Mikroekonometria

Modele i metody analizy danych indywidualnych

redakcja naukowa Marek Gruszczyński

wydanie II rozszerzone

Monika Bazyl, Marek Gruszczyński, Monika Książek,
Marcin Owczarczuk, Adam Szulc, Arkadiusz Wiśniowski, Bartosz Witkowski

Oficyna
a Wolters Kluwer business

Mikroekonometria

Modele i metody analizy danych indywidualnych

redakcja naukowa Marek Gruszczyński

Monika Bazyl, Marek Gruszczyński, Monika Książek,
Marcin Owczarczuk, Adam Szulc, Arkadiusz Wiśniowski, Bartosz Witkowski

wydanie II rozszerzone

Warszawa 2012

Oficyna
a Wolters Kluwer business

Recenzent
Prof. UG dr hab. Krystyna Strzała

Wydawca
Monika Pawłowska

Redaktor prowadzący
Ewa Fonkowicz

Opracowanie redakcyjne
Justyna Łęczyńska

Łamanie
Sławomir Sobczyk

Projekt graficzny okładki i zdjęcie
Barbara Widłak

Autorzy:
Monika Bazyl – rozdział 7 i p. 3.6
Marek Gruszczyński – rozdziały 1 i 3
Monika Książek – rozdział 4
Marcin Owczarczuk – rozdziały 2 (p. 2.1–2.4) i 6
Adam Szulc – rozdział 9
Arkadiusz Wiśniowski – rozdziały 2 (p. 2.1) i 5
Bartosz Witkowski – rozdział 8

ISBN: 978-83-264-3775-5
wydanie 2. rozszerzone

ISBN PDF-a: 978-83-264-5184-3

Wydane przez:
Wolters Kluwer Polska Sp. z o.o.

Redakcja Książek
01-231 Warszawa, ul. Płocka 5a
tel. 22 535 82 00, fax 22 535 81 35
e-mail: ksiazki@wolterskluwer.pl

www.wolterskluwer.pl
księgarnia internetowa www.profinfo.pl

Spis treści

Wstęp

Mikroekonometria to analiza mikrodanych dotyczących zagadnień ekonomicznych, finansowych i społecznych. Mikrodane to dane o pojedynczych uczestnikach życia ekonomicznego: osobach, gospodarstwach domowych, firmach itd., a także o zachowaniach tych uczestników. Coraz większa dostępność mikrodanych stwarza potrzebę sięgania po nowoczesne narzędzia ich analizy, jakich dostarcza mikroekonometria. Jest to ta część ekonometrii, która rozwija się od kilkudziesięciu lat, głównie ze względu na wzrost możliwości obliczeniowych oraz dostępności dużych zbiorów mikrodanych.

Rosnący popyt na analizy mikrodanych odnotowuje się w marketingu, finansach, zarządzaniu, w ekonomii i w innych naukach społecznych. Popularne zastosowania modeli mikroekonometrii obejmują: badanie zdolności kredytowej, prognozowanie odejść klientów i bankructw firm, wybór grup docelowych do kampanii bezpośrednich, badanie preferencji klientów dotyczących marek czy modelowanie szkód i wypłat ubezpieczeniowych.

Niniejsza książka stanowi wprowadzenie do analizy mikrodanych i przedstawia ekonometryczne metody modelowania zmiennych jakościowych lub ograniczonych. Oprócz przeglądu modeli i metod książka zawiera przykłady i ćwiczenia. Jest przeznaczona dla studentów uczelni ekonomicznych mających zajęcia z ekonometrii oraz praktyków analityków korzystających z mikrodanych.

Przygotowanie książki pod tytułem *Mikroekonometria* stanowiło duże wyzwanie dla autorów. Taki tekst powinien bowiem pomieścić zarówno podstawy teoretyczne, jak i przykłady zastosowań, decydujących o zainteresowaniu wielu potencjalnych odbiorców. Nasza książka stara się spełnić te oczekiwania. Jednocześnie mamy świadomość, że nie wszystkie teoretyczne i empiryczne kwestie zostały objaśnione ku zadowoleniu wymagających odbiorców. To jest zawsze sprawa kompromisu – w tym wypadku między objętością książki a jej przeznaczeniem.

W zamyśle autorów jest to podręcznik do przedmiotu „mikroekonometria" wykładanego od niedawna w Polsce, ale jeszcze nieoferowanego powszechnie na studiach ekonomii i zarządzania. To się szybko zmienia, ponieważ coraz częściej mamy do czynienia z dużymi zbiorami mikrodanych, które powinny być analizowane, z perspektywy marketingowej (dane sprzedażowe ze skanerów kasowych) bądź z punktu widzenia obserwatora rynku, na przykład rynku kapitałowego, gdzie mikrodane też tworzą się w sposób automatyczny[1].

Zadanie dla analityka nie jest wcale proste. Mikrodane można analizować według różnych reguł. Ta książka podkreśla regułę ekonometryczną. Znaczy to, że mamy zawsze na myśli pewien model opisujący takie dane, liniowy lub nieliniowy. Jest to model stochastyczny, jak każdy model ekonometryczny. Staramy się opisywać modele mikroekonometrii jako modele behawioralne, to znaczy „podpowiadamy modelom" pewne reguły czy zachowania jednostek, które mają znajdować w nich odzwierciedlenie. Nie są to modele oparte na technikach *data mining*, jakkolwiek obecnie coraz trudniej jest ustalać granice pomiędzy technikami analizy mikrodanych.

