

Przyczynowość, selekcja i endogeniczne oddziaływanie

Paweł Strawiński
Uniwersytet Warszawski
Wydział Nauk Ekonomicznych
pstrawinski@wne.uw.edu.pl

Kwiecień 2007

Streszczenie

W modelach konstruowanych dla pomiaru efektów oddziaływania, lub uwzględniających problem nielosowej selekcji, czynniki nieobserwowane mogą wpływać jednocześnie na wartość zmiennej wynikowej oraz prawdopodobieństwo poddania oddziaływaniu. Ponadto, modele mogą zawierać zmienne endogeniczne. W artykule zdefiniowano pojęcie oddziaływania o endogenicznym charakterze, oraz pokazano, że model z endogenicznym oddziaływaniem oraz model uwzględniający nielosowy mechanizm doboru obserwacji do próby są wariantami wspólnego modelu strukturalnego. Przedstawiono sposób wykorzystania metody funkcji kontrolujących do zgodnego oszacowania parametrów modelu z endogenicznym oddziaływaniem.

Słowa kluczowe: Przyczynowość, Selekcja, Egzogeniczność, Oddziaływanie, Program.

1 Wprowadzenie

Jednym z podstawowych celów prowadzonych badań empirycznych jest wyizolowanie zależności przyczynowo-skutkowych pomiędzy zachodzącymi zjawiskami. W naukach społecznych takich jak socjologia, czy psychologia, oraz przyrodniczych często stosowanym podczas analizy tego zagadnienia narzędziem są kontrolowane eksperymenty. Pozwalają one, w warunkach laboratoryjnych, analizować różne możliwe scenariusze przebiegu zjawiska. Przeprowadzając eksperyment badacze są w stanie kontrolować wszystkie parametry.

W empirycznych badaniach podstawowych dziedzin ekonomicznych metody eksperymentalne są stosunkowo rzadko używane. Dzieje się tak ze względu na występowanie ograniczeń technicznych, społecznych i etycznych. Trudno jest bowiem oddzielić część gospodarki w sposób zapewniający spełnienie warunków eksperymentalnych, bowiem funkcjonuje ona jako całość. Autarkia w czystej postaci w rzeczywistym świecie nie występuje. Nawet gdyby takie odizolowanie było technicznie możliwe, pozostaje pytanie natury etycznej i społecznej, czy wolno dla celów nauki eksperymentować na ludziach. Pomijamy tu kwestię wiarygodności uzyskanych rezultatów. Osoby uczestniczące w eksperymencie wiedząc, że wynik nie ma znaczenia dla ich sytuacji ekonomicznej, mogłyby zachowywać się inaczej niż w warunkach rzeczywistych. Z tego powodu podstawowym typem badania empirycznego w naukach ekonomicznych jest analiza danych ilościowych o charakterze przekrojowym. Z uwagi na ograniczoną możliwość prowadzenia badań eksperymentalnych często nadaje się jej charakter badania obserwacyjnego (ang. *observational study* [2]). Celem badania obserwacyjnego jest wyabstrahowanie związku przyczynowo-skutkowego, w sytuacji gdy nie można zastosować kontrolowanego eksperymentu [12]. Badanie obserwacyjne dotyczy programów, interwencji, czy analizy polityk oddziałujących na otoczenie. Polega ono na porównaniu zachowań jednostek, które są objęte programem czy interwencją z jednostkami, które pozostają poza bezpośrednim oddziaływaniem. Rezultatem podjęcia takich działań jest powstanie naturalnego eksperymentu. Podstawową cechą odróżniającą go od eksperymentu przeprowadzonego w warunkach laboratoryjnych, jest spontaniczne a nie zamierzone jego powstanie.

Z uwagi na spontaniczny charakter naturalnych eksperymentów, badacz nie jest w stanie kontrolować wszystkich parametrów eksperymentu i zweryfikować wszystkich możliwych scenariuszy. Ponadto, jednostki uczestniczące w eksperymencie nie są odizolowane od społeczeństwa i gospodarki, a pozostają jego częścią. W związku z tym narzędzia analityczne służące do analizy danych zostały zmodyfikowane, w sposób umożliwiający uogólnienie wniosków.

Dalsza część artykułu jest zorganizowana jak następuje. W części drugiej przedstawiono problem identyfikacji zależności przyczynowo-skutkowych w badaniach ekonomicznych. W kolejnej części opisano źródła problemu selekcji. Część czwarta obejmuje opis model strukturalny, którego wariantami są model uwzględniający selekcję zaproponowany przez Heckmana, oraz model z endogenicznym oddziaływaniem. Skoncentrowano swoją uwagę na analogiach i istotnych różnicach między tymi wariantami. W części piątej przedstawiona jest metoda funkcji kontrolującej, pozwalająca uzyskać zgodne oszacowania dla parametrów modelu strukturalnego. W zakończeniu zawarte są wnioski i podsumowanie.

