W związku z powyższym, poziom zbalansowania zmiennych, które zostały wykorzystane w modelu, należy uznać za satysfakcjonujący.

Tabela 14. Stopień zbalansowania zmiennych po procedurze doboru grupy kontrolnej

Zmienna			Beneficjenci Phare – średnia	Grupa kontrolna – średnia	Standaryzowana różnica w %[60]
v3	Płeć (mężczyzna)		0,259	0,258	0,3
v5	Stan cywilny (wolny/związany)		0,132	0,138	-1,6
v6	Wykształcenie	brak lub niepełne podstawowe	0,001	0,001	0,0
		podstawowe	0,024	0,025	-0,1
		gimnazjalne	0,012	0,010	1,3
		zasadnicze zawodowe	0,075	0,073	0,7
		średnie	0,694	0,693	0,3
		pomaturalne/policealne	0,072	0,076	-2,0
		wyższe (plus licencjat)	0,121	0,122	-0,2
v7	Samotne wychowywanie dzieci do 7 lat (nie wychowuje/ wychowuje)		0,032	0,031	0,4
v8	Liczba dzieci		0,087	0,089	-0,2
v9	Wiek		21,932	21,952	-1,0
v10	Staż pracy w momencie przystępowania do programu Phare		159,267	150,804	1,8
v17	Zawód	siły zbrojne	0,005	0,005	0,0
		przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy	0,001	0,001	-1,2
		specjaliści	0,080	0,079	0,3
		technicy i średni personel	0,251	0,258	-2,0
		pracownicy biurowi	0,085	0,085	-0,2
		pracownicy usług osobistych i sprzedawcy	0,171	0,175	-1,2
		rolnicy, ogrodnicy, leśnicy i rybacy	0,006	0,004	1,8
		robotnicy przemysłowi i rzemieślnicy	0,065	0,061	1,3
		operatorzy i monterzy maszyn i urządzeń	0,019	0,018	1,1
		pracownicy przy pracach prostych	0,054	0,057	-1,1
		brak zawodu	0,248	0,242	1,4
		brak danych	0,017	0,015	1,6

[60] $SD_{przed} = 100(\bar{X}_1 - \bar{X}_{0K}) \Big/ \sqrt{(s_1^2 + s_{0P}^2)/2}$, gdzie $\bar{X}_1$ i $\bar{X}_{0K}$ to średnie analizowanej zmiennej w grupie beneficjentów i w puli kontrolnej, zaś s_1^2 i s_{0P}^2 to oszacowania wariancji dla grupy beneficjentów i w puli kontrolnej.

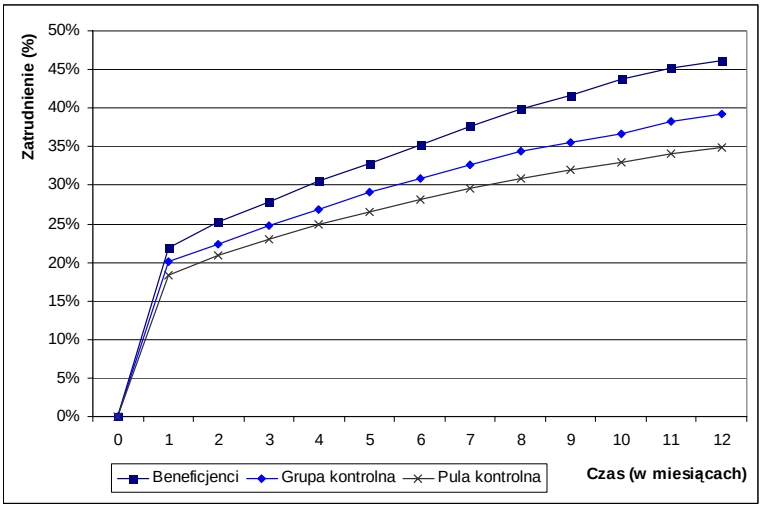
v18	Liczba wykonywanych zawodów		1,413	1,414	0,0
v19	Liczba propozycji pracy ctrzymanych w ciągu roku przed przys:ąpieniem do programu		0,300	0,309	-1,4
v20	Odsetek stawień na wezwanie	0%	0,080	0,086	-2,4
		0,01%–50%	0,103	0,104	-0,2
		50,01%–99,99%	0,083	0,093	-3,8
		100%	0,345	0,343	0,5
		nie dotyczy	0,389	0,374	3,1
v21	Posiadanie prawa jazdy kat. B		0,152	0,152	0,1
v22	Liczba dni przepracowanych przed udziałem w Phare w ramach staży absolwenckich, prac interwencyjnych, robót publicznych		2,733	2,608	0,6
v25	Liczba szkoleń w ciągu rcku przed udziałem w programie		0,032	0,031	0,6
v26	Liczba dni spędzona na szkoleniach, w ciągu roku przed rozpoczęciem programu		1,393	1,291	1,1
v44	Udział w stażu przed programem	brak udziału	0,681	0,669	2,7
		udział później niż rok i w ostatnim roku przed programem	0,029	0,029	0,1
		udział później niż rok przed programem	0,155	0,163	-2,4
		udział w ostatnim roku przed programem	0,135	0,139	-1,2
v45	Liczba dni na bezrobociu przed udziałem w programie		473,867	483,460	-1,6
v46	Liczba dni na zasiłku przed udziałem w programie		299,207	304,967	-1,6
v47	Prawo do zasiłku	bez prawa do zasiłku	0,380	0,378	0,5
		prawo do zasiłku przez okres krótszy niż 1 miesiąc	0,062	0,062	0,0
		prawo do zasiłku przez okres dłuższy niż 1 miesiąc i krótszy niż 3 miesiące	0,121	0,123	-0,5
		prawo do zasiłku przez okres dłuższy niż 3 miesiące	0,437	0,438	-0,1
v48	Praca lub staz +/- 7 dni od programu		0,051	0,013	22,6
v49	Liczba dni bezrobocia ciągłego (w okresie 2 lat przed projektem)		233,624	267,151	-13,7

