Rafał Trzciński

Rozdział 8. **Badania ewaluacyjne – szacowanie efektów programów publicznych**

Ewaluacja definiowana jest na wiele sposobów (por. Olejniczak 2008: 16). Można ją określić jako przedsięwzięcie badawcze „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). Badania ewaluacyjne – a w szczególności ewaluacje typu *ex post* – mają więc za zadanie dostarczać wiedzy na temat skuteczności i efektywności zrealizowanych interwencji. Wnioski w tym zakresie powinny być wykorzystane przy projektowaniu i realizacji kolejnych działań. Taki jest zresztą postulat – aby ewaluacja stanowiła element cyklu polityk publicznych, wywierając wpływ na podejmowane w ramach ich realizacji decyzje (Górniak 2007: 11). Postulaty te znajdują wyraz w propagowanym w ostatnich latach haśle realizacji polityk publicznych „w oparciu o dowody" (ang. *evidence-based policy*), zgodnie z którym administracja, podejmując decyzje w zakresie wydatkowania środków publicznych, powinna bazować na wiedzy, tzw. twardych danych i faktach, pochodzących z przeprowadzonych badań i analiz. W przeciwnym razie decydenci będą musieli polegać jedynie na własnej wiedzy, intuicji, ideologii czy, w najlepszym razie, samej teorii (Banks 2009: 4). Z tego punktu widzenia ewaluacja ma w procesie dostarczania właściwej wiedzy stanowić narzędzie stałego doskonalenia administracji publicznej i podejmowanych przez nią inicjatyw.

Oznacza to, że jednym z podstawowych zadań operacyjnych stawianych przed ewaluacją jest ustalanie rzeczywistych efektów zrealizowanych działań – wdrożonych interwencji, programów, projektów itd. Zadanie to dotyka kluczowego i jednocześnie najbardziej kłopotliwego w naukach społecznych zagadnienia, jakim jest pojęcie przyczynowości oraz identyfikacja zależności przyczynowych. Wyzwanie polega na konieczności badania nie tylko tego, co zmieniło się w rzeczywistości po podjęciu danej interwencji publicznej – np. projektu szkoleniowego dla osób poszukujących pracy, programu wsparcia dla przedsiębiorstw itp. – ale również ustalenia, do jakiego stopnia zaobserwowane zmiany – odpowiednio spadek bezrobocia, zmiana kondycji finansowej

przedsiębiorstw – są konsekwencją zrealizowanych działań. Ewaluacja jest oczywiście szerszą koncepcją, wykraczającą poza ocenę efektów interwencji. Jednak w tym podręczniku skoncentrujemy się na podklasie badań ewaluacyjnych, określanych jako tzw. ewaluacje oddziaływania (ang. *impact evaluation*), których podstawowym celem jest pomiar efektu przyczynowego zrealizowanych programów publicznych. Mogą one znajdować zastosowanie przy ocenie „interwencji" podejmowanych w celu zmian w stratyfikacji społecznej.

W dziedzinie zjawisk społecznych związek przyczynowy ma charakter czysto teoretyczny – nie da się go bezpośrednio zobaczyć. W praktyce badacz może najczęściej zaobserwować jedynie współwystępowanie pewnych zdarzeń lub co najwyżej ich następstwo. Po realizacji programu szkoleniowego dla osób bezrobotnych część z jego uczestników może znaleźć pracę, jednak nie da się wykluczyć sytuacji, w której obu faktów nie łączy żaden związek. Jak wiadomo, o występowaniu związku przyczynowego nie można wnioskować ani z samego faktu współwystępowania zdarzeń (występowanie korelacji nie oznacza zaistnienia relacji przyczynowo-skutkowej), ani (na tej samej zasadzie) z ich następstwa[1]. Powyższy problem znajduje bezpośrednie odzwierciedlenie na poziomie metodologii badań społecznych, w których związki przyczynowe mogą być identyfikowane jedynie w drodze procesu wnioskowania, a nie bezpośredniej obserwacji. Badacz skazany jest więc każdorazowo na pewien – czasem duży – margines niepewności, decydując się na sformułowanie wniosku o zaistnieniu relacji przyczynowo-skutkowej. Ze swej natury wnioskowanie w tym zakresie obarczone jest ryzykiem popełnienia błędu, którego weryfikacja pozostaje poza naszymi możliwościami poznawczymi, co odnosi się nie tylko do nauk społecznych.

8.1. Problem identyfikacji zależności przyczynowych w kontekście koncepcji stanów kontrfaktycznych

Ramy teoretyczne dla wnioskowania w zakresie identyfikacji zależności przyczynowych tworzy tzw. koncepcja stanów kontrfaktycznych (Heckman 2005: 1). Na gruncie badań społecznych, w tym badań ewaluacyjnych, zaczęła być ona szerzej wykorzystywana głównie za sprawą ekonometryków i statystyków takich jak R. Fisher, J. Neyman, W. Cochran, D. Cox, J. Heckman, A.D. Roy, R. Quandt czy D. Rubin. Podejście to przeżywa swój renesans w ostatnim dwudziestoleciu, warto jednak mieć świadomość, że jego korzenie sięgają do filozofii Davida Hume'a (Greenland 2004: 4). Główna idea, jaka stoi za koncepcją stanów kontrfaktycznych, sprowadza się do próby oszacowania hipotetycznych skutków zdarzeń, które stanowią alternatywę względem tego, co faktycznie się wydarzyło. Na przykład problem badawczy o charakterze kontrfaktycznym wyraża się w pytaniu o to, jak potoczyłyby się losy zawodowe uczestników programu szkoleniowego dla bezrobotnych, gdyby nie wzięli oni w nim udziału. Mając przybliżenie takiej hipotetycznej sytuacji, należałoby ją następnie porównać z tym, co faktycznie

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

się wydarzyło, próbując odpowiedzieć, czy fakt udziału w danym programie przekłada się w jakikolwiek sposób na sytuację zawodową jego uczestników.

Idea ta odwołuje się do następującego modelu. Dana jest populacja jednostek I[2]. W danym czasie każda jednostka i, pochodząca z populacji I, może znaleźć się w jednej z dwóch sytuacji, które opisuje zmienna $D \in \{0, 1\}$. Dla uproszczenia, kontynuując wcześniejszy przykład, przyjmijmy, że jednostkami są osoby bezrobotne, zdarzeniem zaś jest dowolny projekt szkoleniowy realizowany w ramach Aktywnych Polityk Rynku Pracy. Każda osoba może wziąć udział w programie szkoleniowym – wtedy $D_i = 1$ – lub może podjąć działanie przeciwne – w szczególności może nie robić nic – wtedy $D_i = 0$. Każdej z tych dwóch sytuacji odpowiada potencjalny skutek, wyrażany przez zmienną Y. W zależności od tego, w jakiej grupie znalazła się dana osoba, Y może przyjąć dla niej jedną z dwóch wartości: Y_{i1} lub Y_{i0}, gdzie Y_{i1} jest wartością, która zostałaby zaobserwowana, gdyby osoba znalazła się w grupie biorącej udział w szkoleniu, Y_{i0} zaś odpowiada wartości, która zostałaby zaobserwowana, gdyby osoba była poza grupą osób przeszkolonych. W przypadku programu aktywizującego bezrobotnych potencjalnym efektem może być np. wysokość zarobków w ciągu sześciu miesięcy od momentu udziału w szkoleniu. W zależności od tego, do której grupy należy jednostka, jeden z wyników – Y_{i1} lub Y_{i0} – jest wynikiem hipotetycznym, nieobserwowanym w rzeczywistości. Załóżmy jednak, że jednostki mają „przypisane" potencjalne wyniki dla każdego z dwóch stanów: dla tego, w którym się rzeczywiście znalazły, oraz dla zdarzenia przeciwnego. Innymi słowy, każda jednostka ma „przypisany" obserwowalny efekt, ale także nieobserwowalny efekt kontrfaktyczny. Dzięki powyższym danym możliwe jest określenie poszukiwanego (przyczynowego) efektu szkolenia dla osoby bezrobotnej. Dla wybranej jednostki i można go wyrazić w różnicy między Y_{i1} a Y_{i0} (Holland 1986: 947), tj.:

$$Y_i = Y_{i1} - Y_{i0} \tag{8.1}$$

W odniesieniu do programu szkoleniowego jest to więc różnica między zarobkami osoby, która wzięła udział w programie szkoleniowym (Y_{i1}), a zarobkami tej samej osoby, w sytuacji gdyby w programie nie uczestniczyła (Y_{i0}). W przypadku gdyby osoba rzeczywiście uczestniczyła w szkoleniu, jest to zestawienie faktycznych zarobków – obserwowalnych po udziale w szkoleniu – z zarobkami hipotetycznymi czy, innymi słowy, kontrfaktycznymi, które osoba posiadałaby w sytuacji przeciwnej do tej, w jakiej faktycznie się znalazła. Oczywiście w praktyce zaobserwowanie w danym czasie dla tej samej osoby skutków dwóch wykluczających się zdarzeń jest niemożliwe. Nie da się jednocześnie uczestniczyć i nie uczestniczyć w szkoleniu. W literaturze przedmiotu sytuacja ta nosi nazwę *fundamentalnego problemu wnioskowania przyczynowego* (Holland 1986: 947).

W teorii powyższy problem nie ma rozwiązania. Jednak w praktyce brak informacji na temat Y_{i0}, w warunkach, gdy znamy Y_{i1}, można potraktować jako problem braku danych (Rosenbaum i Rubin 1983; Heckman i in. 1997: 608). Jeśli tak, to jednym z działań, jakie można podjąć, jest próba imputacji brakujących informacji. Istnieje kilka

[2] Jednostkami mogą być osoby, instytucje, przedsiębiorstwa, a także grupy jednostek, regiony itp.

podejść, wykorzystywanych do tego celu. Jedno z częściej spotykanych sprowadza się do wykorzystania informacji o jednostkach, które nie brały udziału w danym zdarzeniu i w pewnych warunkach mogą stanowić tzw. grupę kontrolną. Rozwiązanie to przenosi problem z poziomu jednostki na poziom populacji, z której dana jednostka pochodzi (Holland 1986: 947).

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:

$$Y_{ATE} = E(Y_1 - Y_0) \tag{8.2}$$

co może być również zapisane jako:

$$Y_{ATE} = E(Y_1) - E(Y_0) \tag{8.3}$$

gdzie $E(Y_1)$ to przeciętny efekt szkolenia (w naszym przykładzie będą to zarobki), w sytuacji gdy wszystkie jednostki w populacji *I* wzięły udział w programie, $E(Y_0)$ zaś jest przeciętnym efektem braku szkolenia, w sytuacji, gdy żadna z osób z populacji *I* nie uczestniczyła w programie. Znów w praktyce ani $E(Y_1)$, ani $E(Y_0)$ nie może być jednocześnie poznane. Jeżeli jednak dana interwencja nie ma charakteru polityki uniwersalnej – tzn. nie obejmuje całej populacji – to możliwe jest zaobserwowanie dla części jednostek wyniku Y_{i1}, a dla pozostałej części wyniku Y_{i0}. W tej sytuacji informacje o osobach uczestniczących w programie mogą być wykorzystane do oszacowania $E(Y_1)$ (ponieważ jest to wartość średnia Y_1 określona dla populacji *I*), z kolei informacje o osobach nieuczestniczących w programie mogą zostać wykorzystane do oszacowania $E(Y_0)$.