2 Przyczynowość

Podstawowe znaczenie dla prawidłowości modelowania ma rozróżnienie na czynniki sprawcze (przyczyny) i efekty ich wystąpienia (skutki). Niejednokrotnie takie rozgraniczenie jest arbitralne. Problem ten głównie dotyczy badań makroekonomicznych, bowiem wiele zjawisk zachodzących na poziomie gospodarki współwystępuje i trudno jest określić pierwotną przyczynę. Badania prowadzone w skali mikro na poziomie jednostkowym też go doświadczają. Na przykład, w badaniach zachowań na rynku pracy i płac przyjmuje się, że status na rynku pracy (zatrudniony, nie zatrudniony) determinuje poziom zarobków. Może jednak relacja przyczynowa jest odwrotna, osoby o niskich dochodach nie chcą pracować, bowiem oczekują, ich ewentualne dochody z pracy nie będą dużo wyższe od obecnych?

Relacja przyczynowo-skutkowe współczesnej statystyce i ekonometrii jest przedstawiana za pomocą notacji używającej wyników hipotetycznych (ang. *counterfactual*), które są sprzeczne z faktami, i nie mogą być obserwowane w rzeczywistości.

Ten sposób analizy przyczynowości wprowadził Neyman [11]. Współczesna wersja jest często w literaturze przypisywana Rubinowi [15].

Głównym elementem modelu jest zmienna losowa opisująca dwa potencjalnie zachodzące stany nazywane oddziaływaniem (ang. *treatment*) i brakiem oddziaływania (ang. *non-treatment*), i zapisywane $d \in \{0, 1\}$. Każdemu stanowi w którym jednostka się znajduje przypisywana jest odpowiednia wartość zmiennej wynikowej (Y_0, Y_1). Badacz obserwuje tylko jedną wartość wyniku Y . Zależność między obserwowanym wynikiem, a wynikami hipotetycznymi przedstawia się następująco

$$Y = dY_1 + (1 - d)Y_0 \quad (1)$$

Taką sytuację można zilustrować klasycznym przykładem modelu płacy. Dla osób pracujących $d = 1$ dochody wynoszą Y . Oczywiście $Y = Y_1$. Gdyby jednak, ta osoba niepracowała to jej hipotetyczne dochody wyniosłyby Y_0 . Problemem komplikującym model jest możliwość, że zarówno Y jak i d zależą od wspólnych (ang. *confounders*), nieobserwowanych czynników, takich jak indywidualne zdolności, oraz obserwowanych np. poziom wykształcenia.

Wartość zmiennej wynikowej Y jest modelowana jako warunkowa wartość oczekiwana, przy znanej realizacji wektora obserwowanych charakterystyk X . Oczywiście stan poddania oddziaływaniu d może być traktowany jako charakterystyka determinująca wartość zmiennej Y . Zmienne zawarte w macierzy X w literaturze anglojęzycznej określa się mianem zmiennych kontrolnych. Ma to swoje uzasadnienie przez analogię do badań eksperymentalnych. Przez porównanie wartości zmiennej wynikowej przy różnych wartościach wektora obserwowanych charakterystyk odtwarzany jest eksperyment laboratoryjny.

Tym co interesuje badaczy to pomiar przeciętnego efektu oddziaływania na jednostki poddane oddziaływaniu (ang. *average treatment on treated*)

$$ATT = E(Y_1 \mid d = 1, X) - E(Y_0 \mid d = 1, X) \quad (2)$$

i na jednostki nie poddane oddziaływaniu (ang. *average treatment on non-treated*)

$$ATN = E(Y_1 \mid d = 0, X) - E(Y_0 \mid d = 0, X) \quad (3)$$

oraz przeciętny efekt oddziaływania (ang. *average treatment effect*)

$$ATE = E(Y_1 | X) - E(Y_0 | X) \quad (4)$$

Zgodnie z propozycją Rosembauma i Rubina [13] przyjmuje się, że oddziaływanie jest dobrze zdefiniowane. Oznacza to, że każda jednostka może albo być poddana oddziaływaniu, albo nie (wykluczone jest częściowe poddanie oddziaływaniu). Ponadto nie występuje współzależność między jednostkami poddanymi oddziaływaniu i jednostkami należącymi do grupy kontrolnej. Brak współzależności można utożsamić z brakiem wpływu programu na stan równowagi ogólnej.

W celu identyfikacji efektów na podstawie zbioru danych empirycznych konieczne jest spełnienie dwóch dodatkowych założeń. Po pierwsze musi istnieć wspólny przedział określoności. Dla pewnego podzbioru $X' \in X$:

$$0 < Pr(d = 0 | X' = x) < 1 \wedge 0 < Pr(d = 1 | X' = x) < 1 \quad (5)$$

Oznacza to, że jednostka charakteryzowana przez wartość wektora $X = x$ zarówno mogłaby znaleźć się w grupie poddanej oddziaływaniu, jak i grupie kontrolnej. Po drugie wartość zmiennej wynikowej musi być niezależna od stanu oddziaływania pod warunkiem znanego wektora X . Modelowany proces selekcji powinien być uzależniony wyłącznie od wielkości obserwowanych.