Mikroekonometria wywodzi się z mikroekonomii. Tego nie przyznają chętnie matematycy lub statystycy, którzy – jak się zdaje – zawładnęli technologią mikroekonometrii. W obecnym świecie mnogości danych niepodobna jednak uciec od pytań o sens modelowania. Można przecież wszystko zautomatyzować, ale kierunek myślenia przy tych automatyzacjach ma decydujące znaczenie. Ta książka wskazuje jeden z takich kierunków. Stoimy na stanowisku, że podstawą wyboru jest racjonalna decyzja dokonana w wyniku analizy możliwości. Uważamy, że jednostki podejmujące decyzję lub będące w określonych sytuacjach wybierają różne rozwiązania, ich grupy są zatem niejednorodne, heterogeniczne. Modelowanie ich decyzji powinno to uwzględniać. Przy tym jest to modelowanie w skali statystycznej. Nie koncentrujemy się na Kowalskim, który decyduje, czy dziś pojedzie do pracy tramwajem, czy rowerem. Interesuje nas duży zbiór różnych Kowalskich i ich różne wybory.

Sądzimy, że nasza książka koresponduje z rozmaitymi tekstami na temat mikroekonometrii i analizy mikrodanych, które ukazały się w ostatnich latach. W miarę pełne omówienie pozycji wydanych przed rokiem 2000, takich jak publikacja Maddali (1983), jest dostępne w pracy Gruszczyńskiego (2001). Autorami współczesnych monografii, do których nawiązujemy w naszej książce, są m.in.: Wooldridge (2002), Cameron i Trivedi (2005), Hensher, Rose i Greene (2005), Winkelmann i Boes (2006), Train (2003), Lee (2010).

Analiza mikrodanych wymaga korzystania z odpowiednich programów. Przy prezentacji naszych tekstów zdecydowaliśmy się przede wszystkim na program Stata. To oznacza, że wiele przykładów jest prezentowanych w języku Stata, można też w tekście spotkać komendy i sugestie różnych rozwiązań związanych z tym programem. Nie przedstawiamy tu jednak wyczerpującego wykładu na temat Stata. Drugim programem, który można wykorzystać w wielu przykładach z tej książki, jest Gretl.

[1] Zob. felieton mojego autorstwa *Będzie coraz więcej liczb* w „Gazecie SGH" pod adresem http://akson.sgh.waw.pl/ thorrel/gazeta/archiwum/pdf/SGH-2009-05.pdf.

Warto podkreślić dostępność dobrych podręczników z mikroekonometrii opartych na programie Stata. Mam tu na myśli autorów takich jak: Cameron i Trivedi (2009), a także Long i Freese (2006). Lżejsza pozycja to praca autorstwa Bauma (2006). Mottem naszej książki są zastosowania. Ich przykłady znajdziemy w tekście. Wiele innych spotkamy, śledząc doniesienia z badań na rozmaite tematy. Jeśli chodzi o zagadnienia ekonomiczne, finansowe i związane z zarządzaniem, dobrym źródłem jest baza artykułów SSRN oraz baza IDEAS. Niektóre zastosowania doczekały się książek z mikroekonometrią w tytule, z nowszych wymienimy prace takich autorów jak Caliendo (2006), Carvalho (2009), Degryse, Kim i Ongena (2009) oraz Lee (2005).

Jak w większości współczesnych tekstów na temat mikroekonometrii, tematy w naszej książce koncentrują się na modelach zmiennych jakościowych i ograniczonych, a główna metoda estymacji to metoda największej wiarygodności. Zwracamy przy tym uwagę, że podstawowe modele opisywane w książce można przedstawić w szerszym kontekście modelowania zmiennej ukrytej (*latent*). Mówi o tym jeden z początkowych rozdziałów. Innym nowym elementem jest modyfikacja metody *maximum score* zasygnalizowana w rozdziale 6. Autorem obu tych omówień jest Marcin Owczarczuk. Pozostali autorzy – Monika Bazyl, Monika Książek, Arkadiusz Wiśniowski oraz niżej podpisany – opracowali kolejne rozdziały, w szczególności opisujące rzadziej prezentowane w Polsce modele zmiennych wielomianowych dla kategorii uporządkowanych i nieuporządkowanych. Autorzy, związani z Zakładem Ekonometrii Stosowanej w Szkole Głównej Handlowej w Warszawie, chętnie przyjmą wszelkie uwagi od życzliwych czytelników.

***

Pierwsze wydanie książki zostało bardzo przychylnie przyjęte przez czytelników. Otrzymaliśmy od Państwa wiele uwag, propozycji udoskonalenia tekstu oraz uzupełnień. Drugie wydanie „Mikroekonometrii" jest rozszerzone o dwa nowe rozdziały, oba napisane przez ekspertów w swoich dziedzinach.

Nowy rozdział ósmy o modelach ekonometrii panelowej, autorstwa Bartosza Witkowskiego z Zakładu Metod Probabilistycznych w Instytucie Ekonometrii SGH, wzbogaca naszą książkę o nowy wymiar, to jest o modelowanie przestrzenno-czasowe. Ten rozdział stanowi kompendium wiedzy na temat modeli dla danych panelowych, z przykładami i propozycjami korzystania z programu Stata. Drugi nowy rozdział (dziewiąty), autorstwa Adama Szulca z Instytutu Statystyki i Demografii SGH, przedstawia tematykę szacowania efektów oddziaływania, niezwykle ważną w mikroekonometrii. Mottem tego rozdziału jest estymacja przez dopasowanie (*matching*). Szacowanie efektów oddziaływania należy do szybko rozwijających się gałęzi mikroekonomii empirycznej, określanych jako metody oceny polityki społecznej (*policy evaluation methods*), czy wręcz jako metody ekonometrii „ewaluacyjnej" (*evaluation econometrics*).