Źródło: opracowanie własne.

Wyniki analiz – efekt netto projektu Alternatywa II

Mając wybrane jednostki kontrolne, można dokonać zestawienia zmiennej wynikowej (zatrudnienie) w grupie beneficjentów oraz w grupie kontrolnej. Takie zestawienie znajduje się na zaprezentowanym poniżej rysunku 5. Pokazuje on przyrost odsetka zatrudnionych, poczynając od pierwszego miesiąca po zakończeniu uczestnictwa poszczególnych osób w projekcie. Dodatkowo pokazano jak kształtowała się wartość zmiennej wynikowej w całej populacji, a więc w sytuacji, gdyby nie tworzyć grupy kontrolnej z wykorzystaniem techniki PSM. Prezentowane dane przedstawiają wynik dla całego projektu Alternatywa II, a więc po połączeniu danych z wszystkich województw.

Z wykresu wynika, że można mówić o pozytywnym wpływie projektu Alternatywa II na zatrudnienie jego beneficjentów. Już od pierwszego miesiąca występuje istotna statystycznie różnica w osiągniętym wyniku pomiędzy grupą beneficjentów a grupą kontrolną. W miesiąc od zakończenia udziału w projekcie zatrudnionych było 21,9% beneficjentów oraz 20% osób z grupy kontrolnej. Kolejne miesiące nieznacznie powiększają różnicę pomiędzy obiema grupami, tak że finalnie po dwunastu miesiącach od wyjścia z projektu zatrudnionych było 46,2% uczestników i 39,2% osób z grupy kontrolnej. Efekt netto Projektu Alternatywa II – liczony dla tego momentu w czasie – wynosi więc ok. 7%. Warto w tym miejscu zwrócić uwagę na różnice w wartościach zmiennej wynikowej u osób z grupy kontrolnej i puli kontrolnej. Wyniki tych ostatnich wskazują na wielkość przeciętnego zatrudnienia wśród wszystkich młodych bezrobotnych w powiatach uczestniczących w projekcie. Jest to więc wskaźnik zatrudnienia wśród osób, które zgodnie z formalnymi kryteriami selekcji, uprawnione były do skorzystania ze wsparcia oferowanego w ramach Phare. Gdyby założyć, że selekcja uczestników do projektu opierała się wyłącznie na tych kryteriach, dokonane oszacowanie efektu netto obciążone byłoby błędem. Wartość tego błędu to różnica pomiędzy odsetkiem zatrudnionych z grupy kontrolnej a odsetkiem osób zatrudnionych z puli kontrolnej. Różnica ta odzwierciedla korzyść z zastosowania techniki PSM do oszacowania efektu projektu. Jak widać w dwunastym miesiącu jest to wartość około 4%.