Pomiar efektów oddziaływania komplikuje się, gdy stan poddania oddziaływaniu jest również funkcją obserwowanych charakterystyk. W ekonometrii zagadnienie egzogeniczności jest kluczowym problemem. Pomimo swojego podstawowego znaczenia dla modelowania ekonometrycznego, pojęcie nie posiada jednoznacznej definicji. Engle, Hendry i Richard [3] zdefiniowali egzogeniczność w odniesieniu do rozkładów obserwowanych zmiennych, funkcji wiarygodności i efektywnych metod estymacji. Zmienna losowa Z jest określana jako słabo egzogeniczna wobec estymatorów nieznanymi parametrów λ , jeśli ocena λ pod warunkiem Z nie powoduje utraty informacji.

Matematycznie można to wyrazić jako przedstawienie rozkładów łącznych za pomocą rozkładów warunkowych. Jeśli zmienne losowe (X, Z) posiadają łączny rozkład, a zmienna Z jest egzogeniczna, to

$$F(X, Z) = F(X | Z) * F(Z) \quad (6)$$

Jeśli dodatkowo zmienna Z nie jest skutkiem w sensie Grangera [4], co oznacza, że żadna inna zmienna nie wpływa na Z , wtedy jest ściśle egzogeniczna.

W sytuacji, gdy celem prowadzonego badania jest ekonomiczna analiza programu egzogeniczność jest definiowana w odmienny sposób. Jej definicja jest powiązana z przyczynowością. Przyjmuje się, że cechy badanych jednostek mogą wpływać na prawdopodobieństwo z jakim jednostka zostanie poddana oddziaływaniu, oraz, że zależność nie zachodzi w drugą stronę. Fakt poddania, bądź nie poddania oddziaływaniu nie powinien wpływać na, inne niż zmienna wynikowa, cechy rozpatrywanych jednostek. W tym przypadku egzogeniczność oznacza, że stan oddziaływania nie wpływa na wartość zmiennych niezależnych.

Sytuację, w której conajmniej jedna z charakterystyk jednostek (zmiennych niezależnych) jednocześnie wpływa na oddziaływanie (ang. treatment) i wartość zmiennej wynikowej będziemy nazywać zjawiskiem endogeniczności. W takim przypadku, bez przyjęcia dodatkowych założeń, nie można odseparować wpływu czynnika endogenicznego, od zmiennej opisującej stan poddania oddziaływaniu.

Lechner [9] pokazał, że w przypadku modelu selekcji problem endogeniczności zmiennych objaśniających można pominąć, ponieważ poprawka korygująca oszacowania o nielosową selekcję uwzględnia ewentualnie występującą endogeniczność. Z kolei [10] warunki, których spełnienie umożliwia identyfikację parametrów modelu.

3 Przyczyny występowania selekcji

Prowadzenie programu pomocy czy podejmowanie interwencji na rynku powoduje oddziaływanie na część populacji. Z reguły takie działania są ukierunkowane do specyficznych grup odbiorców. Pojęcie selekcji zostało wprowadzone do słownika ekonometrii na początku lat siedemdziesiątych ubiegłego wieku. Badacze analizując ekonomiczne determinanty wysokości płacy zauważyli, że traktowanie grupy pracujących osób jako próby losowej prowadzi do nieprawidłowych wyników. Decyzja o partycypacji w sile roboczej jest zdeterminowana przez wybory i charakterystyki decydentów, a więc nie posiada losowego charakteru. W ogólnym przypadku badacz dysponuje próbą losową z populacji i wydzieloną z niej pod-

próbą. O selekcji, bądź samoselekcji mówimy, gdy obserwowane, lub nieobserwowane charakterystyki jednocześnie decydują o przynależności do próby i wielkości analizowanego zjawiska. Obciążenie związane z selekcją pojawia się wtedy, gdy nieobserwowane charakterystyki wpływają jednocześnie na fakt podjęcia pracy i wysokość płacy.