W stosunku do pierwszego wydania uzupełniliśmy znacząco rozważania w rozdziale 2, między innymi o M-estymatory, o dyskusję na temat wielokrotnego testowania hipotez (testowania zbyt wielu modeli) i dyskusję na temat trafności prognoz. W rozdziale 4 wprowadzony został podrozdział o modelu częściowo proporcjonalnych szans -

jako pośrednim pomiędzy standardowym modelem uporządkowanym a uogólnionym modelem uporządkowanym. Omówiono także zdarzający się w zastosowaniach problem rzadkich danych, to znaczy problem bardzo małej lub zerowej liczności obserwacji dla niektórych kombinacji kategorii zmiennych w modelu. We wszystkich rozdziałach poprawiliśmy dostrzeżone błędy i wprowadziliśmy różne uzupełnienia. Dziękujemy czytelnikom, a przede wszystkim naszym studentom z przedmiotu „Mikroekonometria", za wskazanie niedociągnięć tekstu w pierwszym wydaniu. Nowe wydanie książki zawiera indeks. Spis literatury również został uzupełniony.

Wyrażamy nadzieję, że nowe rozszerzone wydanie „Mikroekonometrii" będzie jeszcze lepiej służyć naszym czytelnikom, których niniejszym zachęcamy do dalszego przesyłania uwag i komentarzy.

Marek Gruszczyński

ROZDZIAŁ 1

Wprowadzenie do mikroekonometrii

Pierwszy rozdział stanowi wprowadzenie do problematyki mikroekonometrii. W części wstępnej przedstawiamy różne rodzaje mikrodanych oraz modele, które mogą je opisywać. Dalej następuje przegląd głównych zagadnień mikroekonometrii, to znaczy tematów wyróżniających analizę mikrodanych spośród innych analiz ekonometrycznych. Zwracamy uwagę na strategię modelowania, na specyficzne cechy zbiorów mikrodanych, na kwestię endogeniczności w modelu, a także niejednorodności danych. Wskazujemy, jak dobór próby wpływa na jakość wyników badania mikroekonometrycznego. Pokazujemy też relację mikroekonometrii z innymi dziedzinami analizy danych, takimi jak *data mining* czy ekonometria przestrzenna. Rozdział kończy się laudacją dla Jamesa Heckmana i Daniela McFaddena – laureatów Nagrody im. Alfreda Nobla z dziedziny ekonomii w roku 2000.

1.1. Mikrodane

Zbiory danych liczbowych o pojedynczych jednostkach, takich jak klienci banku, gospodarstwa domowe, rodziny, firmy i tak dalej, to zbiory **mikrodanych**. Mikrodane zawierają informacje obiektywne lub subiektywne. Są to na przykład dane o każdym z 1000 konsumentów, których spytano o ocenę smaku nowej kawy. Płeć konsumenta to informacja obiektywna, a ocena smaku kawy jest subiektywna. Właśnie ta ocena jest przedmiotem zainteresowania odbiorców takich danych. Mikrodane mogą być produktem ubocznym innych działalności, np. polegającej na utrzymywaniu i zarządzaniu danymi z urzędów skarbowych bądź z NFZ czy też z ZUS. Mikrodanymi są niezwykle liczne informacje liczbowe o zachowaniu się klientów operatora telefonii

komórkowej: czas trwania każdej rozmowy, częstotliwość rozmów, koszt rozmów itd. Mikrodanymi są także próby ze zbiorów rozmaitych transakcji, takich jak informacje ze skanerów w supermarketach. Krótko mówiąc – mikrodane nas otaczają. W tej książce opisujemy niektóre sposoby ich analizowania, takie, które zalicza się do mikroekonometrii.

Cechą zbiorów mikrodanych jest ich duża liczebność oraz to, że mają z reguły charakter **danych przekrojowych**. Dane panelowe też zalicza się do mikrodanych, szczególnie wtedy, kiedy wymiar czasowy jest bardzo mały w porównaniu z wymiarem przekrojowym, np. dane o wynikach finansowych 4000 firm z trzech kolejnych lat.

Mikrodane to **dane obserwacyjne**, pochodzą bowiem z badań ankietowych bądź z dostępnych baz danych administracyjnych. Należy je odróżnić od **danych eksperymentalnych** pochodzących z odpowiednich eksperymentów. Dane obserwacyjne mogą być obarczone **błędem selekcji próby**, jeśli ich zbiór nie stanowi próby losowej.

Mikrodane to przede wszystkim **dane przekrojowe**. Są one najprostsze do uzyskania, ale też nie odpowiadają na wszystkie pytania, szczególnie gdy dochodzi wymiar czasowy badanej relacji. Wyjątkiem jest sytuacja, gdy mamy do czynienia z **populacją stacjonarną**, czyli taką, dla której momenty charakteryzujących ją zmiennych są stałe (Cameron, Trivedi, 2005). Jest to jednak pewna abstrakcja, na ogół niezdarzająca się w praktyce.