Rys. 5. Odsetki zatrudnionych w grupie beneficjentów, w grupie kontrolnej oraz w puli kontrolnej, w kolejnych miesiącach od zakończenia udziału w projekcie

Źródło: opracowanie własne na podstawie raportu z badań: PAG Uniconsult, ARC Rynek i Opinia (2007). Ewaluacja ex-post Phare 2003 Spójność Społeczna i Gospodarcza – komponent Rozwój Zasobów Ludzkich.

Szczegółowe dane liczbowe na temat zatrudnienia w grupie beneficjentów i w grupie kontrolnej zawiera poniższa tabela.

Tabela 15. Odsetek zatrudnionych w grupie beneficjentów, w grupie kontrolnej oraz w puli kontrolnej, w kolejnych miesiącach od zakończenia udziału w projekcie oraz oszacowany efekt netto projektu Alternatywa II

Liczba miesięcy dzieląca osoby od momentu wyjścia z projektu	A. Benefi-cjenci	B. Grupa kontrolna	C. Pula kontrolna	Różnica: A–B Efekt netto	Istotność różnicy A–B*
0	0,00%	0,00%	0,00%	0,00%	-
1	21,86%	20,04%	18,26%	1,82%	+
2	25,19%	22,29%	20,83%	2,90%	+
3	27,84%	24,80%	22,98%	3,04%	+
4	30,48%	26,87%	24,94%	3,61%	+
5	32,73%	29,10%	26,56%	3,63%	+
6	35,28%	30,86%	28,09%	4,42%	+
7	37,67%	32,60%	29,59%	5,07%	+
8	39,80%	34,43%	30,84%	5,37%	+
9	41,64%	35,60%	32,00%	6,04%	+
10	43,65%	36,58%	32,97%	7,07%	+
11	45,17%	38,28%	34,04%	6,89%	+
12	46,22%	39,23%	34,93%	6,99%	+

*Zastosowany test t dla prób niezależnych, przy poziomie ufności 0,95. *„+" zależność istotna; „-"zależność nieistotna.*
Źródło: opracowanie własne na podstawie raportu z badań: PAG Uniconsult, ARC Rynek i Opinia (2007). Ewaluacja ex-post Phare 2003 Spójność Społeczna i Gospodarcza – komponent Rozwój Zasobów Ludzkich.

Wnioski z analiz i ich skutki dla oceny projektu Alternatywa II

Analiza pod kątem skuteczności netto pokazała, że projekt Alternatywa II miał dodatni wpływ na zatrudnienie wśród grupy jego uczestników. Z punktu widzenia oceny projektu należy jednak podjąć kwestię wielkości tego wpływu w kontekście poniesionych w projekcie wydatków. Na początek, warto zwrócić uwagę na fakt, że cel projektu – osiągnięcie 40% poziomu zatrudnienia w grupie jego beneficjentów – został osiągnięty. Patrząc więc z perspektywy postawionych w nim celów ilościowych można powiedzieć, że projekt był skuteczny. Do takich też wniosków mogłaby doprowadzić jego pobieżna ocena. Idąc dalej wiadomo jednak, że koszt projektu to ponad 4 miliony euro. Patrząc na globalne zatrudnienie beneficjentów (brutto), można oszacować, że rok po zakończeniu udziału w projekcie pracowało ok. 2600 z nich (47%). Dokonując prostego przeliczenia poniesionych nakładów i uzyskanych efektów otrzymamy, że średni koszt uzyskania zatrudnienia przez jednego beneficjenta brutto wyniósł około 1572 euro. Szacunki te zmienią się jednak diametralnie, gdy odwołamy się do przeprowadzonych analiz z wykorzystaniem techniki PSM. Bezpośrednio w wyniku projektu pracę znalazło ok. 396 osób (7%). Odnosząc tę wartość do kosztu projektu uzyskamy jednostkowy koszt zatrudniania beneficjenta netto, który w tym przypadku wynosi ponad 10 tysięcy euro, a więc jest ok. sześć i pół razy więcej niż wynika to z analizy skuteczności brutto projektu. Czy to dużo, czy mało pozostaje kwestią, którą powinni rozstrzygnąć oceniający całe przedsięwzięcie, a zapewne także instytucje finansujące projekt. Z pewnością warto byłoby odwołać się do innych analiz wykonywanych w obszarze programów rynku pracy, w których podjęto się pomiaru efektu netto działań[59]. Wskazana rozbieżność jest jednak dość istotna i trudno pominąć podczas dokonywania oceny projektu. Naturalne wydają się tu pytania o to, czy interwencja w postaci projektu Alternatywa II była w ogóle zasadna. Być może wydatkowane środki, można było spożytkować w inny, bardziej efektywny sposób. Warto zwrócić uwagę na fakt, że wskazany cel projektu (40% zatrudnienie w grupie uczestników), byłby w zasadzie osiągnięty niezależnie od realizacji projektu, o czym świadczy poziom zatrudnienia osób z grupy kontrolnej (39,2%).