Należy jednak zauważyć, że miara wyrażana przez Y_{ATE} ma istotne ograniczenie. Przedstawia ona efekt danego programu dla przeciętnej, losowo wybranej jednostki pochodzącej z populacji *I*, bez uwzględnienia, czy wzięła ona w nim udział, czy też nie. Częściej przedmiotem zainteresowania badacza będzie raczej efekt przyczynowy ograniczony tylko do osób, które uczestniczyły w ocenianej interwencji. Poszukiwany jest więc tzw. przeciętny efekt oddziaływania na jednostki poddane oddziaływaniu (ang. *treatment on treated effect – ATT*), który wyraża się zazwyczaj w następujący sposób:

$$Y_{ATT} = E(Y_1 - Y_0 | D = 1) = E(Y_1 | D = 1) - E(Y_0 | D = 1) \tag{8.4}$$

gdzie $E(Y_1 | D = 1)$ to średni wynik obserwowany po interwencji w grupie uczestników programu, $E(Y_0 | D = 1)$ zaś to średni wynik braku interwencji również w grupie uczestników programu. $E(Y_0 | D = 1)$ jest oczywiście wartością nieobserwowalną – kontrfaktyczną – niemniej, tak jak w przypadku $E(Y_1)$ lub $E(Y_0)$, przy założeniu, że jesteśmy w stanie ją imputować, może ona zostać oszacowana. W rzeczywistości dane jest $E(Y_0 | D = 0)$, czyli średni efekt obserwowany po interwencji dla osób, które z niej nie skorzystały[3]. Do oszacowania $E(Y_0 | D = 1)$ można więc przy pewnych założeniach, o których mowa dalej, wykorzystać $E(Y_0 | D = 0)$.

[3] Dane w tym sensie, że może zostać oszacowane.

Oczywiście krytyczne pytanie brzmi, w jakim stopniu ten pierwszy efekt odzwierciedla rzeczywiście ten drugi. Odpowiedź na to pytanie związana jest bezpośrednio z problemem występowania tzw. mechanizmów selekcji. Należy przez nie rozumieć sposób, w jaki jednostki – osoby, podmioty gospodarcze itd. – stają się uczestnikami danej interwencji. Proces selekcji mogą warunkować rozmaite czynniki – kryteria formalne stawiane przed beneficjentami programów (np. wiek, okres przebywania na bezrobociu itp.), jak również cechy specyficzne dla różnych jednostek (tzw. autoselekcja). Z punktu widzenia możliwości identyfikacji zależności przyczynowych sytuacją idealną jest, gdy proces selekcji beneficjentów wsparcia jest niezależny od efektów uzyskiwanych przez poszczególne podmioty (Martini 2011: 27). Za stwierdzeniem tym stoi następujące uzasadnienie. Jednostki z grupy będącej uczestnikami danego programu, podobnie jak jednostki nim nieobjęte, można opisać za pomocą zestawu zmiennych – w odniesieniu do osób będą to określone cechy społeczno-ekonomiczne, posiadane umiejętności, motywacje itp. W wielu przypadkach może się okazać, że poszczególne zmienne mogą pozostawać w związku zarówno z prawdopodobieństwem udziału w danej interwencji (np. pewne cechy mogą sprzyjać udziałowi w danym programie), jak i z później obserwowanym skutkiem. Innymi słowy, efekt udziału w danym programie może zależeć od tego, jaka jest dystrybucja cech wpływających na zmienną D (uczestnictwo) i jednocześnie zmienną Y (obserwowana zmiana) w grupie podmiotów poddanych i niepoddanych oddziaływaniu analizowanej interwencji. Z kolei jeśli proces selekcji uczestników programu jest powiązany z uzyskiwanymi przez nich efektami, to $E(Y_0 \mid D = 0)$ będzie niewłaściwym oszacowaniem $E(Y_0 \mid D = 1)$. Obie grupy będą bowiem nieporównywalne pod względem szeregu charakterystyk, i tym samym oszacowanie efektu przyczynowego danej interwencji będzie obciążone. W istocie praktyka pokazuje, że np. w czasie realizacji programów dla osób bezrobotnych rzadko kiedy grupa objęta pomocą jest porównywalna z grupą osób, które nie skorzystały ze wsparcia (Trzciński 2009: 16).

W skrócie, nieuwzględnienie problemu występowania 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 kontroli dodatkowych czynników, może prowadzić do uzyskania błędnych oszacowań efektu przyczynowego, a ostatecznie błędnych wniosków w odniesieniu do analizowanego instrumentu wsparcia.

8.2. Metody pomiaru efektów programów publicznych. Eksperyment i podejścia quasi-eksperymentalne

Najbardziej pożądaną sytuacją z punktu widzenia minimalizacji obciążenia selekcyjnego byłaby więc niezależność procesu selekcji uczestników danego programu od jego efektów. Owa niezależność może być zagwarantowana, gdy selekcja uczestników do danego programu ma charakter losowy, tzn. jedynym czynnikiem mającym wpływ na dobór podmiotów do warunku interwencji jest dowolny mechanizm dający każdej jednostce jednakowe (lub przynajmniej znane) prawdopodobieństwo znalezienia się w grupie uczestników programu. Gdy selekcja uczestników nie ma charakteru losowe-

go, obserwowaną różnicę między beneficjentami i grupą porównawczą zawsze można postrzegać jako sumę dwóch elementów: faktycznego efektu polityki oraz różnicy wynikającej z samego procesu selekcji (Martini 2009: 27). Losowość zapewnia, że jedynym czynnikiem różnicującym jednostki w grupie objętej oddziaływaniem danej interwencji i jednostki z niej wykluczone jest właśnie fakt udziału w danym zdarzeniu. Innymi słowy, przy wykorzystaniu randomizacji $E(Y_0 \mid D = 0)$ powinno być teoretycznie idealnym odzwierciedleniem $E(Y_0 \mid D = 1)$.

W praktyce badań ewaluacyjnych, w których dąży się do ustalenia efektu przyczynowego programów publicznych, powyższa procedura bywa stosowana w postaci eksperymentów zrandomizowanych. W typowej sytuacji sprowadzają się one do utworzenia, z dostępnej populacji jednostek kwalifikujących się do udziału w danym programie, grupy eksperymentalnej oraz grupy kontrolnej. Przy tworzeniu grup wykorzystywany jest oczywiście mechanizm losowy. Grupa eksperymentalna objęta jest dalej oddziaływaniem danej interwencji, natomiast grupa kontrolna jest z niej wykluczona. Po realizacji interwencji obie grupy porównuje się w wymiarze, w jakim oczekiwano uzyskania określonych efektów (Orr 1999: 149).

Choć właściwości eksperymentów w zakresie możliwości ustalania efektów przyczynowych są powszechnie uznane, to ich wykorzystanie w praktyce jest ograniczone. Wynika to z kilku czynników, wśród których wskazać należy przede wszystkim wysokie koszty, długi czas realizacji, problemy etyczne oraz problemy w implementacji terenowej, które mogą ostatecznie skutkować uzyskaniem obciążonych oszacowań efektów przyczynowych. Te oraz inne problemy (szerzej por. Trzciński 2009: 18) powodują, że eksperyment zrandomizowany jest raczej swego rodzaju typem idealnym procedury badawczej, który w rzeczywistości jest niezmiernie trudny do odtworzenia. Z tego względu nie jest to metoda uniwersalna, którą można stosować w każdej sytuacji, gdy poszukiwany jest efekt przyczynowy danego zdarzenia. W tym miejscu należy jednak nadmienić, że ocena eksperymentów i ich statusu nie jest kwestią całkowicie rozstrzygniętą. Aktualnie w środowisku ewaluacyjnym trwa debata[4] na temat tego, w jakim stopniu eksperyment jest jedyną metodą, która może być wykorzystana do ustanawiania wniosków o naturze przyczynowo-skutkowej, stanowiąc tzw. złoty standard (Rossi i in. 1999: 279), a w jakim uprawnione jest zastosowanie do tego celu również innych podejść. Istnieje pokaźne grono zwolenników pierwszego z twierdzeń[5], którzy postrzegają eksperyment jako jedyną właściwą drogę do identyfikacji zależności przyczynowych. Umiarkowane podejście, które przyjmujemy tu za Pattonem, wskazuje, że każdorazowo należy poszukiwać rozwiązań stosownych dla danej sytuacji, unikając metodologicznej ortodoksji (Patton 2011). Możliwość wykorzystania eksperymentów jest kwestią otwartą, i o ile ich zastosowanie znajduje uzasadnienie, należałoby je rekomendować.

Alternatywą dla zastosowania eksperymentów w badaniach skupiających się na pomiarze efektów przyczynowych są tzw. podejścia quasi-eksperymentalne. Ich uwaga

[4] Por. *The 2004 Claremont Debate: Lipsey vs. Scriven Determining Causality in Program Evaluation & Applied Research: Should Experimental Evidence Be the Gold Standard?*

[5] W skrajnym przypadku bywają oni określani mianem *randomistas* (Ravallion 2009: 1).

skupia się podobnie jak w przypadku eksperymentów na znalezieniu właściwego punktu odniesienia dla beneficjentów danego programu, czyli odpowiednika grupy kontrolnej. Wykorzystywane metody różnią się między sobą pod względem trudności w aplikacji oraz z punktu widzenia przyjmowanych założeń, które w większości mają charakter nietestowalny. Poniżej zaprezentowane zostały wybrane podejścia.