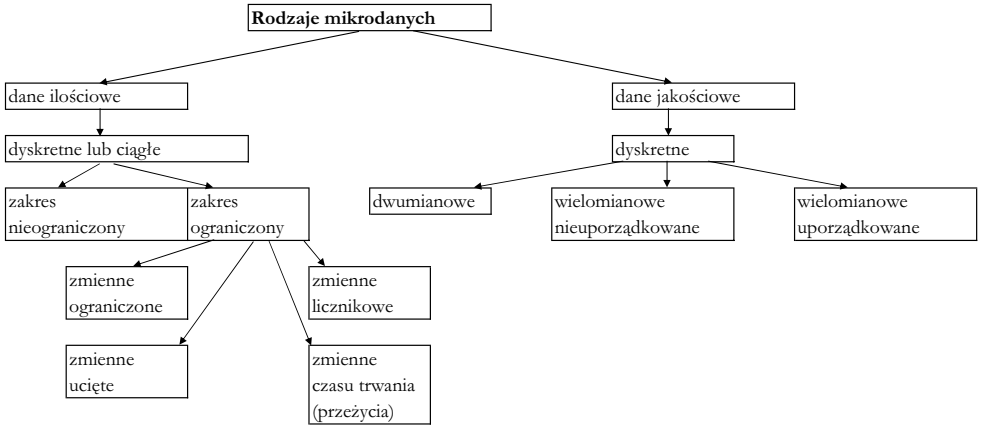
Kolejny rodzaj to **powtarzane dane przekrojowe**, ciąg niezależnych prób, rodzaj panelu z próbą zmieniającą się z okresu na okres. Jeśli populacja nie jest stacjonarna, to poszczególne próby łączy relacja zależna od tego, jak populacja zmienia się w czasie. Jednym z rozwiązań jest konstruowanie na podstawie powtarzanych danych przekrojowych tak zwanych pseudopaneli bądź paneli syntetycznych.

Wreszcie **dane panelowe** (*panel data, longitudinal data*) oznaczają obserwacje na wylosowanej próbie jednostek, dokonywane w ciągu kolejnych okresów. Oczywistą zaletą danych panelowych, jeśli tylko obejmują dostatecznie długą sekwencję obserwacji w czasie, jest możliwość modelowania zarówno relacji statycznych, jak i dynamicznych. Główne problemy paneli to reprezentatywność próby i zmęczenie próby (wypadanie jednostek z panelu).

Ważną cechą mikrodanych jest to, że zmienne, które reprezentują, są często mierzone na **skali nieciągłej**. Są to **zmienne jakościowe**.

Mikrodane dzielimy ogólnie na dane ilościowe (*quantitative*) i dane jakościowe (*qualitative*). Mikroekonometria zajmuje się w szczególności danymi jakościowymi, jako że metody analizy takich danych różnią się znacznie od metod analizy danych ilościowych (por. rys. 1.1). Zanim zagłębimy się w szczegóły, potrzebne będzie szersze spojrzenie na samą dziedzinę mikroekonometrii.

Rys. 1.1 Rodzaje mikrodanych



Źródło: Winkelmann i Boes (2006) oraz opracowanie własne

1.2. Obszar mikroekonometrii, tradycje i piśmiennictwo

„*At its heart economic theory is about individuals (or families or firms) and their interactions in markets and other social settings*[1]*. (...) The field of microeconometrics emerged in the past forty years to aid economists in providing more accurate descriptions of the economy, in designing and evaluating public policies, and in testing economic theories and estimating the parameters of well posed economic models. It is a scientific field within economics that links the theory of individual behaviour to individual data, where individuals may be firms, persons or households.*"

James Heckman, przemówienie noblowskie, 8 grudnia 2000 r. (2000, 2001, 2004).

Można zgodzić się z tą szeroką definicją mikroekonometrii autorstwa Jamesa Heckmana, laureata Nagrody Nobla w dziedzinie ekonomii w roku 2000. Mikroekonometria jest odpowiedzią na wyraźne potrzeby ekonomii i innych nauk społecznych, dotyczące dokładniejszego opisu zjawisk i lepszej weryfikacji rozmaitych hipotez. Odgrywa szczególną rolę w naukach ekonomicznych, ponieważ daje narzędzia analizy zachowań jednostek przy użyciu mikrodanych, przy czym jednostkami mogą być firmy, pojedyncze osoby czy gospodarstwa domowe.

Jest to jedna z wielu dziedzin współczesnej nauki, które swoją popularność i przydatność zyskały dzięki komputerom. Ostatnie 40–50 lat były okresem wielkiego rozwoju mikroekonometrii, głównie za sprawą dostępności mikrodanych oraz możliwości obliczeniowych. Podstawy metodyczne sięgają przeszłości znacznie odleglejszej. Jeśli szukać „ojców" mikroekonometrii, to są nimi wszyscy, którzy w badaniach z zakresu

[1] Słowa *or families or firms* zostały usunięte z tekstu przemówienia noblowskiego Jamesa Heckmana w późniejszych wydaniach.

ekonomii korzystali z mikrodanych. Należą do nich z pewnością Engel (1857), Stone (1953), Houthakker (1957) czy Tobin (1958). Typowych przykładów zastosowań mikroekonometrii w ekonomii szuka się najczęściej w ekonomii pracy; chodzi tu o klasyczną zależność płacy od poziomu wykształcenia czy liczby dzieci urodzonych przez kobietę od jej możliwości na rynku pracy.

Heckman (2004) podaje następujące powody, dla których mikroekonometria przyczyniła się do rozwoju wiedzy ekonomicznej:

- Mikroekonometrycy rozwinęli nowe narzędzia umożliwiające podjęcie wyzwań, jakie stanęły przed ekonometrią w postaci dużych zbiorów mikrodanych.
- Mikroekonometria wzbogaciła metody analizy szeregów czasowych poprzez konstrukcję modeli, które połączyły modele ekonomiczne dla pojedynczych uczestników życia ekonomicznego z danymi na temat indywidualnych zachowań (dane panelowe).
- Badania mikroekonometryczne niezmiennie dowodzą istnienia różnorodności i heterogeniczności indywidualnych zachowań. Ta heterogeniczność ma wielkie znaczenie dla teorii ekonomii oraz dla praktyki ekonometrycznej.
- Mikroekonometria ma ponadto duże znaczenie jako narzędzie naukowej oceny rozmaitych programów polityki społecznej.