W praktyce obciążenie próby spowodowane przez nielosową selekcję może pojawić się w rezultacie występowania jednego z wielu czynników. Do najważniejszych z nich, po pierwsze, należy zaliczyć samoselekcję. Występuje ona, gdy badane jednostki będą same decydować, czy znaleźć się w jednej grupie, czy w drugiej, np. dokonując wyboru pracuje-nie pracuje. Dużo częściej selekcja towarzyszy prowadzeniu programu pomocy. Instytucja prowadząca program, zebrawszy wstępne deklaracje uczestnictwa, może w trakcie dalszego postępowania arbitralnie zadecydować komu przydzielić pomoc. Oczywiście proces selekcji może być bardziej rozbudowany, np. mieć charakter wieloetapowy. Nawet w przypadku w pełni obiektywnej procedury przydzielania pomocy wystąpi selekcja. Jest ona związana ze sposobem rozstrzygnięcia problemu komu przydzielić pomoc. Posiada ona dwa źródła. Jeżeli pomoc zostanie przydzielona wszystkim jednostkom ubiegającym się o nią, to zaburzenie losowości wystąpi podczas składania deklaracji o chęci uczestnictwa w programie. Grupa ubiegających się o pomoc będzie składać się z jednostek aktywnych społecznie, a więc wystąpi samoselekcja. Gdy pomoc nie trafi do wszystkich, oczywiste jest, że wystąpi selekcja. W kontekście budowanego modelu ważne jest, że zarówno z punktu widzenia jednostek oczekujących pomocy, jak i osób lub instytucji oceniających skuteczność lub efektywność podejmowanych działań, proces ten można traktować jako samoselekcję niezależnie od jego prawdziwej natury.

Kolejnym źródłem selekcji są arbitralne decyzje analityków prowadzących badania. Podejmowane są one podczas pracy z danymi empirycznymi. Czyszczenie zbiorów z obserwacji nietypowych lub mało wiarygodnych oraz przypadkowe usunięcia części obserwacji może powodować zaburzenie losowości próby. Powstałe zaburzenia zbioru danych mogą być przedstawione jako wynik występowania jednego z dwóch zjawisk. Po pierwsze, dane mogą być dobierane do próby za pomocą nieobserwowanego przez badaczy nielosowego procesu. Jeżeli proces jest niezależny

od działań podejmowanych przez badane jednostki, to mówi się w takim przypadku o występowaniu selekcji. Gdy proces doboru jest uzależniony od reakcji bądź cech jednostek, takie zjawisko nazywamy samoselekcją. Po drugie, zaburzenie losowego charakteru próby może wystąpić jako świadome lub przypadkowe pominięcie części obserwacji w momencie tworzenia modelu ekonometrycznego opisującego badane zjawisko.

Problem, którego nie można odróżnić od selekcji, może pojawić się jako efekt niewłaściwej specyfikacji modelu. Jest nim brak egzogeniczności zmiennych. Jeżeli w grupie zmiennych objaśniających znajdują się zmienne endogeniczne to równanie będzie posiadało nieprawidłową specyfikację. Jednak testy statystyczne bardzo często będą wskazywać na wystąpienie problemu związanego z nielosowym doбором obserwacji do próby

4 Model selekcji

Podstawą modelu selekcji jest teoria użyteczności. Zgodnie z nią jednostki dokonując wyboru kierują się wartością funkcji użyteczności dla danej alternatywy. Heckman przełożył na język matematyki koncepcję Roy'a [14] zakładającą, że obserwowane są wyłącznie wybory najlepszych alternatyw. Pomysł potraktowania problemu nielosowej selekcji analogicznie do problemu błędnej specyfikacji formy funkcyjnej modelu, oraz powiązania go z mikroekonomiczną teorią użyteczności, wyznacza ramy metodologiczne analizy. Proces budowy podstawowej wersji modelu nielosowej selekcji rozpoczyna się od równania badanego zjawiska, które zazwyczaj jest tradycyjnym równaniem regresji

$$y_i^* = X_i' \beta + \varepsilon_i \quad (7)$$

gdzie X_i jest wektorem obserwowanych charakterystyk, β wektorem szukanych parametrów modelu, a ε_i składnikiem losowym. Tym co odróżnia (7) od równania regresji jest ocenzurowanie zmiennej zależnej. Obserwujemy wartości zmiennej wynikowej y_i tylko dla wyselekcjonowanej części próby. Drugim składnikiem jest równanie opisujące mechanizm wyboru obserwacji do próby. Zakłada się, że zależy

od obserwowanych charakterystyk jednostek

$$d_i^* = Z_i' \gamma + \nu_i \quad (8)$$

gdzie Z_i jest wektorem obserwowanych charakterystyk, γ wektorem nieznanych parametrów wymagających oszacowania, a ν_i składnikiem losowym. Zmienna zależna d_i^* jest wskaźnikiem podjętej decyzji o zaliczeniu obserwacji do próby.

W celu uproszczenia postaci funkcyjnej i interpretacji parametrów równania (8) definiuje się wartość progową zmiennej decyzyjnej $\bar{d}_i$. Wykorzystuje się ją do zastąpienia nieobserwowanej wartości zmiennej decyzyjnej indykatorową zmienną zero-jedynkową

$$d_i = \begin{cases} 1 & \text{gdy } d_i^* \geq \bar{d}_i \\ 0 & \text{gdy } d_i^* < \bar{d}_i \end{cases} \quad (9)$$

Czasem przyjmuje się normalizację $\bar{d}_i = 0$, ale nie jest to konieczne. Z reguły rozkłady prawdopodobieństwa definiowane są z dokładnością do stałej. Celem normalizacji jest uproszczenie wyprowadzenia postaci analitycznej formy funkcyjnej.