8.2.1. Metoda różnicy w różnicach

Stosunkowo prostą strategią estymacji efektów danej interwencji jest procedura określana mianem różnicy w różnicach (*Difference-in-Differences* – DiD). Jej główna idea sprowadza się do założenia, że potencjalne efekty programów publicznych podlegają pewnej dynamice w czasie – zarówno w grupie beneficjentów wsparcia, jak i w grupie z niej wykluczonej. Co prawda nie wiemy, jak zmieniłaby się sytuacja beneficjentów danego programu, gdyby nie wzięli oni w nim udziału (tj. w sytuacji kontrfaktycznej), ale można założyć, że jej dobrym przybliżeniem jest trend wzrostu obserwowany w grupie podmiotów, które ze wsparcia nie skorzystały. Biorąc za przykład program wsparcia przedsiębiorstw, polegający na dofinansowaniu procesu modernizacji firm (zakup nowych maszyn i urządzeń itp.), można oczekiwać, że ewentualna zmiana poziomu zatrudnienia (jako jeden z oczekiwanych efektów interwencji) dotyczyć będzie zarówno podmiotów, które skorzystały z programu, jak i tych, które znalazły się poza nim. O ile tylko jesteśmy skłonni przyjąć założenie, że bez udziału w programie trend rozwoju beneficjentów byłby taki jak w grupie porównawczej, o tyle możliwe jest oszacowanie wielkości efektu interwencji. W praktyce sprowadza się to do porównania skali zmiany sytuacji beneficjentów interwencji publicznej w stosunku do zmiany obserwowanej w tym samym czasie u wytypowanej grupy porównawczej. Można to wyrazić w następujący sposób:

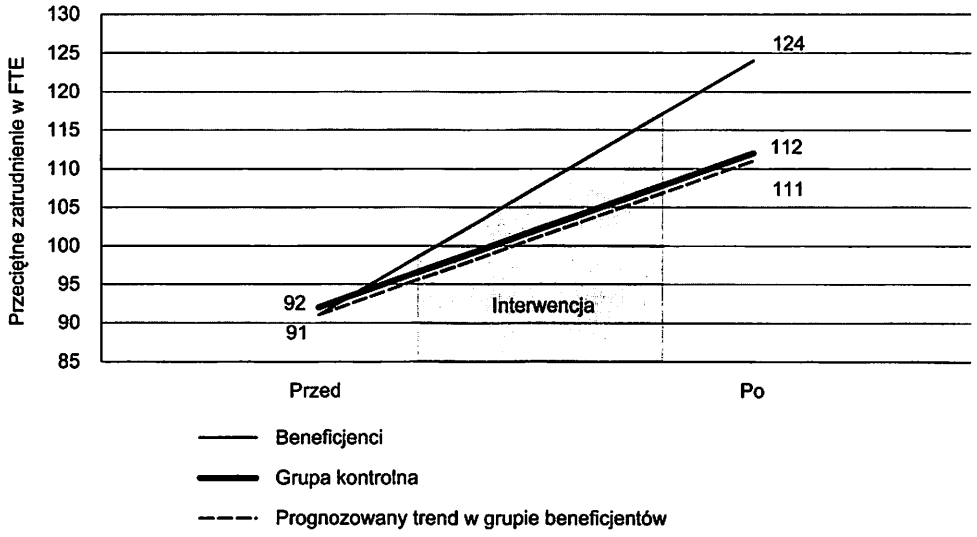
$$\text{Efekt interwencji} = (Y_{1,\,PO} - Y_{1,\,PRZED}) - (Y_{0,\,PO} - Y_{0,\,PRZED}) \tag{8.5}$$

gdzie $Y_{1,\,PO}$ odpowiada wartości zmiennej utożsamianej z potencjalnym efektem interwencji dla grupy objętej programem, oszacowanym po udziale w programie, $Y_{1,\,PRZED}$ zaś to wartość tej samej zmiennej, ustalona przed programem, natomiast $Y_{0,\,PO}$ oraz $Y_{0,\,PRZED}$ to analogiczne oszacowania zmiennej ustalone dla grupy, która nie była włączona do danej interwencji (Bertrand i in. 2004: 249).

Dla rozpatrywanego tu przykładowo programu wsparcia dla przedsiębiorstw procedurę szacowania efektu z wykorzystaniem metody różnicy w różnicach przedstawia wynik jednego z badań nad wpływem dotacji bezpośrednich na funkcjonowanie przedsiębiorstw[6]. W przeprowadzonej ewaluacji jednym z wymiarów analizy była zmiana za-

[6] Chodzi o badanie zrealizowane przez Polską Agencję Rozwoju Przedsiębiorczości (PARP). Więcej informacji na temat tego programu zawiera publikacja: Pokorski J. (red.) (2010). *Ocena instrumentów wsparcia bezpośredniego przedsiębiorstw. Podsumowanie wyników ewaluacji wybranych Działań SPO WKP*. Wybrane wyniki badania – zob.: http://www.google.pl/url?sa=t&rct=j&q=impact%20seminar%20warsaw%20rafa%C5%82&source=web&cd=2&ved=0CFAQFjAB&url=http%3A%2F%2Fec.europa.eu%2Fregional_poli-

trudnienia we wspartych podmiotach oraz w wytypowanej grupie kontrolnej. Uzyskany wynik przedstawiony jest na rysunku 8.1.



Rysunek 8.1. Przykład wykorzystania estymacji DiD w badaniu wpływu dotacji bezpośrednich na wzrost poziomu zatrudnienia w przedsiębiorstwach

Źródło: opracowanie własne na podstawie badania PARP przeprowadzonego w 2011 r.

W przedstawionym przykładzie dana jest liczba pracowników zatrudnionych w grupie beneficjentów wsparcia oraz w grupie podmiotów, które z programu nie skorzystały. Wyniki przeprowadzonego pomiaru odnoszą się do dwóch punktów w czasie – przed udziałem w interwencji i po nim. Firmy, które skorzystały ze wsparcia, przed przystąpieniem do programu zatrudniały przeciętnie 91 pracowników, natomiast podmioty, które znalazły się poza interwencją – 92. Po programie przedsiębiorstwa zatrudniały odpowiednio średnio 124 oraz 112 osób.

Proste porównanie sytuacji typu przed-po wskazywałoby, że beneficjenci wsparcia zwiększyli zatrudnienie średnio o 33 osoby. Jednak wartość ta będzie najprawdopodobniej niewłaściwym oszacowaniem efektu przyczynowego interwencji. Uznanie tego wyniku za faktyczny efekt wiązałoby się bowiem z przyjęciem dodatkowego założenia, że w grupie beneficjentów nie ma żadnej naturalnej dynamiki, czyli innymi słowy, bez udziału w programie poziom zatrudnienia pozostałby na stałym poziomie. Założenie to wydaje się mało realne[7], zwłaszcza że w opisywanym przykładzie oba pomiary dzielił okres około pięciu lat.

cy%2Fimpact%2Fevaluation%2Fconf_doc%2Fwarsaw_12122011%2Ftrzcinski_12_12_11_ie.ppt&ei=V_q7T8uhJ4KJ0AXdw-SfAw&usg=AFQjCNHj_sWygJP--8pdnVH9zDsLJhCrGw – stan na 29 maja 2012 r.

[7] Choć niestety przyjmowanie tego typu założeń w badaniach ewaluacyjnych pozostaje wciąż spotykaną praktyką.

Podobnie porównanie sytuacji beneficjentów z sytuacją grupy kontrolnej po zakończeniu realizacji interwencji nie wydaje się właściwym oszacowaniem efektu programu. W tym przypadku byłby to wynik 12 dodatkowych miejsc pracy względem grupy porównawczej. Chcąc przyjąć tę wartość za miarę efektu programu, należałoby jednak założyć, że nie występuje problem obciążenia selekcyjnego, o którym mowa była wcześniej. Założenie to również jest trudne do utrzymania, jak bowiem wspomniano, beneficjenci programów publicznych zazwyczaj różnią się od podmiotów, które nie skorzystały ze wsparcia. Co prawda w zaprezentowanym przykładzie poziom zatrudnienia „na wejściu" jest w obu grupach niemal identyczny, jednak jest to konsekwencja przyjętej procedury doboru grupy kontrolnej, o czym mowa dalej. W tym miejscu, bez dodatkowych informacji, nie jesteśmy w stanie stwierdzić, czy podmioty z obu grup nie różnią się pod względem innych wymiarów, takich jak np. branża, skala działalności itp.

Strategia zastosowana w podejściu różnicy w różnicach do pewnego stopnia niweluje oba wymienione powyżej problemy. Z jednej strony dopuszcza istnienie naturalnej dynamiki niezależnie od samej interwencji, a więc bierze np. pod uwagę oddziaływanie czynników makroekonomicznych, tak długo, jak owo oddziaływanie wywiera jednakowy wpływ na grupę beneficjentów i grupę porównawczą (Bryson i in. 2002: 7). Z drugiej strony w strategii DiD do pewnego stopnia brane są pod rozwagę potencjalne różnice pomiędzy porównywanymi grupami (trend grupy kontrolnej jest dopasowywany do wyjściowej sytuacji beneficjentów).

Jaki jest więc możliwy efekt programu w przykładzie powyżej? Prognozowany trend beneficjentów w sytuacji braku udziału w interwencji przedstawiony został na rysunku 8.1 linią przerywaną. Wskazuje on, że w przypadku braku interwencji faktyczni beneficjenci wsparcia prawdopodobnie zatrudnialiby 111 pracowników. Porównując tę sytuację z tym, co jest rzeczywiście obserwowane po realizacji działania, można dojść do wniosku, że program wpłynął na wzrost zatrudnienia we wspartych przedsiębiorstwach w wymiarze 13 dodatkowo zatrudnionych osób. Analogiczny wynik otrzymamy po podstawieniu właściwych danych do przedstawionego wcześniej równania. Należy powtórzyć, że wynik ten może być rzeczywiście utożsamiany z efektem interwencji, jeśli spełnione będzie kluczowe założenie przyjmowane w podejściu DiD, stanowiące, że trend obserwowany w grupie kontrolnej jest właściwym przybliżeniem sytuacji kontrfaktycznej beneficjentów danego programu wsparcia (tzw. założenie o równoległości trendów).

Spełnienie powyższego założenia w praktyce zależeć będzie od szeregu czynników, w tym w dużej mierze od tego, w jaki sposób dobrana zostanie grupa kontrolna. Oczywiście sposób doboru grupy kontrolnej w przypadku podejścia DiD pozostaje kwestią otwartą – poczynając od najprostszej z możliwych sytuacji, w której porównywana jest grupa uczestników programu z wszystkimi podmiotami, które w nim nie uczestniczyły. Biorąc jednak pod uwagę problemy potencjalnego obciążenia selekcyjnego, które będzie miało wpływ na trend, w jakim mogą rozwijać się porównywane populacje, musimy zdawać sobie sprawę, że nie będzie to rozwiązanie optymalne. Bardziej racjonalnym podejściem jest kontrola istotnych charakterystyk w obu porównywanych grupach,

tak aby zagwarantować w wymiarze obserwowalnych zmiennych brak istotnych różnic. W przedstawionym powyżej przykładzie dobór grupy kontrolnej oparto na warunkowym prawdopodobieństwie udziału w interwencji, oszacowanym w ramach techniki *propensity score matching*, która została scharakteryzowana w kolejnej części tego rozdziału. Dzięki temu udało się zniwelować istotne, wstępne różnice pomiędzy porównywanymi populacjami. Im staranniej dobrana jest grupa kontrolna, tym bardziej przekonujące będą oszacowania efektu przyczynowego, co niestety niekoniecznie musi znaczyć, że będą bliższe prawdy. Przyjęte założenie o równoległości trendów jest bowiem nieweryfikowalne w praktyce. Chcąc je uwiarygodnić, można dokonać dodatkowych analiz, polegających na sprawdzeniu trendów porównywanych grup, w czasie poprzedzającym realizację danego programu. Procedura ta wymaga jednak dodatkowych danych (Albouy 2004: 4), które niestety nie zawsze są dostępne.