Sam termin „mikroekonometria" pojawia się w zachodniej literaturze ekonometrycznej od połowy lat osiemdziesiątych (zob. np. Pesaran, 1987; Ronning, 1991 lub Blundell, 1996). Nazwa wynika z potrzeby wyodrębnienia tej części ekonometrii, która obejmuje metody wykorzystania mikrodanych do analizy zagadnień ekonomicznych. Są to metody wyraźnie różne od klasycznych, przede wszystkim w przypadku, gdy mikrodane są podstawą modelowania zmiennych jakościowych lub ograniczonych. Mikroekonometria zdobywa popularność także na poziomie akademickim. Są już pierwsze podręczniki do ekonometrii, w których wyraźnie oddzielono analizę danych przekrojowych od analizy szeregów czasowych, np. podręcznik Wooldridge'a (2003). Takie w istocie są też podręczniki Stocka i Watsona (2007) oraz Heija, de Boera, Fransesa, Kloeka i van Dijka (2004), a także w jakiejś mierze podręcznik Mycielskiego (2009). Monografie dotyczące samej mikroekonometrii wymieniono też we wstępie do tej książki. Wskazano tam na pozycje literatury dotyczące programu Stata. Warto też dodać, że monografia Henshera, Rose'a i Greene'a (2005) na temat stosowanej analizy wyboru jest w istocie poświęcona obsłudze i zastosowaniu innego programu: NLOGIT.

Część mikroekonometrii wykorzystuje dane panelowe. W naszej książce modele ekonometrii panelowej omawia rozdział 8. Z monografii dotyczących ekonometrii panelowej można podać książki takich autorów, jak Arellano (2003), Baltagi (2008), Hsiao (2003), Hsiao, Lahiri, Lee i Pesaran (1998), Lee (2002).

W literaturze polskiej do mikroekonometrii nawiązują autorzy i redaktorzy kilku monografii, na przykład Gruszczyński (2001), Wiśniewski (1986), Walesiak (1996), Gatnar i Walesiak (2004), Marzec (2008). Jest też monografia Hozera (1993), gdzie za mikroekonometrię uznaje się wszelkie zastosowania ekonometrii w mikroekonomii. Warto dodać, że wiele współczesnych badań z zakresu ekonomii, finansów i zarządzania

w Polsce opiera się na mikrodanych przekrojowych bądź na danych panelowych. Znaczącą ich część można zaliczać do obszaru mikroekonometrii. Są one coraz częściej publikowane w renomowanych czasopismach naukowych, a także w krajowych monografiach konferencyjnych.

1.3. Modele mikroekonometrii

Mikrodane, które są podstawą modelowania w mikroekonometrii, można klasyfikować na różne sposoby. Winkelman i Boes (2006) proponują następującą klasyfikację mikrodanych ilościowych i jakościowych, a właściwie modeli, które mają opisywać zmienne mające takie cechy (por. rys. 1.1):

1. Mikrodane ilościowe: dyskretne lub ciągłe
   a) zakres nieograniczony
   b) zakres ograniczony
      - ograniczone zmienne zależne
      - zmienne czasu trwania (przeżycia)
      - zmienne licznikowe
2. Mikrodane jakościowe: dyskretne
   a) dwumianowe (binarne)
   b) wielomianowe nieuporządkowane
   c) wielomianowe uporządkowane

Podstawą klasyfikacji mikrodanych jest model, który objaśnia zmienną o określonej postaci. Nieco inny podział proponuje Gruszczyński (2001). Bierze się pod uwagę jedynie zmienne jakościowe i ograniczone, których modele podane są w czterech grupach. Tabela 1.1 przedstawia jedną z wersji tej klasyfikacji.

W poniższych przykładach przedstawimy kilka modeli mikroekonometrii, które można znaleźć w obu klasyfikacjach.

Przykład 1.1. Czy pracujesz?
Możliwe odpowiedzi:
$Y = 1$ tak
$Y = 0$ nie
Wyobraźmy sobie pewną zmienną ciągłą Y^* oznaczającą skłonność do podejmowania pracy. Jest to zmienna nieobserwowalna (ukryta). O zmiennej ukrytej dowiemy się więcej z rozdziału 2. Przyjmijmy, że jeśli $Y^* \geqslant 0$, to pracuję, a jeśli $Y^* < 0$, to nie pracuję. Obserwujemy jednak tylko wartości Y, wartości zmiennej binarnej.
Wniosek: do opisu zmiennej jakościowej Y powinniśmy zastosować model dwumianowy.
Co modelujemy? Prawdopodobieństwa $p = P(Y^* \geqslant 0)$, czyli $p = P(Y = 1)$ oraz $1 - p = P(Y^* < 0)$, czyli $1 - p = P(Y = 0)$. Szczegóły w rozdziale 3.

Tab. 1.1 Systematyka modeli zmiennych jakościowych i ograniczonych

1. Modele dwumianowe	**2. Modele wielomianowe**
liniowy model prawdopodobieństwa (LMP)	**Modele kategorii uporządkowanych**
probitowy	uporządkowany model logitowy i probitowy
logitowy	uogólniony model uporządkowany
logarytmiczno-liniowy	modele danych sekwencyjnych
gompitowy (krzywej Gompertza)	
komplementarny log-log	**Modele kategorii nieuporządkowanych**
burritowy (rozkładu Burra)	wielomianowy model logitowy i probitowy
ucięty LMP	warunkowy model logitowy (McFadden)
krzywej Urbana	zagnieżdżony model logitowy
	mieszany model logitowy
3. Modele zmiennych ograniczonych	**4. Modele licznikowe i czasu trwania**
regresja ucięta	regresja Poissonowska
modele tobitowe (klasyczne)	model rozkładu ujemnego dwumianowego
dwugraniczny model tobitowy	model czasu trwania
model doboru próby (Heckman)	model licznikowy ucięty
modele efektów oddziaływania	model płotkowy

Źródło: Gruszczyński (2001).