Tabela 16. Podsumowanie efektywności projektu Alternatywa II

Wydatkowanie w projekcie	€ 4 090 702,03
Odsetek zatrudnionych brutto	46%
Szacowana liczba beneficjentów, którzy znaleźli zatrudnienie brutto	2602
Średni koszt uzyskania jednostkowego zatrudnienia brutto po programie	€ 1 572
Oszacowany efekt netto zatrudnienia	7%
Szacowana liczba beneficjentów, którzy znaleźli zatrudnienie w związku z udziałem w programie – na podstawie analizy efektu netto	396
Średni koszt uzyskania zatrudnienia po programie w związku z udziałem programie – na podstawie analizy efektu netto	€ 10 330

Źródło: opracowanie własne na podstawie raportu z badań: PAG Uniconsult, ARC Rynek i Opinia (2007). Ewaluacja ex-post Phare 2003 Spójność Społeczna i Gospodarcza – komponent Rozwój Zasobów Ludzkich.

[59] Problemem jest oczywiście znalezienie badań, w których zastosowano metodę PSM i które tym samym mogłyby posłużyć do rekomendowanych porównań. Wydaje się jednak, że liczba takich opracowań może w niedługim czasie wzrosnąć. Chlubnym przykładem jest tu raport opracowany przez Ministerstwo Pracy i Polityki Społecznej w 2008 r. – *Zatrudnienie w Polsce 2007. Bezpieczeństwo na elastycznym rynku pracy*. Przedstawia on m.in. wyniki analiz efektu netto realizowanych w Polsce Aktywnych Polityk Rynku Pracy (ALMP). Przy szacowaniu efektu netto ALMP wykorzystano technikę PSM.

Podczas całościowej oceny projektu, na pewno należy wziąć również pod uwagę fakt, że powyższa analiza jest jednowymiarowa. Oszacowany został efekt netto projektu, który wskazuje, w jakim stopniu Projekt Alternatywa II wpłynął na obserwowany w grupie jego beneficjentów poziom zatrudnienia. Pewien niedosyt może tu wywoływać brak informacji na temat jakości pracy, którą posiadają beneficjenci projektu. Brak jest oczywiście takiej informacji również dla dobranej grupy kontrolnej. Przy ocenie jakości pracy można by wziąć pod uwagę takie kwestie jak np.: wysokość zarobków[60], forma zatrudnienia (rodzaj umowy), odległość od miejsca zamieszkania czy choćby satysfakcję z posiadanej pracy, a pewnie i wiele innych. Tego typu dane mogłyby rzucić nowe światło na oszacowany efekt netto projektu.

(Kontr)przykład wykorzystania techniki PSM

W tej części przedstawione zostały wyniki innej ewaluacji, w której również do utworzenia grupy kontrolnej wykorzystano technikę PSM. Badanie objęło wcześniejszą, wobec opisanej powyżej, edycję Phare – 2002 SSG RZL[61]. Prezentowana w tym miejscu ewaluacja stanowi przykład niepowodzenia zastosowania techniki PSM.

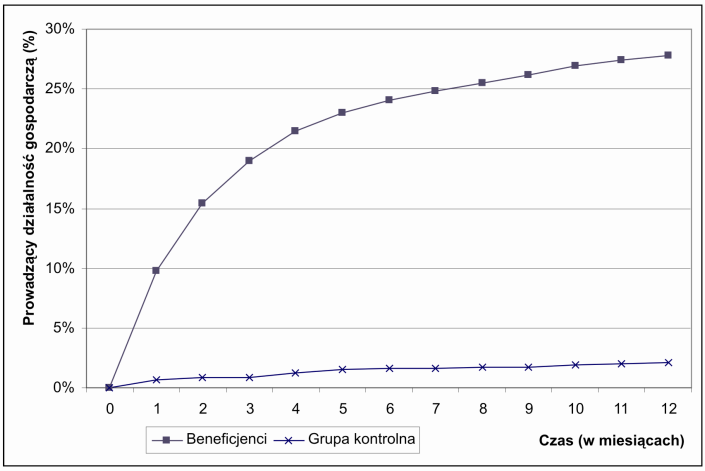
Co istotne ewaluacja Phare 2002 bazowała na tym samym źródle danych, co analizy wykonane dla Projektu Alternatywa II (SI PULS). Ogólna, koncepcja badania była zbliżona do ewaluacji Projektu Alternatywa II. Zakres danych włączonych do modelu był niemalże identyczny[62]. *Propensity score* zostało oszacowane z wykorzystaniem regresji logistycznej. Grupę kontrolną dobrano z wykorzystaniem metody najbliższego sąsiada. Wszystkie zmienne predykcyjne udało się zbalansować w grupie beneficjentów i grupie kontrolnej w zadowalającym stopniu. To, co różnicowało obie ewaluacje, to przedmiot badania. W ramach Phare 2002, tzw. Podprojekt II, ukierunkowany był na wsparcie osób bezrobotnych, chcących rozpocząć własną działalność gospodarczą. Zakres wsparcia dla uczestników projektu obejmował szkolenia z zakresu przedsiębiorczości na różnym poziomie zaawansowania oraz różne formy doradztwa indywidualnego. Wyniki z przeprowadzonych analiz prezentuje poniższy rysunek.