8.2.2. Technika *propensity score matching*

Strategie redukcji obciążenia selekcyjnego przyjmowane w technikach quasi-eksperymentalnych podążają za stosunkowo intuicyjną ideą. Sprowadza się ona do próby eliminacji obserwowalnych różnic pomiędzy porównywanymi populacjami – tzn. grupą objętą daną interwencją oraz dobraną grupą kontrolną. Jedno z możliwych rozwiązań stanowią techniki oparte na schemacie prób dopasowanych według cech (ang. *matched samples*). W podejściach tych wychodzi się z założenia, że wszystkie istotne różnice między grupą, która wzięła udział w programie, i grupą nieobjętą interwencją można w całości wytłumaczyć w kategoriach obserwowalnych charakterystyk (Bryson i in. 2002: 10). W związku z tym obciążenie selekcyjne można minimalizować przez „upodobnienie" jednostek z obu grup na wektorze pewnego zbioru cech X.

W praktyce powyższe podejście realizowane jest przez poszukiwanie dla każdej jednostki z grupy uczestniczącej w danym działaniu przynajmniej jednej identycznej pod pewnym względem jednostki z populacji wyłączonej z interwencji. Łączenie odbywa się na podstawie wartości zmiennych charakteryzujących podmioty w obu grupach. W pary dobierane są jednostki o takich samych wartościach wszystkich zmiennych, składających się na określony przez badacza wektor zmiennych X. Z tego powodu procedura ta bywa czasem nazywana poszukiwaniem tzw. statystycznych bliźniaków (Cichocki i Tyrowicz 2010: 83). Gdy mamy finalnie dwie identyczne grupy (z punktu widzenia przyjętego koszyka zmiennych), możliwe jest ich porównanie w celu ustalenia nieobciążonego efektu interwencji.

Aby wynik przeprowadzonego porównania mógł odzwierciedlać poszukiwany efekt przyczynowy danego programu, muszą być spełnione dwa założenia. Po pierwsze, należy założyć, że (Y_0, Y_1) i D są niezależne od siebie, ze względu na wartości zmiennych w X, tj.:

$$(Y_0, Y_1) \perp D \mid X \tag{8.6}$$

gdzie, tak jak wcześniej, Y_0 to wynik braku udziału w interwencji (tzn. skutek znalezienia się w grupie kontrolnej), Y_1 to wynik udziału w interwencji, „$\perp$" oznacza niezależność, D jest zmienną identyfikującą przynależność grupową, X jest zaś wspomnianym powyżej wektorem zmiennych charakteryzujących jednostki. Oczywiście Y_0 obserwujemy tylko dla podmiotów, które nie wzięły udziału w danej interwencji, a Y_1 tylko dla jednostek z grupy beneficjentów wsparcia. Jeśli powyższe założenie jest spełnione, to wynik braku udziału w programie dla jednostek w nim uczestniczących (tj. w sytuacji kontrfaktycznej) jest taki sam jak obserwowany wynik braku udziału w programie dla osób z dobranej grupy porównawczej (czyli w sytuacji, którą możemy faktycznie zaobserwować), tj.:

$$E(Y_0 \mid X, D = 1) = E(Y_0 \mid X, D = 0) = E(Y_0 \mid X) \tag{8.7}$$

Innymi słowy, warunkowo względem zmiennych X wynik obserwowany w grupie podmiotów nieuczestniczących w interwencji przedstawia sytuację kontrfaktyczną dla beneficjentów danego działania (Heckman i in. 1997: 610). Warto zauważyć, że założenie 8.6 działa również niejako w drugą stronę, tzn. na jego mocy warunkowo względem X wynik Y_1 – obserwowany w grupie beneficjentów ($D = 1$) – odpowiada kontrfaktycznemu wynikowi uczestnictwa w programie dla grupy kontrolnej ($D = 0$). Przyjmując założenie 8.6, można więc odpowiedzieć na pytanie o to, jaki byłby los jednostek nieuczestniczących w programie, gdyby (przeciwnie niż w rzeczywistości) wzięły one udział w interwencji. Zazwyczaj ten ostatni warunek nie jest przedmiotem zainteresowania ewaluatorów – do oszacowania przeciętnego efektu oddziaływania interwencji na jednostki poddane interwencji (8.4) wystarczy więc słabsza wersja formuły 8.6, którą można wyrazić w sposób następujący (Heckman i in. 1997: 611):

$$Y_0 \perp D \mid X \tag{8.8}$$

W praktyce, mimo że 8.8 jest słabszym założeniem od 8.6, jest ono założeniem mocnym i w rzeczywistości nietestowalnym. Przyjmuje się w nim bowiem, że wszelkie ewentualne różnice między grupą eksperymentalną i grupą kontrolną, występujące na zmiennych nieobserwowanych (niezmierzonych bądź niedających się zmierzyć lub po prostu nieuwzględnionych w wektorze X), są nieistotne. Kluczem do spełnienia założenia 8.8 jest więc posiadanie odpowiedniego zbioru danych, który obejmuje wszystkie istotne charakterystyki „odpowiedzialne" za uczestnictwo jednostek w testowanym „działaniu" i za jego efekty. Gdy dane, którymi dysponujemy, nie zawierają wszystkich zmiennych „odpowiedzialnych" za uczestnictwo i wynik interwencji (działania), założenie 8.8, a więc i 8.6, będzie niespełnione.

W praktyce niezwykle trudne jest uwzględnienie i dokonanie pomiaru wszystkich zmiennych spełniających powyższe wymogi. W przypadku programów dotyczących wsparcia dla rynku pracy trudno jest, przykładowo, uchwycić charakterystyki odnoszące się do indywidualnych motywacji uczestniczących w nich osób. Jednak jeżeli założenie 8.8 jest spełnione, dopasowanie jednostek z grupy „interwencyjnej" i grupy kontrolnej na podstawie wartości X odgrywa analogiczną rolę jak mechanizm randomizacji

w ramach klasycznego eksperymentu, tj. proces selekcji do grupy interwencji może być postrzegany jako losowy (Bryson i in. 2002: 10). Jest to pożądana właściwość, która pozwala wyeliminować ewentualne obciążenie wynikające z obecności różnych mechanizmów selekcji.

Drugie założenie przyjmowane w technikach *matched samples* mówi o tym, że każda jednostka w populacji może zarówno uczestniczyć w ocenianej interwencji, jak i znaleźć się poza nią (Rosenbaum i Rubin 1983: 43). Można to zapisać jako:

$$0 < \Pr(D = 1 \mid X) < 1, \forall X \tag{8.9}$$

W literaturze metodologicznej warunki 8.6 i 8.9 łącznie noszą miano założenia warunkowej niezależności (*conditional independence assumption* – CIA)[8].

Ograniczeniem tego podejścia w zastosowaniu praktycznym są trudności doboru grupy kontrolnej, które rosną w miarę zwiększania liczby zmiennych uwzględnionych w wektorze X. Zestaw kontrolowanych cech powinien być możliwie najszerszy, co koliduje jednak z możliwościami znalezienia i dobrania w pary identycznych jednostek. Pogodzenie tych dwóch sprzecznych warunków, a więc i praktyczna realizacja technik *matched samples*, wymaga relatywnie dużych zbiorów danych, w porównaniu z wielkością grupy beneficjentów. W przypadku zmiennych binarnych liczba wariacji możliwych cech rośnie zgodnie z wzorem 2^n, gdzie n to liczba kontrolowanych zmiennych. Na przykład dokładne dopasowanie grupy kontrolnej na podstawie 20 dwuwartościowych cech może teoretycznie generować konieczność posiadania ponaddwumilionowego zbioru potencjalnych jednostek kontrolnych. Problem oczywiście komplikuje się, jeśli zmienne, charakteryzujące jednostki, mają więcej niż dwie wartości (np. wykształcenie) lub są zmiennymi ciągłymi (np. wysokość zarobków). Dlatego nawet gdy dysponujemy relatywnie dużymi zbiorami danych, często może okazać się, że utworzenie grupy kontrolnej będzie niemożliwe z uwagi na brak jednostek kontrolnych, które odpowiadałyby pod względem wytypowanych cech beneficjentom wsparcia.

Z tego właśnie powodu poszukiwania badaczy zwróciły się w kierunku innych rozwiązań, które z jednej strony zawierałyby analogiczny mechanizm minimalizacji obciążenia selekcyjnego, a z drugiej były łatwiejsze w praktycznym zastosowaniu. W rezultacie wypracowano podejście zwane techniką *propensity score matching* (PSM). Rozwiązanie to sprowadza się do pomysłu, aby dokładne dopasowanie według ustalonego wektora zmiennych X zastąpić zbalansowaniem X, tj. utworzyć grupę kontrolną, która będzie miała taki sam rozkład zmiennych w X jak grupa objęta interwencją (Rosenbaum 2004: 18). Podejście to przedstawione zostało po raz pierwszy w 1983 r. przez Paula Rosenbauma i Donalda Rubina w artykule *The central role of the propensity score in observational studies for causal effects*. Zgodnie z ich propozycją, zbalansowanie zmiennych można uzyskać nie tylko przez łączenie według X, ale również według określonej funkcji X, która ma tzw. właściwości balansujące (ang. *balancing score*). Funkcja ta musi spełniać następujący warunek (Rosenbaum i Rubin 1983: 42):

[8] (Rosenbaum i Rubin 1983), *selection on observables* (Barnow i in. 1980), *conditional independence* (Lechner 1999), *exogeneity* (Imbens 2004). Zob. również Guo i in. (2006: 362).

$$X \perp D \mid b(X) \tag{8.10}$$

Informuje on o tym, że warunkowy rozkład X względem wartości $b(X)$ jest taki sam dla grupy objętej interwencją ($D = 1$) jak dla jednostek w grupie kontrolnej ($D = 0$). Innymi słowy, zmienne zawarte w X tracą swoje właściwości predykcyjne odnośnie do tego, które jednostki znajdą się w ewaluowanym programie (Rosenbaum 2004: 18). Jak dowodzą Rosenbaum i Rubin, jedną z takich funkcji jest tzw. *propensity score* (Rosenbaum i Rubin 1983: 42). Zdefiniowana jest ona jako warunkowe prawdopodobieństwo doboru obserwacji do zdarzenia ($D = 1$), względem wektora zmiennych X:

$$P(X) = Pr(D = 1 \mid X) \tag{8.11}$$

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:

$$(Y_0, Y_1) \perp D \mid P(X) \tag{8.12}$$

oraz

$$0 < Pr(D = 1 \mid P(X)) < 1 \; \forall \; P(X) \tag{8.13}$$

Technika PSM, podobnie jak technika *matched samples*, ma więc na celu utworzenie grupy kontrolnej, składającej się z jednostek w jak największym stopniu podobnych do tych, które objęte zostały ocenianą interwencją. Zasadnicza różnica polega na tym, że w przypadku techniki PSM dopasowywanie jednostek parami odbywa się na podstawie wartości tylko jednej zmiennej, tj. wspomnianego *propensity score*. Technika ta jest więc sposobem na redukcję liczby 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 2009: 187).