Przykład 1.2. Czy zgadzasz się ze stwierdzeniem: każda mama powinna zrezygnować z pracy i wychowywać swoje dziecko?
Możliwe odpowiedzi:
$Y = 1$ zdecydowanie się nie zgadzam, $Y = 2$ nie zgadzam się, $Y = 3$ nie mam zdania na ten temat, $Y = 4$ zgadzam się, $Y = 5$ zdecydowanie zgadzam się z tym stwierdzeniem. Tutaj zmienną ukrytą jest Y^* skłonność do wyrażania zgody na to stwierdzenie (zmienna ciągła). Nadal mamy zmienną ciągłą Y^*, ale więcej punktów na osi Y^*, które nas interesują. Oznaczmy je jako τ_1, τ_2, τ_3, τ_4 (każdy większy od poprzedniego):

$$Y = 1 \quad \text{dla} \quad Y^* < \tau_1, \qquad Y = 2 \quad \text{dla} \quad \tau_1 \leqslant Y^* < \tau_2, \quad Y = 3 \quad \text{dla} \quad \tau_2 \leqslant Y^* < \tau_3,$$
$$Y = 4 \quad \text{dla} \quad \tau_3 \leqslant Y^* < \tau_4, \quad Y = 5 \quad \text{dla} \quad \tau_4 \leqslant Y^*.$$

Wniosek: powinniśmy zastosować model wielomianowy kategorii uporządkowanych.
Co modelujemy? Prawdopodobieństwa p_1, p_2, p_3, p_4, p_5 dla (odpowiednio) Y=1, 2, 3, 4 i 5. Szczegóły w rozdziale 4.

Przykład 1.3. Jaką wykonujesz pracę?
Możliwe odpowiedzi:
$Y = A$ praca fizyczna, $Y = B$ praca biurowa,
$Y = C$ praca menedżerska, $Y = D$ inna praca.

Tutaj mamy 4 prawdopodobieństwa: p_A, p_B, p_C i p_D, które odpowiadają skłonności do wykonywania pracy danego rodzaju. Wartości zmiennej Y nie da się uszeregować w jakiejkolwiek obiektywnej kolejności.
Wniosek: powinniśmy zastosować model wielomianowy kategorii nieuporządkowanych.

Co modelujemy? Ilorazy prawdopodobieństw p_B/p_A, p_C/p_A oraz p_D/p_A, gdzie A przyjęto za kategorię bazową. Szczegóły w rozdziale 5.

Przykład 1.4. Ile jest dzieci do lat 6 w Twojej rodzinie?
Możliwe odpowiedzi:
$Y = 0$, $Y = 1$, $Y = 2$, $Y = 3$, rzadziej więcej.
Jest to zmienna o wartościach dyskretnych, ograniczona z dołu. Odpowiedzi $Y = 0$ jest na ogół najwięcej, $Y = 1$ nieco mniej itd. Można przyjąć, że zmienna Y ma rozkład Poissona.
Wniosek: możemy zastosować model regresji Poissona.
Co modelujemy? Wartość oczekiwaną zmiennej Y, to jest wartość oczekiwaną w rozkładzie Poissona. Szczegóły w rozdziale 7.

Przykład 1.5. Ile kosztował samochód, który Twoja rodzina kupiła w zeszłym roku?
Możliwe odpowiedzi: $Y = 0$ lub $Y > 0$.
Jest to zmienna o wartościach ciągłych nieujemnych z dużą liczbą zer. Z punktu widzenia ekonomii te dwie informacje ($Y = 0$ lub $Y > 0$) mówią o tym, czy rodzina wybrała brzegowe (*corner solution*), czy też wewnętrzne (*interior*) rozwiązanie zadania maksymalizacji użyteczności gospodarstwa domowego.
Wniosek: powinniśmy zastosować model tobitowy (model rozwiązań brzegowych).
Co modelujemy? Zmienną Y o warunkowym rozkładzie dyskretno-ciągłym, który jest połączeniem rozkładu ciągłego (wartości $Y > 0$) oraz rozkładu jednopunktowego ($Y = 0$). Szczegóły w rozdziale 6.

Inne przykłady modeli mikroekonometrycznych można znaleźć w pracy Gruszczyńskiego (2001, rozdział 1). Trzeba pamiętać, że w najszerszym ujęciu model mikroekonometryczny to w istocie każdy model typu regresyjnego oparty na mikrodanych. Jeśli szacujemy model, w którym płaca jest liniową funkcją charakterystyk danej osoby, to mamy na początku na myśli zwykły model liniowej regresji, w którym zmienna objaśniana jest zmienną ilościową (ciągłą). Dopiero próba, którą dysponujemy, a także adekwatność założeń modelu (np. kwestia endogeniczności) decydują o metodzie analizy.

1.4. Główne zagadnienia mikroekonometrii

1.4.1. Ekonomia a strategia modelowania w mikroekonometrii

Tradycyjne podejście do strategii modelowania w ekonometrii, w którym podstawą jest duże zaufanie do teorii ekonomicznej i jej postulowanych modeli, obecnie coraz częściej ustępuje miejsca podejściu aplikacyjnemu, w którym punktem wyjścia są dane. Heij i inni (2004) w nowoczesnym podręczniku do ekonometrii twierdzą, że

podstawą modelowania ekonometrycznego nie powinno być testowanie konkretnej teorii ekonomicznej, lecz wykorzystanie danych do lepszego zrozumienia interesującego nas zjawiska. Modele to jedynie konstrukcje, które mogą się zmieniać pod wpływem informacji, jakie niosą dane. Włączenie do modelu charakterystyk tych danych może prowadzić do lepszego zrozumienia opisywanego procesu ekonomicznego. Poza wszystkim, teoria ekonomii często nie podpowiada konkretnych postaci modeli, co pozwala na dość swobodne podejście do specyfikacji.