---

60 Abstrahując od oceny projektu, w tym miejscu należy także zauważyć, że pewną ułomnością modelu mógł być brak informacji o wysokości wynagrodzenia przed przystąpieniem do projektu. Autorzy, tacy jak James Heckman, wskazują, że jest to istotny predyktor uczestnictwa osób bezrobotnych w programach rynku pracy. De facto nie wiadomo, na ile założenie warunkowej niezależności (CIA) zostało spełnione, czy nie pominięto jakichś istotnych zmiennych. Stąd można powiedzieć, że oszacowany efekt netto interwencji jest raczej przybliżeniem rzeczywistego efektu przyczynowego.

61 Ewaluacja zrealizowana została na zlecenie PARP w 2006 r.

62 W ewaluacji Phare 2002 SSG RZL wykorzystano w modelu prawdopodobieństwa jedną zmienną więcej niż w modelu utworzonym w ramach ewaluacji projektu Alternatywa II – niepełnosprawność osoby.



Rys. 6. Odsetek osób prowadzących własną działalność gospodarczą w grupie beneficjentów oraz w grupie kontrolnej, w kolejnych miesiącach od zakończenia udziału w programie

Źródło: PBS DGA (2006) Raport z ewaluacji ex-post komponentu regionalnego programu Phare 2002 Spójność Społeczna i Gospodarcza – komponent Rozwój Zasobów Ludzkich.

Jak widać istnieje drastyczna różnica pomiędzy wynikami jednostek w grupie beneficjentów i w grupie kontrolnej. Niemal 30% uczestników projektu, w 12 miesięcy od jego zakończenia, prowadziło własną działalność gospodarczą. Równocześnie na podobny krok zdecydował się tylko niewielki odsetek osób z grupy kontrolnej (ok. 2%). Do wniosków na temat wielkości efektu netto podprojektu należy podchodzić jednak ostrożnie. Obserwowane wyniki mogą z jednej strony świadczyć o tym, że rozpoczęta przez beneficjentów działalność gospodarcza była wyłącznym efektem oddziaływania Phare 2002. Jak piszą jednak autorzy raportu ewaluacyjnego, wniosek taki byłby sprzeczny z wynikami równolegle prowadzonego badania ankietowego, w którym ponad połowa beneficjentów prowadzących działalność gospodarczą przyznała, że założyłaby przedsiębiorstwa bez względu na udział w programie (PARP, 2006: 112). Uwzględnić tu oczywiście należy wszelkie ograniczenia badania wpływu interwencji metodą wywiadu kwestionariuszowego – w tym deklaratywność odpowiedzi. Niemniej jednak wydaje się raczej mało prawdopodobne, żeby opisywany podprojekt faktycznie aż w takim stopniu przyczynił się do aktywizacji zawodowej beneficjentów w zakresie zakładania własnej działalności gospodarczej. Oferowane w nim wsparcie sprowadzało się jedynie do szkoleń w zakresie przedsiębiorczości i pomocy doradczej. Brakowało w nim natomiast takiego instrumentu wsparcia, jak bezpośrednie wsparcie finansowe.