Implementacja techniki PSM jest w praktyce procesem składającym się zwykle z trzech głównych etapów (Guo 2006: 362). Pierwszy z nich polega na estymowaniu szukanych wartości *propensity score*. W tym celu wykorzystać można np. model regresji logistycznej, w którym zmienną zależną jest fakt przynależenia do grupy objętej oddziaływaniem interwencji, z kolei zmiennymi niezależnymi są cechy, które, jak już wspomniano, w założeniu mają wpływać z jednej strony na przewidywany wynik (Y), z drugiej na uczestnictwo (D) w danym programie.

Drugim etapem jest dokonanie doboru jednostek do grupy kontrolnej na podstawie wyliczonego warunkowego prawdopodobieństwa udziału w programie. Dobór jednostek do grupy kontrolnej może odbywać się na wiele sposobów. Jednym z najprostszych jest tzw. metoda najbliższego sąsiada, a więc dopasowanie jednostek najbardziej podobnych, tj. o najbliższej wartości *propensity score*[9]. Efektem procedury łączenia jest oczywiście

[9] Jest wiele różnych sposobów/algorytmów doboru grup kontrolnych w technice PSM. Oprócz wspomnianej metody najbliższego sąsiada do najbardziej popularnych należą metoda z limitem (ang. *nearest neighbor*

otrzymanie grupy kontrolnej, która zgodnie z założeniem powinna mieć zbalansowane wszystkie zmienne wykorzystane w modelu prawdopodobieństwa. Grupa kontrolna będzie więc w zakresie wybranego zestawu cech podobna do istniejącej grupy interwencji. Istotnym elementem tego etapu jest weryfikacja, jak dalece łączenie doprowadziło faktycznie do upodobnienia porównywanych populacji. Trzecim etapem jest analiza efektów przeprowadzona na podstawie porównania grupy interwencji z utworzoną grupą kontrolną. Można tu zastosować zwykłe porównanie średnich, jak również bardziej rozbudowane podejścia, takie jak np. opisany wcześniej estymator DiD.

W przypadku Polski technika PSM została m.in. wykorzystana w badaniu ewaluacyjnym, które posłużyło za przykład omówionego powyżej estymowania efektów z wykorzystaniem podejścia DiD (por. rysunek 8.1). Analizowano efekty działania: *Wzrost konkurencyjności małych i średnich przedsiębiorstw poprzez inwestycje*, które było wdrażane w ramach unijnego *Sektorowego Programu Operacyjnego Wzrost Konkurencyjności Przedsiębiorstw*, mającego na celu zwiększenie konkurencyjności polskich MSP (małych i średnich przedsiębiorstw) przez modernizację ich oferty technologicznej i produktowej. Przedsiębiorstwa będące beneficjentami programu otrzymały bezzwrotną pomoc finansową, pokrywającą istotną część kosztów realizowanych projektów inwestycyjnych.

Głównym celem badania ewaluacyjnego było oszacowanie efektów udzielonych dotacji, w tym m.in. analiza wpływu dotacji bezpośrednich na zwiększenie zatrudnienia w MSP. Jako grupę kontrolną wytypowano przedsiębiorstwa, które ubiegały się o wsparcie (tj. złożyły wniosek o dofinansowanie projektu inwestycyjnego), jednak go nie otrzymały (tzw. wnioskodawcy nieskuteczni). Ponieważ obie grupy istotnie różniły się między sobą pod względem wielu cech, ocena efektów polegająca na porównaniu grupy uczestników programu z grupą podmiotów wykluczonych narażona była na znaczący błąd pomiaru. Ewentualna poprawa sytuacji w grupie beneficjentów mogła wynikać bowiem z ich wcześniejszego potencjału, posiadanych zasobów i możliwe, że byłaby zaobserwowana niezależnie od otrzymanego wsparcia finansowego. Dlatego w celu minimalizacji błędu oszacowania efektów interwencji, wynikających z różnic pomiędzy obiema porównywanymi grupami (tj. z mechanizmów selekcji), zastosowano technikę PSM, której zadaniem było zrównanie (uczynienie podobnymi) obu grup w zakresie wybranych charakterystyk.

Jeden z wyników tego badania przedstawiliśmy już w ramach omówionego powyżej podejścia DiD. W tym miejscu zaprezentujemy na przykładzie tej samej ewaluacji bezpośredni skutek zastosowania techniki PSM. Jak zostało wspomniane, zadaniem metody PSM jest tzw. zbalansowanie rozkładów kontrolowanych zmiennych, tak aby porównywane grupy – interwencji oraz kontrolna – były w jak największym stopniu do siebie podobne. W tabeli 8.1 zamieszczone zostały estymowane średnie wartości zmiennych kontrolnych, uwzględnionych w modelu regresji logistycznej, na której podstawie oszacowana została wartość *propensity score*. Wartości średnie prezentowane są odpowied-

with caliper), metoda z promieniem (ang. *radius matching*) i metoda Kernel (ang. *Kernel matching*). Szerzej omawia to Trzciński (2009: 41).

nio dla: grupy beneficjentów; całej populacji podmiotów nieskutecznie ubiegających się o dotację (firmy, które złożyły wniosek o dofinansowanie do programu, ale w wyniku procesu selekcji ostatecznie nie otrzymały wsparcia); oraz grupy kontrolnej, wyselekcjonowanej (z wykorzystaniem techniki PSM) spośród całej populacji nieskutecznych wnioskodawców. Porównanie rozkładów zmiennych w grupie beneficjentów odpowiednio z całą populacją podmiotów, które nie otrzymały pomocy, a następnie z grupą kontrolną, pozwala odpowiedzieć na pytanie, czy w wyniku zastosowanej procedury udało się rzeczywiście uzyskać zbalansowane rozkłady zmiennych włączonych do modelu. Innymi słowy, chodzi o ustalenie, w jakim stopniu udało się zminimalizować pierwotne różnice występujące między jednostkami nieuczestniczącymi w interwencji a jej beneficjentami. Utrzymywanie się znaczących różnic wskazywałoby na konieczność ponownej estymacji modelu, np. pod kątem ponownego wyboru algorytmu łączenia jednostek w pary, czy nawet cofnięcia się do etapu szacowania *propensity score* (Caliendo i Kopeinig 2005: 16).

Oceny jakości doboru grupy kontrolnej można dokonywać zgodnie z propozycją przedstawioną przez Rosenbauma i Rubina. Polega ona na oszacowaniu, a następnie porównaniu dwóch miar. Pierwsza z nich to wystandaryzowana procentowa różnica między średnimi danej zmiennej predykcyjnej grupy uczestników programu i całej populacji wykluczonej ze wsparcia (Rubin 2006: 212):

$$SD_{przed} = 100(\bar{X}_1 - \bar{X}_{0P}) \Big/ \sqrt{(s_1^2 + s_{0P}^2)/2},$$

gdzie $\bar{X}_1$ i $\bar{X}_{0P}$ są średnimi analizowanej zmiennej X w grupie beneficjentów i w populacji nieskutecznych wnioskodawców, s_1^2 i s_{0P}^2 zaś identyfikują odpowiednie oszacowania wariancji dla obu tych grup. W drugim kroku należy oszacować analogiczną różnicę, tyle że przedmiotem porównania jest tu grupa beneficjentów z dobraną grupą kontrolną:

$$SD_{po} = 100(\bar{X}_1 - \bar{X}_{0K}) \Big/ \sqrt{(s_1^2 + s_{0P}^2)/2},$$

gdzie $\bar{X}_{0K}$ są średnimi wartościami analizowanej zmiennej X w grupie kontrolnej (pozostałe parametry zostały zdefiniowane w formule powyżej). Obie wartości zostały przedstawione w ostatnich dwóch kolumnach tabeli 8.1. Problemem, który trzeba arbitralnie rozstrzygnąć, jest ustalenie akceptowalnej wartości obciążenia kontrolowanej zmiennej. Zazwyczaj ocena dokonywana jest za pomocą „reguły kciuka", choć w literaturze przedmiotu można spotkać opinie, że zredukowanie obciążenia poniżej 3%–5% jest satysfakcjonujące (Caliendo i Kopeinig 2005: 15).

Tabela 8.1. Stopień zbalansowania zmiennych kontrolnych w wyniku zastosowania techniki *propensity score matching*. Średnie wartości i różnice między grupami

Zmienna	Grupa beneficjentów programu	Populacja nieskutecznych wnioskodawców	Grupa kontrolna	Standaryzowana różnica w %, przed doborem grupy kontrolnej (SD_{przed})	Standaryzowana różnica w %, po doborze grupy kontrolnej (SD_{po})
Przychód	50409,243	74320,048	62275,931	–4,74	–2,35
Udział podmiotów, które zanotowały wzrost przychodów względem wcześniejszego okresu sprawozdawczego	0,718	0,592	0,729	26,72	–2,35
Suma aktywów	32676,701	75478,179	30319,083	–2,78	0,15
Udział podmiotów, które zanotowały wzrost aktywów względem wcześniejszego okresu sprawozdawczego	0,696	0,601	0,698	20,12	–0,35
Wartość pomocy *de minimis*	13802,669	9826,166	11593,883	5,22	2,90
Wiek firmy w latach	10,315	10,407	10,126	–1,12	2,28
Zatrudnienie osób ogółem	70,344	44,771	71,154	47,44	–1,50
Udział podmiotów, które zanotowały wzrost zatrudnienia względem wcześniejszego okresu sprawozdawczego	0,657	0,565	0,659	18,88	–0,46
Przeciętna różnica zatrudnienia pomiędzy dwoma okresami sprawozdawczymi poprzedzającymi udział w programie	7,291	4,205	7,727	17,62	–2,49
Odsetek kobiet	0,270	0,333	0,268	–25,82	0,98
Liczba podpisanych umów w ramach programów realizowanych w ramach Phare	1,147	0,446	1,058	53,95	6,87

Łączna wartość środków wypłaconych w ramach Phare		15502,530	5301,365	13934,875	46,73	7,18
Udział podmiotów, które przed wnioskiem do działania 2.3 SPO WKP złożyły wniosek do działania 2.1 SPO WKP (usługi doradcze)		0,064	0,021	0,058	21,45	2,77
Udział podmiotów, które przed wnioskiem do działania 2.3 SPO WKP złożyły wniosek do działania 2.1 SPO WKP (usługi doradcze), a w efekcie podpisały umowy o dofinansowanie		0,038	0,008	0,034	20,24	2,60
Udział podmiotów, które przed udziałem w programie posiadały podpisaną umowę na działania szkoleniowe dla pracowników firmy w ramach działania 2.3 SPO RZL		0,083	0,044	0,089	15,92	–2,28
Udział podmiotów, które posiadały kredyt na realizację inwestycji		0,699	0,532	0,675	34,68	4,86
Województwo	Dolnośląskie	0,074	0,072	0,078	0,53	–1,71
	Kujawsko-pomorskie	0,069	0,053	0,075	6,61	–2,55
	Lubelskie	0,034	0,048	0,038	–7,01	–1,96
	Lubuskie	0,029	0,028	0,035	0,76	–3,32
	Łódzkie	0,081	0,068	0,081	4,85	0,00
	Małopolskie	0,078	0,087	0,068	–3,23	3,44
	Mazowieckie	0,138	0,122	0,149	4,81	–3,30
	Opolskie	0,028	0,023	0,024	3,11	2,12
	Podkarpackie	0,054	0,067	0,053	–5,64	0,23
	Podlaskie	0,039	0,033	0,037	3,35	1,48
	Pomorskie	0,056	0,055	0,059	0,65	–1,21
	Śląskie	0,107	0,133	0,106	–8,01	0,34