Takie rozmycie tradycyjnego podejścia do modelowania może nie być szczególnie pożądane w makroekonometrii, ale już ekonometria finansowa zajmująca się analizą szeregów czasowych danych o dużej częstości zdaje się zmierzać w kierunku tej drugiej strategii modelowania.

A jak jest w mikroekonometrii? Cameron i Trivedi (2005) uważają, że są dwa podejścia do korzystania z teorii ekonomii w mikroekonometrii. Pierwsze to podejście strukturalne, gdzie celem analizy jest identyfikacja i oszacowanie pewnych podstawowych parametrów, które charakteryzują badane zależności, np. funkcję kosztów czy funkcję produkcji. Przyjmuje się wówczas – na podstawie teorii ekonomii – wiele założeń dotyczących samej specyfikacji modelu czy też własności składników losowych. Jeśli posługujemy się danymi zagregowanymi, to oceny parametrów otrzymuje się przy znacznie silniejszych (niekoniecznie spełnionych) założeniach w porównaniu z korzystaniem z mikrodanych. Mikrodane pozwalają na większą elastyczność przy specyfikowaniu modelu. W drugim podejściu, które autorzy nazywają opartym na postaci zredukowanej, celem analizy jest modelowanie zależności pomiędzy zmiennymi wynikowymi (endogenicznymi) a zmiennymi, które uznaje się za egzogeniczne. Główną cechą tego podejścia jest to, że nie zawsze bierze się pod uwagę wszystkie współzależności między zmiennymi. Badanie ogniskuje się bowiem na predykcji zmiennej objaśnianej na podstawie zmiennych objaśniających. Nie chodzi o interpretację przyczynową parametrów modelu.

Naszym zdaniem umieszczenie modelu w nurcie określonej teorii ekonomicznej powinno być pierwszym celem badawczym. Mikrodane dają tu pewną przewagę nad danymi zagregowanymi. Nie zawsze jednak można posłużyć się solidną teorią. Z tego powodu wiele empirycznych modeli budowanych dla mikrodanych to modele oparte na dość luźnych, niekoniecznie spójnych hipotezach behawioralno-ekonomicznych.

Mikroekonometria, która korzysta z mikrodanych, daje też pewną przewagę nad podejściami korzystającymi z makrodanych. Cameron i Trivedi (2005) wskazują, że makroekonometria opiera się czasami na mocnych założeniach, np. na założeniu o jednostce reprezentatywnej (*representative agent*). Często dane zagregowane mają właśnie odzwierciedlać zachowanie takich jednostek, co nie zawsze ma sens. Z punktu widzenia mikroekonomii, analiza ilościowa oparta na mikrodanych może być uznana za bardziej realistyczną niż analiza bazująca na danych zagregowanych. Po pierwsze pomiar takich zmiennych jest na ogół bardziej bezpośredni i ma większe odniesienie do weryfikowanej właśnie teorii. Po drugie hipotezy dotyczące zachowania ekonomicznego zwykle są wyprowadzane z teorii behawioralnych. Przy

korzystaniu z danych zagregowanych trzeba wówczas dokonywać wielu aproksymacji i upraszczających założeń. Po trzecie realistyczny obraz aktywności gospodarczej powinien dopuszczać szerokie spektrum różnych wyników, które są konsekwencją naturalnej heterogeniczności jednostek i które można prognozować na podstawie danej teorii. Mikrodane zatem mogą opisywać modele, które są bardziej realistyczne niż modele oparte na makrodanych.

1.4.2. Założenia modelu regresji dla danych przekrojowych

Modele mikroekonometrii omawiane w podręcznikach koncentrują się na zmiennych jakościowych oraz na modelach nieliniowych. Nie znaczy to jednak, że prawidłowy opis zmiennej ilościowej za pomocą modelu liniowego w przypadku mikrodanych przekrojowych powinien pozostać niezauważony. W tym punkcie wskazujemy, jak formułuje się założenia dla klasycznego modelu regresji $y_i = x_i'\beta + \epsilon_i$, gdzie y_i jest i-tą obserwacją przekrojową dla zmiennej objaśnianej, natomiast x_i jest wektorem obserwacji dla zmiennych objaśniających.

Cameron i Trivedi (2005) proponują zestaw założeń dla klasycznego modelu regresji liniowej dla mikrodanych przekrojowych. Są to założenia dotyczące procesu generowania danych (*DGP – data generating process*). Na ogół rozumie się, że ***DGP*** jest zasadny w przypadku szeregów czasowych. Dla danych przekrojowych proces generowania danych jest definiowany rzadziej.

Podstawowe założenia (można je utożsamiać z założeniami dla metody najmniejszych kwadratów MNK) są następujące:

1. W modelu regresji y względem x dane (y_i, x_i) **są niezależne i nie mają identycznych rozkładów względem *i*** (*inid: independent and not identically distributed*). To założenie wyraźnie odrzuca przyjmowane w klasycznych tekstach „założenie o jednakowych wartościach zmiennych objaśniających w powtarzanych próbach". Ma ono sens jedynie w przypadku danych eksperymentalnych (gdzie x można utożsamiać z poziomem oddziaływania: *treatment*; zob. niżej). W mikroekonometrii mamy często do czynienia z próbami nielosowymi (co powoduje *inid*, w kontraście do prób losowych, które skutkują *iid: independent and identically distributed*). W parze (y_i, x_i) jedynie y_i jest zmienną losową o wartościach zależnych od wartości x_i. Zauważmy, że to założenie wyklucza dane w postaci szeregów czasowych, które są na ogół zależne względem i.