Zmienna, której wartości obserwujemy w próbie y_i powstaje jako

$$y_i = y_i^* * d_i \quad (10)$$

Zestawiając powyższe równania możemy przedstawić model nielosowego doboru w postaci strukturalnej [1]

$$\begin{cases} y_i^* = X_i' \beta + \varepsilon_i \\ d_i^* = Z_i' \gamma + \nu_i \\ d_i = \begin{cases} 1 & \text{gdy } d_i^* \geq \bar{d}_i \\ 0 & d_i^* < \bar{d}_i \end{cases} \\ y_i = y_i^* * d_i \end{cases} \quad (11)$$

W celu estymacji parametrów przyjmuje się, że ε_i oraz ν_i są składnikami losowymi o średniej zero, wariancjach odpowiednio σ_ε^2 i σ_ν^2 oraz kowariancji $cov(\varepsilon_i, \nu_i) = \sigma_{\varepsilon\nu}$. W przypadku, gdy kowariancja jest równa zero, równania (7) i (8) są niezależne i nie występuje problem selekcji w modelu. Dwa pierwsze równania w (11)

są matematycznym opisem zachodzącego zjawiska ekonomicznego lub społecznego. Dwa kolejne narzucają ograniczenia na parametry modelu.

W populacji generalnej zachodzi zależność

$$E(y_i | X_i) = X_i' \beta \quad (12)$$

ale model nie jest tworzony na podstawie obserwacji całej populacji. Co więcej, nie jest on tworzony na podstawie próby losowej z tej populacji, lecz bazuje na wyselekcjonowanej w sposób nielosowy podpróbie. Metoda doboru obserwacji jest opisywana przez regułę selekcji $d_i^* \geq \bar{d}_i$. W związku z tym jest ona uwzględniana podczas obliczania wartości oczekiwanej.

$$E(y_i | X_i, d_i^* \geq \bar{d}_i) = X_i' \beta + E(\varepsilon_i | X_i, d_i^* \geq \bar{d}_i) \quad (13)$$

Przyjmuje się, że wartość progowa $\bar{d}_i$ jest identyczna dla każdej jednostki. Ponieważ wartość d_i^* nie jest bezpośrednio obserwowana, budowany jest probabilistyczny model determinujący jej wielkość. Bez utraty ogólności rozważań, możemy znormalizować wartość progową $\bar{d}_i = 0$. Ten zabieg upraszcza analityczny zapis modelu [5]. Po normalizacji parametrów modelu obserwujemy y_i jeśli $d_i^* \geq 0$.

Kontynuacja wyprowadzenia wzoru (13) polega na przekształcaniu jego prawej strony. Wstawiając wartości znormalizowane uzyskujemy

$$X_i \beta + E(\varepsilon_i | X_i, d_i^* \geq \bar{d}_i) = X_i' \beta + E(\varepsilon_i | X_i, \nu_i \geq -Z_i' \gamma) \quad (14)$$

Z założeń modelu wynika, że $E(\varepsilon_i | X_i) = 0$, ale $E(\varepsilon_i | \nu_i \geq -Z_i' \gamma) \neq 0$, ponieważ $cov(\varepsilon, \nu) = \sigma_{\varepsilon_i, \nu_i} \neq 0$, czyli kowariancja między składnikami losowymi jest różna od zera. Z tego powodu obliczone metodą standardową estymatory modelu (12) będą obciążone i niezgodne.

Model efektów oddziaływania (ang. *treatment effect*) jest bardzo zbliżony do modelu selekcji próby i może być traktowany jako jego wersja z nałożonymi restrykcjami. Został on zaadaptowany do potrzeb nauk ekonomicznych w drugiej połowie lat dziewięćdziesiątych ubiegłego wieku. Jego zaletą jest łatwość wykorzystania w modelowaniu i ocenie skuteczności prowadzonych działań o charakterze programów. Nakłada on dodatkowe ograniczenia przyjmując, że nieznane wartości

parametrów są jednakowe w grupie poddanej oddziaływaniu i grupie kontrolnej. Model efektów oddziaływania traktowany jest jak model zawierający ocenzone zmienne endogeniczne [16]. Zmodyfikowane równanie (7) przybiera formę:

$$y_i = X_i' \beta + \theta d_i + \varepsilon_i \quad (15)$$

gdzie d_i jest zmienną wskazującą czy badana jednostka została poddana oddziaływaniu. Jest ona endogeniczna, ponieważ jej wartość uzależniona jest od cech zapisanych w macierzy X . Model może być zapisany z wykorzystaniem następującej postaci strukturalnej

$$\begin{cases} y_i = X_i' \beta + \theta d_i + \varepsilon_i \\ d_i^* = Z_i' \gamma + \nu_i \\ d_i = \begin{cases} 1 & \text{gdy } d_i^* \geq \bar{d}_i \\ 0 & d_i^* < \bar{d}_i \end{cases} \end{cases} \quad (16)$$

Problem nielosowej selekcji obserwacji został sprowadzony do problemu endogenicznej zmiennej zero-jedynkowej.