	Świętokrzyskie	0,025	0,028	0,022	–1,89	2,07
	Warmińsko-mazurskie	0,039	0,036	0,039	2,02	0,00
	Wielkopolskie	0,116	0,105	0,107	3,60	3,01
	Zachodniopomorskie	0,033	0,043	0,029	–5,16	2,04
Charakterystyka zakresu głównej działalności gospodarczej	Produkcja artykułów spożywczych i napojów	0,042	0,078	0,037	–14,96	2,11
	Produkcja drewna i wyrobów z drewna oraz z korka (z wyłączeniem mebli), wyrobów ze słomy i materiałów używanych do wyplatania	0,053	0,032	0,050	10,46	1,11
	Działalność wydawnicza; poligrafia i reprodukcja zapisanych nośników informacji	0,070	0,027	0,067	20,37	1,56
	Produkcja wyrobów gumowych i z tworzyw sztucznych	0,118	0,045	0,117	26,81	0,20
	Produkcja metalowych wyrobów gotowych, z wyłączeniem maszyn i urządzeń	0,130	0,061	0,128	23,76	0,76
	Produkcja maszyn i urządzeń, gdzie indziej niesklasyfikowana	0,069	0,031	0,066	17,64	1,53
	Produkcja mebli; działalność produkcyjna, gdzie indziej niesklasyfikowana	0,069	0,034	0,067	15,66	1,01
	Przetwórstwo przemysłowe (z wyłączeniem działów 15, 20, 22, 25, 28, 29, 36)	0,228	0,150	0,238	20,04	–2,42
	Budownictwo	0,064	0,122	0,070	–19,79	–2,11
	Handel hurtowy i detaliczny; naprawa pojazdów samochodowych, motocykli oraz artykułów użytku osobistego i domowego	0,070	0,185	0,074	–35,10	–1,35

	Obsługa nieruchomości, wynajem i usługi związane z prowadzeniem działalności gospodarczej	0,036	0,049	0,029	–6,77	3,31
	Inna działalność	0,050	0,187	0,056	–43,06	–1,76
Obszar realizacji projektu	Miejski	0,641	0,721	0,650	–17,33	–1,91
	Wiejski	0,349	0,261	0,340	19,06	1,82
	Brak danych	0,011	0,017	0,010	–5,86	0,47

Źródło: opracowanie własne na podstawie badania PARP przeprowadzonego w 2011 r.

Z zestawienia wielkości przedstawionych w dwóch ostatnich kolumnach tabeli wynika, że zastosowana procedura doboru grupy kontrolnej pozwoliła znacząco zminimalizować pierwotne różnice występujące między populacją jednostek, które nie otrzymały wsparcia, a beneficjentami programu dotacji dla przedsiębiorstw. Większość różnic między beneficjentami a dobraną grupą kontrolną nie przekracza 3%, mimo że w punkcie wyjścia obie populacje istotnie się między sobą różniły – z punktu widzenia obserwowalnych charakterystyk były nieporównywalne ze sobą. Największe redukcje różnic wystąpiły w odniesieniu do liczby zatrudnionych osób, udziału w innych programach wsparcia dla przedsiębiorstw (Phare, SPO RZL[10] itd.) i przynależności rozpatrywanych podmiotów do branży gospodarki.

Wspomnijmy o ograniczeniach techniki PSM. Wynikają one z jej założeń. Krytycznym elementem jest założenie o warunkowej niezależności (CIA). Założenie to jest nietestowalne w praktyce badawczej, a jedyny wyjątek stanowi sytuacja, w której istnieje możliwość porównania badań quasi-eksperymentalnych z wynikami analiz wykorzystujących dane eksperymentalne (np. Heckman i in. 1997: 606; Dehejia i Wahba 2002: 151). Niespełnienie założenia o warunkowej niezależności grozi błędnym oszacowaniem efektu przyczynowego, stąd też posługując się PSM, należy każdorazowo rozważyć, jak dalece kontrolujemy wszystkie istotne cechy determinujące partycypację w ramach ocenianego programu i jego efekty. Technika PSM eliminuje obciążenie estymowanych skutków interwencji wynikające z tego, że grupa interwencyjna różni się w zakresie obserwowanych charakterystyk od grupy kontrolnej. Jednak problematyczne pozostaje niekontrolowane obciążenie wynikające z różnic dla zmiennych nieobserwowalnych. W przeciwieństwie do metody eksperymentalnej, która (przynajmniej w punkcie wyjścia) zapewnia brak różnic w zakresie wszystkich charakterystyk – zarówno tych obserwowalnych, jak i nieobserwowalnych – technika PSM pozwala na upodobnienie grup jedynie w zakresie zmiennych, które zostały uwzględnione w modelu ewaluacji[11]. Ma to fundamentalne znaczenie dla trafności wyników analiz realizowanych w badaniach quasi-eksperymentalnych.

Potencjalne problemy z właściwą identyfikacją modelu można zobrazować na przykładzie ewaluacji skuteczności instrumentów wsparcia dla osób bezrobotnych. Badanie dotyczyło jednego z programów finansowanego z przedakcesyjnego funduszu Phare 2002 Spójność Społeczno-Gospodarcza: Rozwój Zasobów Ludzkich, realizowanego w latach 2004–2005[12]. W programie wspierane były osoby bezrobotne chcące rozpocząć własną działalność gospodarczą. Zakres wsparcia dla uczestników projektu obejmował szkolenia z zakresu przedsiębiorczości na różnym poziomie zaawansowania oraz różne formy doradztwa indywidualnego. Jednym z wymiarów, na którym skupiała się ewaluacja, była skuteczność instrumentu, tj. analiza, w jakim stopniu program miał wpływ na to, że jego beneficjenci faktycznie rozpoczęli własną działalność gospodarczą. W ewa-

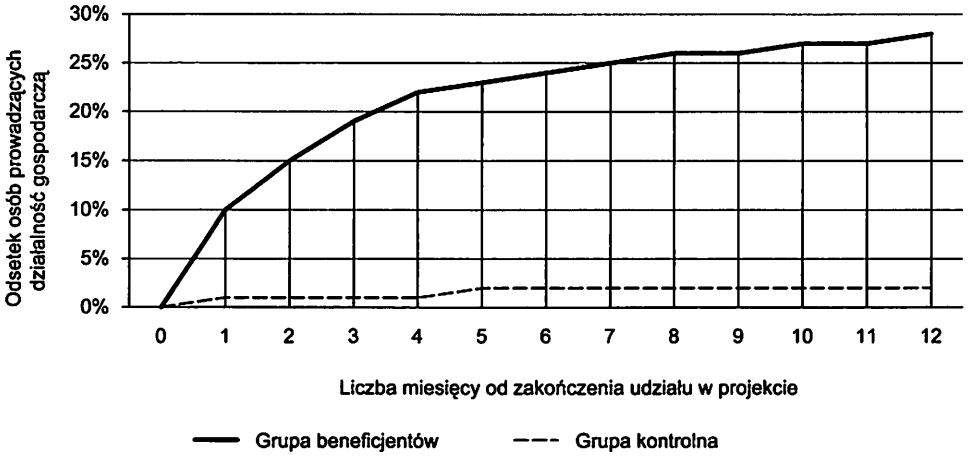
[10] Sektorowy Program Operacyjny: Rozwój Zasobów Ludzkich.

[11] W przypadku PSM zmienne nieobserwowalne mogą być uwzględnione pośrednio, o ile tylko pozostają one w związku z cechami podlegającymi kontroli.

[12] Ewaluacja zrealizowana na zlecenie Polskiej Agencji Rozwoju Przedsiębiorczości w 2006 r.

luacji wykorzystane zostały gromadzone przez powiatowe urzędy pracy dane dotyczące rejestracji bezrobotnych, udzielania wsparcia, monitorowania itp.[13]

Dokonując ewaluacji, podjęto decyzję, że grupa kontrolna zostanie dobrana z pozostałej grupy osób zarejestrowanych jako bezrobotne. Oczywiście porównanie grupy beneficjentów z szerszą grupą bezrobotnych obarczone byłoby dużym błędem pomiaru. Z tego powodu zdecydowano się na dobór wyselekcjonowanej grupy kontrolnej z wykorzystaniem techniki PSM. Chcąc wyeliminować obciążenie selekcyjne, wytypowano 22 zmienne kontrolne obejmujące m.in.: cechy społeczno-gospodarcze; cechy związane z zatrudnieniem, aktywnością zawodową i rodzajem tej aktywności, cechy związane z wcześniejszym podnoszeniem kwalifikacji i aktywnością szkoleniową, cechy odnoszące się do względnej motywacji osób do poszukiwania pracy oraz cechy dotyczące posiadanych umiejętności (Trzciński 2009: 52). Zmienne te pokrywały więc szerokie spektrum charakterystyk mogących wpływać na udział w programie i na aktywność zawodową bezrobotnych. W wyniku zastosowanych procedur udało się otrzymać dwie grupy, nieróżniące się istotnie w zakresie wymienionych cech. Oszacowanie efektu przyczynowego (efektu interwencji) przedstawione jest na rysunku 8.2.



Rysunek 8.2. Wpływ wsparcia szkoleniowego na przedsiębiorczość osób bezrobotnych – przykład obciążonego efektu interwencji z wykorzystaniem techniki PSM

Źródło: opracowanie własne na podstawie raportu z badań: PBS DGA (2006) *Raport z ewaluacji* ex-post *komponentu regionalnego programu Phare 2002 Spójność Społeczna i Gospodarcza – Komponent Rozwój Zasobów Ludzkich.*

Łatwo zauważyć, że istnieje istotna różnica między wynikami uzyskanymi dla jednostek w grupie beneficjentów i w grupie kontrolnej. Niemal 30% uczestników projektu w ciągu 12 miesięcy od jego zakończenia prowadziło własną działalność gospodarczą.

[13] Na potrzeby ewaluacji korzystano z Systemu Informatycznego PULS, zawierającego informacje o wszystkich bezrobotnych, w tym tych, którzy zdecydowali się na udział w programie Phare.