Najbardziej prawdopodobnym wytłumaczeniem dla zaistniałej rozbieżności wyników grupy beneficjentów i grupy kontrolnej jest brak kluczowej informacji w Systemie Informatycznym PULS. Do podprojektu przyjmowane były osoby, które zadeklarowały chęć założenia działalności gospodarczej. Phare miało stanowić w tym celu jedynie element wspomagający. Tak więc decyzja o tym, czy dana osoba będzie próbować założyć własną firmę, podejmowana była zapewne najczęściej jeszcze przed przystąpieniem do projektu. W SI PULS brakuje danych, które wskazywałyby na plany osób bezrobot-

nych związane z zakładaniem własnej działalności gospodarczej. Jak widać, pominięcie tej jednej informacji całkowicie zaburzyło szacunki efektu netto projektu.

Pomimo zrównania porównywanych grup ze względu na zmienne obserwowalne wydaje się, że niespełnione zostało założenie o warunkowej niezależności (*CIA*). W związku z tym prezentowane oszacowania efektu netto należy uznać za niewiarygodne. Powyższy przykład pokazuje, jak ważne jest umieszczenie w modelu wszystkich istotnych zmiennych predykcyjnych.

Zakończenie

Technika PSM pozwala dobrać do grupy interwencji, grupę kontrolną tak, że obie będą porównywalne w kategorii obserwowalnych charakterystyk. Technika ta oferuje skuteczny i efektywny sposób na redukcję obciążenia selekcyjnego podczas szacowania efektu przyczynowego działań. Wysuwanie wniosków na temat efektu przyczynowego będzie możliwe, o ile tylko spełnione są kluczowe założenia dla techniki PSM. Utrzymanie tych założeń zależy w praktyce od wielu czynników. Kluczem do ich spełnienia jest poprawna identyfikacja mechanizmu odpowiedzialnego za uczestnictwo jednostek w danym zdarzeniu oraz powiązanych z nim efektów. Z tego punktu widzenia ważnym czynnikiem warunkującym możliwość poprawnego wykorzystania techniki PSM jest dostępność odpowiednich danych. O ile tylko powyższe założenie jest spełnione, technika PSM może pomóc w odpowiedzi na ważne pytania badawcze, w tym jedno najważniejsze: *czy oceniana interwencja działa, tj., czy rzeczywiście wywołała efekty, dla których została powołana?*

PSM zdobywa coraz większą popularność ze względu na stosunkową prostotę w aplikacji i jednocześnie duże możliwości w obszarze wyciągania wniosków na temat przyczynowego oddziaływania zdarzeń. Szczególnie przydatna może się ona okazać w przypadku, gdy niemożliwe lub nieuprawnione jest wykorzystanie metody eksperymentalnej. Z tego względu pojawiają się opinie, że PSM jest w pewnym sensie tak fundamentalne dla badań obserwacyjnych, jak randomizacja dla badań eksperymentalnych. Oferuje bowiem *następny po najlepszym z rozwiązań (tj. po metodzie eksperymentalnej), nadzwyczaj prosty i skuteczny sposób na redukcję czy nawet eliminację obciążeń wyników, gdy randomizacja nie jest możliwa lub nie jest wykorzystywana w danym badaniu* (Gelman i in., 2004: 14). Z tych powodów technika PSM znajduje zastosowanie w coraz większej liczbie obszarów badawczych różnych nauk – w socjologii, psychologii, epidemiologii, politologii i innych.

W niniejszej pracy zaprezentowany został przykład wykorzystania techniki PSM w badaniach ewaluacyjnych interwencji publicznych. Można spodziewać się, że jej popularność, z uwagi na wymienione zalety, będzie wciąż rosła. Choć, jak już wielokrotnie zostało to podkreślone, możliwość zastosowania techniki PSM zależy od tego, jakimi danymi dysponuje badacz. Kwestią bardziej pierwotną jest, czy badacz w ogóle dysponuje jakimiś danymi. Z dotychczasowych doświadczeń realizacji chociażby programów finansowanych ze środków Unii Europejskiej wynika, że nie przykłada się zbyt wielkiej wagi do zapewnienia możliwości pomiaru efektu netto podejmowanych działań. Tym samym nie gromadzi się danych, które dalej mogłyby być wykorzystane na etapie oceny podjętego przedsięwzięcia. Dopóki pod tym względem nie zajdą istotne zmiany, zakres wykorzystania techniki PSM do rzetelnego pomiaru oddziaływania interwencji publicznej może okazać się jedynie incydentalny.