Model endogenizujący efekt oddziaływania posiada jedną zasadniczą różnicę w stosunku do modelu selekcji próby. Inne jest założenie odnoszące się do rozkładu zmiennej zależnej równania zjawiska. W modelu z endogenicznym efektem oddziaływania zakłada się, że znany jest cały rozkład zmiennej zależnej, podczas gdy w modelu selekcji jest on oceniony i dostępny wyłącznie dla wyselekcjonowanej podpróby. Drugie założenie, o jednakowych wartościach nieznanymi parametrów dla grupy poddanej oddziaływaniu i pozostającej poza jego wpływem, może w prosty sposób zostać uchylone. W tym celu należy dołączyć do równania zjawiska dodatkowy zestaw zmiennych objaśniających $X_i'^*$, taki że $X_i'^* = X_i' d_i$. Wtedy równanie zjawiska zapiszemy jako

$$y_i = X_i' \beta + X_i'^* \beta^* + \eta_i \quad (17)$$

wówczas parametry β^* mierzą różnice między jednostkami poddanymi, a nie poddanymi oddziaływaniu. Efekt netto oddziaływania jest zdefiniowany jako

$$\theta = X_i'^* \beta^* = X_i' d_i \beta^* \quad (18)$$

Jest to przeciętny efekt oddziaływania programu na jednostki poddane oddziaływaniu, skorygowany za pomocą zmiennej d_i o mechanizm selekcji. W rezultacie otrzymywana jest wartość estymatora efektu netto. Parametry modelu będą zidentyfikowane o ile w każdym z równań występuje zmienna, która jest warunkowo niezależna od składnika losowego równania zjawiska [10].

Poważnym ograniczeniem modelu jest założenie o jednorodności mechanizmu selekcji. W praktyce selekcja ma bardzo zróżnicowaną naturę [7]. Jednostki poddane oddziaływaniu programu mogą w różnym stopniu odczuć jego działanie. Analogicznym problemem występującym w modelu efektów oddziaływania, który jest rozszerzeniem modelu selekcji, jest heterogeniczność reakcji na bodziec. Oba zjawiska powodują niezgodność estymatorów, pozostawiając nierozwiązanymi problemy wynikające ze zjawiska selekcji.

5 Metoda funkcji kontrolującej

Metoda funkcji kontrolującej jest zastosowaniem sposobu korygowania estymatora z uwagi na występujący nielosowy dobór obserwacji do próby do rozwiązania problemu braku egzogeniczności zmiennych objaśniających. Istotą metody jest ustalenie pewnej formy funkcyjnej. W procesie budowy funkcji kontrolującej wykorzystywane są zarówno wielkości obserwowane X jak również takie, których nie są dostępne dla badaczy i z tego powodu nieobserwowane U . Zakłada się, że można je odseparować:

$$X \cap U = \emptyset \quad (19)$$

Przez X_i oznaczony jest wektor obserwowanych cech badanych jednostek. Dla każdej z nich obserwowany jest jeden z dwóch wyników (Y_0, Y_1) . Niech $g_i(x_1, \dots, x_n)$ będzie addytywną funkcją swoich argumentów, taką że:

$$\forall i \quad E(g_i(X)) = E(X) \quad (20)$$

Czyli funkcja $g_i(\cdot)$ nie zmienia wartości oczekiwanej zmiennej losowej na którą działa. Każdy możliwy do osiągnięcia wynik może być przedstawiony jako funkcja wektora obserwowanych cech X_i i czynników nieobserwowanych U_i . Dla każdej

rozpatrywanej w badaniu jednostki obserwowane jest

$$Y_0 = g_0(X_i, U_0) \quad \vee \quad Y_1 = g_1(X_i, U_1) \quad (21)$$

Założenia (19),(20), oraz (21) definiują równania strukturalne, które są niezależne od czynników nieobserwowanych [8]. Używając notacji zdefiniowanej przez (21) do równania efektu oddziaływania na jednostki poddane działaniu programu można pokazać, że występuje zależność między tym efektem a błędem losowym.

$$E(Y_1 - Y_0 \mid X, D = 1) = g_1(X) - g_0(X) + E(U_1 - U_0 \mid X, D = 1) \quad (22)$$

W ten sposób uzyskany element po prawej stronie równania oznaczany jest w literaturze $s(x)$ i nazywany jest obciążeniem związanym z nielosową selekcją, bądź nielosowym doбором obserwacji do próby.