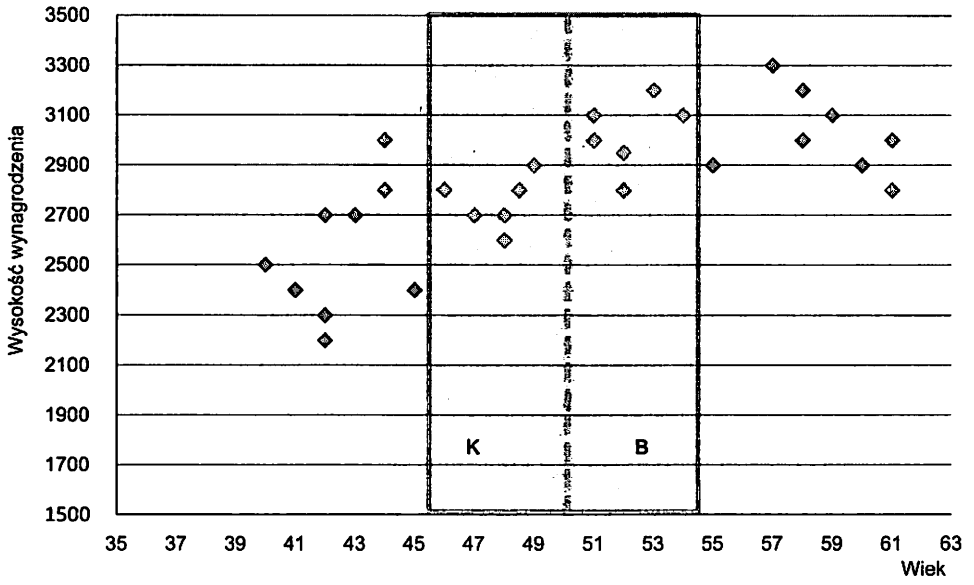
Równocześnie na podobny krok zdecydowała się tylko co dwudziesta osoba z grupy kontrolnej (ok. 2%). Jednak do wniosków na temat wielkości potencjalnego efektu należy podchodzić bardzo ostrożnie. Obserwowane wyniki mogą – z jednej strony – świadczyć o tym, że rozpoczęta przez beneficjentów działalność gospodarcza była w przeważającej mierze wyłącznym efektem oddziaływania ocenianego programu. Z drugiej strony pozytywny wynik testu może być rezultatem procedury doboru grupy kontrolnej, która nie wyeliminowała obciążenia selekcyjnego. Potwierdzeniem tego przypuszczenia mogą być wyniki zrealizowanego równolegle badania ankietowego, w którym ponad połowa beneficjentów prowadzących działalność gospodarczą przyznała, że założyłaby firmę bez względu na udział w programie (PARP 2006: 112). Trzeba oczywiście pamiętać o wszelkich ograniczeniach badania wpływu interwencji metodą wywiadu kwestionariuszowego – w tym przede wszystkim o deklaratywności odpowiedzi respondentów – niemniej wydaje się raczej mało prawdopodobne, żeby oceniany tu program aż w takim stopniu przyczynił się do aktywizacji beneficjentów. Należy podkreślić, że oferowane w nim wsparcie sprowadzało się jedynie do szkoleń oraz doradztwa w zakresie przedsiębiorczości. Brakowało w nim natomiast takiego instrumentu wsparcia jak dotacje, pożyczki itp.

Możliwym wyjaśnieniem różnicy między grupą eksperymentalną i grupą kontrolną było nieuwzględnienie istotnych informacji w ramach rozpatrywanego modelu. Dotyczyłoby to głównie cech związanych z motywacją do podjęcia własnej działalności gospodarczej. Informacje pozyskane od osób odpowiedzialnych za realizację interwencji wskazują, że do programu przyjmowane były osoby, które zadeklarowały chęć założenia działalności gospodarczej. Tak więc decyzja o założeniu własnej firmy podejmowana była zapewne jeszcze przed przystąpieniem do projektu. Phare miało stanowić jedynie element wspomagający. Otóż informacji tych brakowało w bazie danych, z której korzystano w badaniu. Można domniemywać, że pominięcie tej informacji zaburzyło szacunki efektu projektu. Jeżeli mimo zrównania porównywanych grup ze względu na obserwowalne zmienne założenie o warunkowej niezależności nie zostało spełnione, prezentowane oszacowania efektu przyczynowego interwencji należałoby uznać za błędne i tym samym bezużyteczne z punktu widzenia oceny zastosowanego instrumentu wsparcia.

W praktyce nie ma gotowego przepisu na sposób doboru zmiennych do modelu. Kluczowym czynnikiem będzie oczywiście specyfika danej interwencji i kontekst jej realizacji, który powinien być w miarę możliwości odtworzony, zanim podejmie się działania zmierzające do doboru grupy kontrolnej. Ogólna rekomendacja w tym zakresie mówi, że dobór odpowiednich zmiennych kontrolnych powinien opierać się na uznanych teoriach społecznych, ekonomicznych, doświadczeniach oraz wiedzy wynikającej z wcześniejszych badań przeprowadzonych w zakresie danego typu interwencji. Przydatne mogą okazać się również pozyskane od samych potencjalnych odbiorców wsparcia informacje na temat ich indywidualnych motywów uczestnictwa lub jego braku. Informacje od osób zarządzających lub administrujących daną interwencją również mogą być tu pomocne (Bryson i in. 2002: 14).

8.2.3. Technika nieciągłości w równaniu regresji

Zgoła odmienną procedurę doboru grupy kontrolnej przyjmuje się w technice nieciągłości w równaniu regresji (ang. *regression discontinuity design* – RD). O ile w przypadku PSM badacz stara się wytypować zmienne, które wpływają na uczestnictwo w określonej interwencji i na jej efekt, o tyle problemem pozostaje dobór zmiennych „odpowiedzialnych" za selekcję do grupy interwencyjnej i grupy kontrolnej. W podejściu RD zmienna, bądź zmienne, decydująca o tym, kto otrzyma wsparcie, jest znana. Najczęściej powiązane są one z odgórnie narzuconymi warunkami selekcji do danej interwencji. Będzie tak w każdej sytuacji, w której przewidziany jest pewien stały próg graniczny (tzw. punkt odcięcia – *cut-off point*); dzieli on populację na jednostki, które mogą zostać objęte danym programem, oraz jednostki, które są z niego wykluczone. Próg graniczny może być definiowany na wartościach jednej lub kilku znanych zmiennych. Przykładem takiego progu może być np. określony wiek, liczba pracowników w przedsiębiorstwie, wynik przeprowadzonego testu – wówczas kryteria dostępu do danej interwencji ustala się dla określonej wartości zmiennej(-ych). Konkretnym przykładem są programy określane mianem 50+, które oferują wsparcie szkoleniowe – np. w postaci kursów komputerowych, językowych itd. – osobom powyżej 50. roku życia[14]. Osoby młodsze nie mogą ubiegać się o udział w programie. Korzystając z tego przykładu, ogólną intuicję stojącą za metodą RD można przedstawić na rysunku 8.3.



Rysunek 8.3. Idea nieciągłości w równaniu regresji

Źródło: opracowanie własne.

[14] Por. http://www.efs.gov.pl/Wiadomosci/Documents/prezentacja_EFS_dla_osob_50_plus.pdf – stan na 29 maja 2012 r.

Na rysunku 8.3 przedstawiony jest rozkład wielkości wynagrodzenia względem wieku. Powyżej linii 50 lat znajdują się zarobki beneficjentów programu, poniżej – osób nieuprawnionych do skorzystania ze wsparcia. Przyjmuje się, że oszacowania efektu interwencji można dokonać przez porównanie marginalnej podgrupy beneficjentów, znajdujących się w pobliżu progu granicznego, i podgrupy jednostek wykluczonych ze wsparcia, znajdujących się po drugiej stronie progu, zakładając, iż takie niewielkie różnice dla zmiennej „odpowiedzialnej" za selekcję nie powinny mieć wpływu na zmienną identyfikującą „efekt" programu. Tak więc oszacowanie efektu sprowadzałoby się do porównania osób znajdujących się tuż na prawo od przerywanej linii, „wpadających" w wyróżniony na szaro obszar (B), z osobami, które znajdują się tuż po lewej stronie przerywanej linii (K) – osoby te należą do grupy interwencyjnej i grupy kontrolnej, ale pod względem zmiennej selekcyjnej są bardzo podobne.

Widać, że istotnym ograniczeniem tego podejścia jest możliwość oszacowania efektu przyczynowego jedynie lokalnie, a więc w odniesieniu do przyjętego obszaru podobieństwa. W powyższym przykładzie można byłoby więc mówić o pozytywnym wpływie programu wsparcia na wysokość wynagrodzenia osób, z zaznaczeniem, że wniosek ten odnosi się jedynie do populacji osób w wieku 50–55 lat. Drugą praktyczną komplikacją jest możliwość uśredniania wyników dla stosunkowo małych grup. Dążąc bowiem do maksymalnego upodobnienia grupy beneficjentów i grupy kontrolnej pod względem wartości zmiennej selekcyjnej, rozsądnie byłoby tak określić jej górną i dolną wartość, aby znacząco nie odbiegała od ustalonego progu granicznego. Jednak w typowej sytuacji, im bardziej obie populacje będą do siebie podobne (tj. im bliższe progu granicznego będą dopuszczalne różnice w wartościach zmiennej selekcyjnej), tym mniejsza liczebnie będzie grupa beneficjentów i grupa kontrolna. Problem ten jest oczywiście pochodną wielkości całej populacji, a szerokością obszaru obejmującego jednostki można oczywiście dowolnie manipulować. Jednak zawężanie zakresu różnic między porównywanymi grupami na zmiennej selekcyjnej, a tym samym zmniejszanie liczebności obu grup, może rodzić problemy z uzyskaniem odpowiedniej mocy testów statystycznych stosowanych do wykrycia efektów danej interwencji. Dlatego aby móc generalizować wyniki na całą grupę beneficjentów, niezbędne są dodatkowe założenia w odniesieniu do relacji między zmienną selekcyjną a wynikami, co umożliwia zastosowanie jakiegoś modelu regresji.

Pierwsze zastosowanie i omówienie metody RD przedstawili Thistlethwaite i Campbell w pracy z 1960 r. (Hahn i in. 2001: 201). Technika ta znów zyskuje na popularności, szczególnie w ostatnich latach, w kontekście rozwoju podejść quasi-eksperymentalnych, stosowanych w ewaluacji programów publicznych. Jedną z przyczyn takiego stanu rzeczy jest pokaźna liczba interwencji, w których mamy do czynienia z arbitralnie przyjętym punktem selekcji, decydującym o tym, kto może zostać beneficjentem danego programu wsparcia. Bardzo często jest nim wspominany wiek, ale również takie cechy jak dochód, wielkość przedsiębiorstw czy region.