Co istotne – możliwości techniki PSM wykraczają poza aplikację w badaniach ewaluacyjnych. Jest ona bowiem użytecznym narzędziem minimalizacji błędów szacunku efektu przyczynowego, wynikającego z obecności mechanizmów selekcji. Z tego względu wszędzie tam, gdzie występuje inny niż losowy dobór jednostek do grupy interwencji, technika PSM może znaleźć zastosowanie. Wśród już wspomnianych zastosowań pojawiają się np. ciekawe propozycje wykorzystania jej w rozwijających się obecnie badaniach internetowych, które „cierpią" na nielosowy dobór respondentów.

Niezależnie od zastosowania warto mieć na względzie jedną uwagę. Technika PSM jest jedynie narzędziem w ręku analityka, który dysponując odpowiednimi danymi i przeprowadzając poprawnie proces analizy, może wypowiadać się na temat szacowanego efektu przyczynowego ocenianych dzia-

łań. W procesie wyciągania wniosków, czy też szerzej – dokonywania oceny działań, stanowi to jednak dopiero punkt wyjścia. Zwykle kolejnym krokiem po oszacowaniu efektu interwencji jest odpowiedź na pytanie, dlaczego wystąpiła taka, a nie inna wielkość efektu przyczynowego. Na to pytanie technika PSM może odpowiedzieć w bardzo ograniczonym zakresie.

Bibliografia

- Blalock H.M. (1977). *Statystyka dla socjologów*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Bryson A., Dorsett R., Purdon S. (2002). *The use of propensity score matching in the evaluation of active labour market policies*. Pobrany 16.03.2009 r. z www.dwp.gov.uk/asd/asd5/WP4.
- Bukowski M. (red.) 2008. *Zatrudnienie w Polsce 2007. Bezpieczeństwo na elastycznym rynku* pracy. Pobrany 16.03.2009 r. z http://www.mpips.gov.pl/_download.php?f=userfiles%2FFile%2FAnalizy%2FZwP_2007.pdf.
- Caliendo M., Kopeinig S. (2005). *Some Practical Guidance for the Implementation of Propensity Score Matching*. Berlin: German Institute for Economic Research IZA.
- Dehejia R.H., Wahba S. (1999). *Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs*. Journal of the American Statistical Association, Vol. 94, No. 448., s. 1053–1062.
- Dehejia R.H., Wahba S. (2002). *Propensity score matching methods for non-experimental causal studies*. Review of Economics and Statistics, Vol. 84, No.1, s. 151–161.
- Guo S. (2006). *Develop Innovative Methods in Secondary Analyses of Child Welfare Databases – Children's Bureau Discretionary Grants Program*. Grantee's Final Report
- Guo S., Barth R.P., Gibbons C. (2006). *Propensity score matching strategies for evaluating substance abuse services for child welfare clients*. School Children and Youth Services Review, Vol. 28, s. 357–383.
- Górniak J. (2005). *Ewaluacja jako czynnik doskonalenia zarządzania strategicznego w administracji publicznej: projekt SAPER*. W: A. Haber i in., *Ewaluacja programów o charakterze społeczno-gospodarczym finansowanych z funduszy strukturalnych*, s. 83-102 Warszawa: Ministerstwo Gospodarki i Pracy – Departament Koordynacji Polityki Strukturalnej.
- Górniak J. (2007). *Ewaluacja w cyklu polityk publicznych*. W: S. Mazur (red.), *Ewaluacja funduszy strukturalnych – perspektywa regionalna*. Kraków: Uniwersytet Ekonomiczny w Krakowie, Małopolska Szkoła Administracji Publicznej.
- Górniak J., Keler K. (2008). *Rola systemów wskaźników w ewaluacji*. W: K. Olejniczak, M. Kozak, B. Ledzion (red.), *Teoria i praktyka ewaluacji interwencji publicznych. Podręcznik akademicki*, s. 109–128. Warszawa: Wydawnictwa Akademickie i Profesjonalne, Akademia Leona Koźmińskiego.
- Greenland S. (2004). *An overview of methods for causal inference from observational studies*. W: A. Gelman, X. Meng (red.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An essential journey with Donald Rubin's statistical family*, s. 3–14. West Sussex: John Wiley & Sons Ltd.
- Haber A., Witowski W. (red.) (2006). *Phare Spójność Społeczno-Gospodarcza. Podsumowanie programu*. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości, Ministerstwo Rozwoju Regionalnego.
- Heckman J.J, Smith J.A. (1995). *Assessing the Case for Social Experiments*. The Journal of Economic Perspectives, Vol. 9, No. 2., s. 85–110.
- Heckman J.J., Ichimura H., Todd P.E. (1997). *Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme*, Review of Economic Studies, Vol. 64, s. 605–654.
- Holland P.W. (1986). *Statistics and Causal Inference*. Journal of the American Statistical Association, Vol. 81, No. 396., s. 945–960.