Metoda funkcji kontrolującej pozwala na jednoczesne uwzględnienie w modelu służącym do pomiaru oceny programu problemu endogeniczności zmiennych i problemu nielosowej selekcji do próby. Myśl ta po raz pierwszy, chociaż nie wprost, została sformułowana przez Heckmana [5]. Istotą modelu (22) jest fakt, że pozwala on w prosty sposób endogenizować, czyli włączyć do wnętrza modelu, mechanizm selekcji [8]. Dołączając selekcję do całkowitego efektu programu otrzymujemy:

$$E(Y \mid X, D, s(X)) = E(Y_1 p + Y_0(1 - p) \mid X, D, s(X)) = \quad (23)$$

gdzie p jest proporcją jednostek poddanych działaniu programu.

$$\begin{aligned} &= g_0(X) + p(g_1(X) - g_0(X)) \\ &\quad + p(E(U_1 \mid X, D = 1, s(X))) \\ &\quad + (1 - p)(E(U_0 \mid X, D = 0, s(X))) \end{aligned} \quad (24)$$

Z założeń modelu wynika, że

$$(X, D) \perp (U_0, U_1) \wedge (X, p) \perp (U_0, U_1) \quad (25)$$

ponieważ p jest skalar, D jest endogeniczne, czyli jest funkcją czynników obserwowanych, a $X \cap U = \emptyset$.

Wykorzystując warunek (25), można zapisać warunkowe wartości oczekiwane składników nieobserwowanych jako funkcje procesu selekcji

$$\forall j \in \{0, 1\} \quad E(U_j \mid X, D = j, s(X)) = k_j(s(X)) \quad (26)$$

Nazywane są one funkcjami kontrolującymi [6].

Funkcje selekcji $s(X)$ mogą być dla uproszczenia traktowane jako funkcje indykatorem. W tym podejściu zakłada się racjonalność zachowań badanych jednostek. Przypisują one każdemu poziomowi zmiennej wynikowej (Y_0, Y_1) oraz każdemu stanowi uczestnictwa w programie pewne poziomy użyteczności. Jednostka staje się uczestnikiem programu, gdy oczekiwana użyteczność z faktu uczestnictwa w programie przekracza poziom graniczny $\bar{U}$. Przyjmuje się, że poziomem granicznym jest maksymalna wartość funkcji użyteczności jaką może uzyskać jednostka bez konieczności uczestniczenia w programie lub oczekiwana użyteczność z faktu braku uczestnictwa w rozpatrywanym programie. Dodatkowo zakładana jest niezależność między graniczną wartością użyteczności a mechanizmem selekcji.

$$s(X) \cap \bar{U} = \emptyset \quad (27)$$

Szczegółowe założenia dotyczące rozkładów zmiennych losowych ($U_j, \bar{U}$) determinują formy funkcyjne $k_j(\cdot)$.

Zapisując całkowite efekty programów za pomocą notacji (23) przedstawiamy je jako dwa równania selekcji zastosowane do subpopulacji poddanej oddziaływaniu programu i subpopulacji pozostającej poza programem

$$\forall j \in \{0, 1\} \quad E(Y_j \mid X, D = j, s(X)) = g_0(X) + k_j(s(X)) \quad (28)$$

Jak zostało pokazane, aby stworzyć funkcję kontrolującą nie potrzebujemy dodatkowych zmiennych. Informacja zgromadzona w zbiorze danych jest zazwyczaj wystarczająca. Konieczność posiadania dodatkowych zmiennych zastępowana jest dodatkowym równaniem, które jest określane jako funkcja kontrolująca lub równanie selekcji.

Metoda funkcji kontrolującej daje podobne wyniki do metody zmiennych instrumentalnych. Jej zaletą jest łatwość rozszerzenia na większą liczbę czynników

odpowiadających za proces selekcji. Należy jednak wziąć pod uwagę, że każde nowe równanie z wielokrotnia liczbę parametrów wymagających oszacowania. Ponadto aby otrzymać zgodne estymatory za pomocą metod parametrycznych należy założyć niezależność różnych funkcji kontrolujących [8]. Problemy te są bardziej związane ze sposobami estymacji niż z metodologią prowadzenia badań.

Niewątpliwą zaletą funkcji kontrolujących jest fakt, iż nie jest wymagana egzogeniczność czynników kontrolujących. Rozpatrując to zagadnienie w bardziej szczegółowy sposób można zauważyć, że te same czynniki mogą być składnikami równania selekcji i równania modelu. Zostawia to większą swobodą działania badaczom, ponieważ dużo łatwiej jest przekonująco uzasadnić że dany czynnik wpływa na badane zjawisko, niż że jest tego wpływu pozbawiony i występuje zupełnie niezależnie od niego.

W prostych modelach zawierających jedną funkcję kontrolującą z reguły nie ma problemu z doбором jej składników. Naturalnymi kandydatami są, np. zmienne które w literaturze ekonomicznej są przedstawiane jako dobre instrumenty w kontekście badanego zjawiska [17]. Jednakże do uwzględnienia bardziej skomplikowanych mechanizmów selekcji potrzebne są założenia a-priori wynikające z teorii ekonomicznej, albo trudne do spełnienia restrykcje dotyczące funkcyjnych.