8.3. Ustalenie efektu przyczynowego w świetle procesu ewaluacji interwencji

Zaprezentowane powyżej metody pozwalają na dokonanie oszacowania efektu przyczynowego wybranych interwencji publicznych. Abstrahując w tym miejscu od wiarygodności dokonanych oszacowań – te bowiem powinny być każdorazowo przedmiotem osobnej, krytycznej analizy – należy zwrócić uwagę na kwestię wykorzystania otrzymanych wyników w całościowym procesie ewaluacji. Mając oszacowany efekt przyczynowy, jesteśmy w stanie powiedzieć, w jakim stopniu dany program okazał się skuteczny, a więc w jakim stopniu miał wpływ na osiągnięcie założonych celów. Dokonując analizy kosztów i korzyści w zakresie wydatkowanych środków, można byłoby również udzielić odpowiedzi na pytanie o relację między uzyskanymi efektami a poniesionymi kosztami. Czy jednak na podstawie danych informujących o wielkości efektów jesteśmy w stanie odpowiedzieć na pytania, dlaczego osiągnięto taki a nie inny efekt – czy mógłby być on większy, a jeśli tak, to kiedy, dla kogo i w jakich okolicznościach? Niestety badania skupiające się tylko na pomiarze efektów przyczynowych pozostawiają powyższe pytania bez odpowiedzi. Natomiast jeśli ewaluacje mają służyć poprawie projektowania i realizacji polityk publicznych, to oprócz informacji o tym, czy zastosowane instrumenty wsparcia zadziałały, muszą one wskazywać również, dlaczego tak się stało. Oczywiście odpowiedź na pytanie o to, czy zastosowane dotychczas instrumenty się sprawdziły, poprzedza udzielenie odpowiedzi na pytanie o przyczyny takiego stanu rzeczy. Jednak jeśli chcemy traktować ewaluację jako użyteczne źródło wiedzy w procesie ulepszania polityk publicznych, konieczne jest uzyskanie odpowiedzi na oba pytania łącznie. Idea ta jest coraz intensywniej propagowana pod hasłem konieczności realizacji tzw. ewaluacji oddziaływania opartej na teorii (ang. *Theory-based impact evaluation* – TBIE), w której powinno się podejmować próby odpowiedzi właśnie na pytania: dlaczego dany program wygenerował określone efekty? Jakie warunki musiałyby zaistnieć, aby były one inne (większe)? Kto najbardziej skorzystał na uczestnictwie w danym programie i dlaczego? Co należałoby zmienić, aby uczynić instrument bardziej wydajnym itp.? Które zakładane mechanizmy przyczynowe zadziałały w rzeczywistości, a które nie?

Chcąc ująć ten problem szerzej, należy zauważyć, że każda interwencja publiczna – program, projekt itd. – jest swoistą teorią zmiany. Decydując się na jej realizację, przyjmuje się liczne założenia o tym, że w danych warunkach podjęcie określonych działań, z wykorzystaniem takich, a nie innych zasobów, przyczyni się do rozwiązania pewnych zdiagnozowanych problemów. W tym kontekście „rolą ewaluacji jest sprawdzenie, czy w świetle uzyskanych efektów dana teoria programu daje się utrzymać, czy też należy ją odrzucić" (Górniak 2007: 13). Ewaluację można więc postrzegać jako proces weryfikacji teorii ocenianych programów. Teorię programu można najogólniej określić jako zbiór założeń na temat procesu przechodzenia od konkretnych działań do oczekiwanych efektów. Przyjmowane założenia mają zwykle charakter hierarchiczny, do tego często są ze sobą powiązane i ostatecznie układają się w ciąg przyczynowo-skutkowy. Przykłady takich założeń dla programów rynku pracy polegających na wsparciu szkoleniowym

osób bezrobotnych można byłoby sformułować w następujący sposób: (1) problem ze znalezieniem pracy wynika z braku odpowiednich kompetencji w grupie osób bezrobotnych, (2) dzięki organizacji szkoleń jesteśmy w stanie wyposażyć osoby bezrobotne w wystarczające kompetencje potrzebne do znalezienia pracy, (3) tematyka zorganizowanych szkoleń odpowie na rzeczywiste potrzeby rynku pracy, (4) osoby rekrutowane do programu nie będą posiadały odpowiednich kompetencji do znalezienia pracy, itd.

Wielu założeń nie podaje się wprost w dokumentach programowych, znaczna część z nich przyjmowan jest *implicite* (Leeuw 2003: 6). Dlatego na etapie ewaluacji często konieczne jest dokonanie rekonstrukcji teorii programów – tj. odtworzenia kompletnej logiki stojącej za daną interwencją. Proces ten można typowo zobrazować za pomocą matrycy logicznej programu. W wierszach matrycy umieszczone są hierarchicznie kolejne poziomy oczekiwanych zmian uzyskanych w wyniku realizacji programu. Kolumny wskazują kolejno na: obszary zamierzonych zmian wyrażone najczęściej w określonych wskaźnikach; źródła, dzięki którym możliwe będzie uzyskanie informacji, czy oczekiwana zmiana faktycznie zaszła; założenia, które warunkują uzyskanie określonych zmian (bądź efektów). Uzyskanie zmian na każdym z poziomów pozwala na osiągnięcie zmian na poziomie wyższym (Kierzkowski 2002: 57). W ogólnym ujęciu odpowiednie zasoby (wkład – finansowy, ludzki itp.) pozwalają na uzyskanie określonych produktów (np. zorganizowanie szkoleń dla bezrobotnych) oraz rezultatów (np. przeszkolenie osób bezrobotnych), które przy spełnieniu właściwych założeń mogą przeistoczyć się w odpowiednie efekty (np. spadek stopy bezrobocia w grupie objętej wsparciem), będące faktycznym celem danej interwencji.

Tabela 8.2. Schemat matrycy logicznej

Cel	Wskaźniki osiągnięć	Źródła informacji	Założenia
Efekty			
Rezultaty			
Produkty			
Wkład			

Źródło: opracowano na podstawie: Kierzkowski 2002.

Istotnym elementem ewaluacji, w drodze do zrozumienia oddziaływania programów publicznych, a w dalszej kolejności także możliwości generalizacji wyników, jest zrozumienie kontekstu realizacji analizowanej interwencji (White 2009: 10). Zadanie to wymaga dokładnego poznania motywów realizacji danego programu oraz jego otoczenia, przez które należy rozumieć kontekst społeczny, ekonomiczny, ale i kulturowy. Prace wykonane w tym zakresie pozwolą z jednej strony na właściwe zdiagnozowanie potencjalnych efektów programu, w tym na wskazanie wymiarów, w których uzyskane wyniki mogą się różnicować. Z drugiej strony rzeczywiste poznanie kontekstu pozwoli zbadać, w jakim zakresie i w jakich okolicznościach uzyskane efekty programu są replikowalne,

tzn. pozwoli ustalić warunki brzegowe dla możliwości odtworzenia efektów analogicznej interwencji w innym miejscu i czasie. Powyższe ustalenie ma kluczowe znaczenie z punktu widzenia możliwości wykorzystania wyników ewaluacji w praktyce.

Realizacja ewaluacji oddziaływania opartych na teorii wymaga zastosowania całego spektrum podejść badawczych. W odpowiedzi na pytanie o to, dlaczego uzyskano w wyniku realizacji interwencji dane efekty, nieocenione okażą się wszelkiego rodzaju techniki jakościowe. TBIE jest więc w praktyce zadaniem złożonym, wykraczającym daleko poza ilościowy pomiar oddziaływania interwencji. Można je scharakteryzować jako przedsięwzięcie, które jednoczy w jednym procesie badawczym zadanie ustalania efektów (*effects of causes*) i ich przyczyn (*causes of effects*) – (Holland 1986: 945). W tym miejscu warto za Howardem White'em – jednym z orędowników tego podejścia – podsumować tę część, wskazując na główne składowe procesu ewaluacji oddziaływania opartej na teorii (White 2009: 7). Składają się na niego odpowiednio:

1) odtworzenie logiki interwencji (teorii programu),
2) zrozumienie kontekstu, w jakim realizowana jest interwencja,
3) identyfikacja obszarów, w których mogą różnicować się efekty programu wsparcia,
4) dokonanie pomiaru efektów z wykorzystaniem wiarygodnego stanu kontrfaktycznego,
5) analiza faktyczna – tego, co faktycznie się wydarzyło w toku realizacji interwencji,
6) korzystanie z wielu metod, w tym jakościowych, adekwatnych do danego problemu badawczego (*mixed methods*).

8.4. Podsumowanie

Powyżej przedstawiono rozwiązania metodologiczne służące pomiarowi efektów przyczynowych interwencji publicznych. Nurt badań typu *impact evaluation*, w których stosuje się opisane podejścia, jest obecny w Polsce od kilku lat. Liczba przykładów ewaluacji stosujących metody oparte na podejściu kontrfaktycznym stale rośnie i można ostrożnie przewidywać, że trend ten utrzyma się również w najbliższym czasie. Sprzyja temu szczególnie realizacja programów finansowanych ze środków unijnych. Nie bez znaczenia pozostają tu również działania podejmowane przez Komisję Europejską, zmierzające do popularyzacji rygorystycznych podejść metodologicznych w grupie krajów członkowskich. Godny polecenia jest tu chociażby portal poświęcony ewaluacji programów unijnych z wykorzystaniem metod bazujących na podejściu kontrfaktycznym[15]. Obserwowany trend należy oczywiście oceniać pozytywnie. Wiedza uzyskana w badaniach nad wpływem programów publicznych ma szansę przełożyć się na skokowy wzrost jakości realizowanych polityk publicznych, wpisując się w nurt kształtowania polityk opartych na właściwych dowodach (*evidence-based policy*). Kluczem do sukcesu w tym zakresie pozostaje jednak umiejętne wykorzystanie opisywanych technik w ramach szerszego procesu ewaluacji opartej na teorii. Niestety w tym momencie trudno

[15] Por.: http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/sourcebooks/method_techniques/counterfactual_impact_evaluation/index_en.htm – stan na 29 maja 2012 r.

o przykłady tego typu badań. Należy mieć jednak nadzieję, że sytuacja w tej dziedzinie będzie również systematycznie ulegać poprawie[16].

Na koniec warto zauważyć, że potencjalne zastosowanie opisywanych w niniejszym rozdziale podejść metodologicznych nie musi zawężać się wyłącznie do badań ewaluacyjnych. Zaprezentowane techniki mają w dużej mierze charakter uniwersalny i z powodzeniem mogą być stosowane również w innych niż polityki publiczne dziedzinach analizy. Ogólna rekomendacja wskazywałaby na możliwość ich wykorzystania wszędzie tam, gdzie poszukiwany jest efekt pewnego zdarzenia, do którego selekcja uczestników nie ma charakteru losowego. Na przykład mogą to być takie zdarzenia jak reklama, kampania społeczna/wyborcza/informacyjna itp. Warto również zwrócić uwagę na potencjał drzemiący w technice PSM w odniesieniu do badań realizowanych w internecie, w których możliwość uogólniania wyników na szersze populacje jest bardzo ograniczona z uwagi na nielosowy dobór grupy badawczej.

Poważne ograniczenie wykorzystania omawianych tu podejść stanowi dostępność właściwych danych, które pozwolą na spełnienie przyjmowanych założeń. Niemniej jeśli tylko wymagane dane będą dostępne, to zastosowanie technik może znacząco usprawnić wnioskowanie w zakresie identyfikacji relacji przyczynowo-skutkowych.

[16] Jednym z nielicznych wyjątków jest tu ekspertyza zrealizowana przez Seweryna Krupnika (2008).