- Hosmer D.W., Lemeshow S. (2000). *Applied Logistic Regression,* 2nd Edition. New York: Wiley Series in Probality and Statistics.
- Karpiński J. (1985). *Przyczynowość w badaniach socjologicznych*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Konarski R., Kotnarowski M. (2007). *Zastosowanie metody propensity score matching w ewaluacji ex-post*. W: A. Haber. (red.), *Ewaluacja ex-post. Teoria i praktyka badawcza,* s. 156–183. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości.
- Komisja Europejska (1997). *Ewaluacja programów wydatków Unii Europejskiej: Przewodnik Ewaluacja w połowie okresu i ex-post*. Pobrany 01.10.2006 z doc.ukie.gov.pl/cd-rom/cd2/dokumenty/04_fundusze/wiecej/00_ocena_ex_post.doc.
- Lechner M. (2002). *A note on the common support problem in applied evaluation studies,* Discussion paper no. 2001-01. St. Gallen: University of St. Gallen.
- Lissowski G., Haman J., Jasiński M. (2008). *Podstawy statystyki dla socjologów*. Warszawa: Wydawnictwo Naukowe Scholar..
- Orr L.L. (1999) *Social experiments. Evaluating Public Programs With Experimental Methods*. London: SAGE Publications Inc.
- PAG Uniconsult, ARC Rynek i Opinia (2007). *Ewaluacja ex-post Phare 2003 Spójność Społeczna i Gospodarcza – komponent Rozwój Zasobów Ludzkich*. Raport z badania przeprowadzonego na zlecenie Polskiej Agencji Rozwoju Przedsiębiorczości. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości.
- PBS DGA (2006), *Ewaluacja ex-post komponentu regionalnego programu Phare 2002 Spójność Społeczna i Gospodarcza – Komponent Rozwój Zasobów Ludzkich*. Raport z badania przeprowadzonego na zlecenie Polskiej Agencji Rozwoju Przedsiębiorczości. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości.
- Rosenbaum P.R. (2002). *Observational Studies*, 2nd edition. New York: Springer-Verlag.
- Rosenbaum P.R. (2004). Matching in observational studies,. W: A. Gelman, X. Meng (red.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. An essential journey with Donald Rubin's statistical family,* s. 15–24. West Sussex: John Wiley & Sons Ltd.
- Rosenbaum P.R. (2005). Observational Study. W: *Encyclopedia of Statistics in Behavioral Science* Volume 3, s. 1451–1462. Chichester: John Wiley & Sons, Ltd.
- Rosenbaum P.R., Rubin D.B. (1983). *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika, Vol. 70, No. 1., s. 41–55.
- Rosenbaum P.R., Rubin D.B. (1983). *Reducing bias in observational studies using subclassification on the propensity score*, Journal of the American Statistical Association, 79, s. 516–524.
- Rossi P.H., Freeman H.E., Lipsey M.W. (1999). *Evaluation. A systematic approach,* 6th edition. Thousand Oaks: Sage Publications Inc.
- Rubin D.B. (2006). *Matched sampling for Causal Effects*. New York: Cambrige University Press.
- Rubin D.B. (2005). *Causal inference using potential outcomes: design, modeling, decisions*. Journal of the American Statistical Association, Vol. 100, s. 322–331.
- Rubin D.B., Thomas N. (1996*). Matching Using Estimated Propensity Scores: Relating Theory to Practise*. Biometrika, Vol. 52, No. 1., s. 249–264.
- Smith J.A., Todd P.E. (2005). *Does matching overcome LaLonde's critique of nonexperimental estimators?* Journal of Econometrics, Vol. 125, s. 305–353.

- Smith, H. L. (1997). *Matching with multiple controls to estimate treatment effects in observational studies*. Sociological Methodology, 27, s. 325–353.
- Stanisz A. (2007). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*. Tom 2. *Modele liniowe i nieliniowe*. Kraków: StatSoft Polska.
- Strawiński P. (2008). *Quasi-eksperymentalne metody ewaluacji*. W: A. Haber, M. Szałaj (red.), *Środowisko i warsztat ewaluacji*. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości.
- Sułek A. (1979). *Eksperyment w badaniach społecznych*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Winship Ch., Morgan S.L. (1999). *The Estimation of Causal Effects from Observational Data*. Annual Review of Sociology, Vol. 25., s. 659–706.

Spis tabel

Spis ilustracji