6 Podsumowanie

Metody analizy danych pochodzących z badania obserwacyjnego różnią się od standardowych narzędzi ekonometrycznych. Podstawowe znaczenie dla prawidłowości analizy ma rozróżnienie na zmienne sprawcze (przyczyny) i efekty działań. W sytuacji, gdy stan poddania oddziaływaniu nie zależy od charakterystyk jednostki i zmiennych sprawczych, to naturalny eksperyment posiada cechy kontrolowanego eksperymentu. Sytuacja jest bardziej skomplikowana, gdy stan oddziaływania jest skorelowany ze zmiennymi sprawczymi. Taką sytuację określamy jako oddziaływanie o endogenicznym charakterze.

Przedstawiono rozwiązanie problemu selekcji zaproponowane przez Heckmana. Wyjaśniono dlaczego dane pochodzące z badań dotyczących naturalnych ekspe-

rymentów podlegają nielosowej selekcji. Ukazano sposób wykorzystania modeli uwzględniających nielosowy mechanizm selekcji do analizy danych z badań obserwacyjnych. Ponadto pokazano, że model uwzględniający nielosowy mechanizm doboru obserwacji do próby i model pomiaru efektów oddziaływania o endogenicznym charakterze można przekształcić do jednakowej postaci strukturalnej. W zasadzie te dwa modele, pomimo różnych podstaw teoretycznych, są dwoma wariantami tego samego zagadnienia. Przeprowadzona unifikacja pozwala powiązać fakt uczestnictwa w programie bezpośrednio z teorią użyteczności.

Z drugiej strony model nielosowej selekcji może być rozpatrywany jako narzędzie pozwalające na uzyskanie oszacowań nieznanych parametrów modelu efektów oddziaływania w sytuacji, gdy informacja o rozkładzie zmiennej zależnej jest ograniczona. Co prawda uzyskane estymatory nie będą efektywne, ale będą zgodne.

Bibliografia

- [1] Amemiyia Takeshi (1984) "Tobit Models: A Survey", *Journal of Econometrics*, vol. 24. pp. 3-61.
- [2] Cochran W. (1965) "The Planning of Observational Studies of Human Populations", *Journal of Royal Statistical Society*, seria A, vol. 128. str. 134-155.
- [3] Engle R., Hendry D., Richard J-F. (1983) "Exogeneity", *Econometrica*, vol. 51/2. str. 277-304.
- [4] Granger CW. (1969) "Investigating Causal Relationship by Econometric Models and Cross-Spectral Methods", *Econometrica*, vol. 37/2. str. 424-438.
- [5] Heckman James J. (1979) "Sample Selection Bias as a Specification Error", *Econometrica*, vol. 47/1. str. 153-162.
- [6] Heckman James, Robb Richard (1985) "Alternative Identifying Assumptions in Econometric Models of Selection Bias", *Advances in Econometrics*, vol 5. pp. 243-287.
- [7] Heckman James J. (1990) "Varieties of Selection Bias", *American Economic Review*, vol. 80/2. pp. 313-318.
- [8] Heckman James J., LaLonde Robert, Smith Jeffrey (1999) "The Economics and Econometrics of Active Labor Market Programs" w Ashenfelter, Orley, Card, David [red] *Handbook of Labor Economics vol 3a*, Elsevier.
- [9] Lechner Michael (2005) "Note on Endogenous Control Variables in Evaluation Studies", *Universitat St. Gallen Discussion Paper*, nr 2005-16.
- [10] Lewbel Artur (2005) "Endogenous Selection or Treatment Model Estimation" *Boston College Working Paper*
- [11] Neyman Jerzy (1923) "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9", angielskie tłumaczenie z *Roczniki Nauk Rolniczych*, tom X, str 1-51.

- [12] Rosenbaum Paul (1999) *Observational Studies*, Springer-Verlag. New York-Berlin-Heidelberg.
- [13] Rosenbaum Paul, Rubin Donald B. (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects.", *Biometrika*, vol. 70. str. 41-55.
- [14] Roy Andrew D. (1951) "Some Thoughts on the Distribution of Earnings" *Oxford Economic Papers* New series, vol. 3/2. str. 135-146.
- [15] Rubin Donald B. (1973) "Matching to Remove Bias in Observational Studies", *Biometrics*, vol. 29/1. str. 159-183.
- [16] Vella Francis (1998) "Estimating Model with Sample Selection Bias: A Survey", *Journal of Human Resources*, vol. 33/1. str. 127-169.
- [17] Vella Francis, Verbeek Marno (1998) "Estimating and Interpreting Models with Endogenous Treatment Effect", *Journal of Business and Economic Statistics*, vol. 17. str. 473-478.