2. Postać modelu $y_i = x_i'\beta + \epsilon_i$ jest **prawidłowa** (*correct specification*). Co to znaczy? To założenie mówi o tym, że model jest liniowy względem x, że wektor x zawiera wszystkie potrzebne zmienne objaśniające oraz że nie ma błędu w zmiennych, to znaczy x_i są zmierzone poprawnie. Są to bowiem te same zmienne, które zarządzają *DGP*. Dodatkowo parametry β nie zmieniają swoich wartości względem i.

3. **Zmienne objaśniające mogą być stochastyczne**, przy czym mają skończoną wartość drugiego momentu. To założenie jest ważne zawsze wówczas, gdy korzystamy z danych pochodzących z badań ankietowych, a nie z wyników eksperymentu.

4. Warunkowa wartość oczekiwana błędów ϵ_i względem zmiennych objaśniających x_i jest równa zeru: $E(\epsilon_i|x_i) = 0$. To podstawowe założenie razem z założeniem 2 implikuje, że $E(y_i|x_i) = x_i'\beta$. Merytoryczne znaczenie tego założenia polega na tym, że wszystkie zmienne objaśniające, które nie są uwzględnione w modelu (czyli ϵ_i), nie są skorelowane ze zmiennymi x, to jest ze zmiennymi uwzględnionymi w modelu, a ponadto mają wartość oczekiwaną równą zeru. To jest założenie o **słabej egzogeniczności**. Jeśli nie jest ono spełnione, znaczy to, że jedna ze zmiennych objaśniających (egzogenicznych) lub ich większa liczba jest zależna od y, a zatem jest endogeniczna.

5. Błędy ϵ_i są heteroskedastyczne, warunkowo względem zmiennych objaśniających, przy czym $Var(\epsilon_i|x_i) = \sigma_i^2$, natomiast macierz warunkowej wariancji i kowariancji wektora błędów ϵ_i jest diagonalna, co zapisujemy jako $Diag\{\sigma_i^2\}$. To założenie mówi o **warunkowo heteroskedastycznych błędach modelu regresji**.

Powyższy zestaw założeń można przyjąć jako standard dla stosowania MNK w modelach liniowych opartych na danych przekrojowych. Każde odejście od założeń daje konsekwencje w postaci niepożądanych własności estymatorów (MNK i podobnych).

Klasyczna MNK nie okazuje się w tym przypadku rozwiązaniem najlepszym (najefektywniejszym), głównie ze względu na założenie 5. Uogólniona MNK (lub ważona MNK) są teoretycznie bardziej poprawne. Cameron i Trivedi (2005) proponują też korzystanie z **regresji kwantylowej** opartej na innej niż MNK funkcji straty.

1.4.3. Korelacja a przyczynowość. Relacje przyczynowe a analiza *ceteris paribus*

Przy specyfikacji modeli z wieloma zmiennymi często posługujemy się współczynnikami korelacji. Trzeba jednak zawsze pamiętać, że współczynnik korelacji mierzy asocjację statystyczną między zmiennymi. Jeśli nie ma dla tej asocjacji jakiegoś mocnego uzasadnienia teoretycznego, to korelacja między Y i X jest po prostu miarą opisującą, na ile wspólnie zmieniają się Y i X. Bardzo rzadko na tej podstawie można uznać, że mamy do czynienia z przyczynowością, to znaczy: X „jest przyczyną" Y. W ekonomii trudno nawet znaleźć bezdyskusyjny przykład. Znajdziemy go najwyżej w ekonomii eksperymentalnej. Nieco łagodniejsza w tym względzie może być opinia o korelacji cząstkowej, „po usunięciu wpływu" innych zmiennych.

Co z tego wynika? Po pierwsze to, że wielkość współczynnika korelacji nie powinna być utożsamiana z poziomem przyczynowości. Po drugie odpowiedzi na pytanie

o przyczynowość trzeba poszukiwać w inny sposób. Wooldridge (2002) w monografii na temat mikroekonometrii zwraca uwagę na pojęcie *ceteris paribus*, które można tłumaczyć jako: „zakładając niezmieniony poziom innych czynników". Za Wooldridgem przyjmijmy, że pierwotnym celem naszych dociekań jest ustalenie, czy zmiana wartości jednej zmiennej, na przykład w, powoduje zmianę wartości innej zmiennej, na przykład y. Wówczas, jeśli koncentrujemy się na wartości przeciętnej zmiennej y, to analiza *ceteris paribus* oznacza, że zależy nam na oszacowaniu $E(y|w, \mathrm{c})$, czyli wartości oczekiwanej zmiennej y, warunkowej względem w i c. Wektor c oznacza tutaj zbiór zmiennych kontrolnych, które chcielibyśmy traktować jako ustalone przy analizie wpływu w na y. Zmienne kontrolne są wyróżnione dlatego, iż uważamy, że w jest zmienną skorelowaną także z innymi czynnikami, które wpływają na y. Jeśli w jest zmienną ciągłą, to interesujemy się wielkością $\partial E(y|w, \mathrm{c})/\partial w$, którą nazywa się cząstkowym wpływem w na $E(y|w, \mathrm{c})$. Jeśli w jest zmienną skokową, to wtedy trzeba ustalić, jak $E(y|w, \mathrm{c})$ zmienia się dla różnych wartości w, przy ustalonych c.

Nie jest łatwo ustalić, które zmienne należy traktować jako kontrolne (c). Błąd w tej mierze oznacza możliwość ustalenia błędnej relacji między y i w. Jeśli c można obserwować, to pozostaje nam wybór właściwego c. Jednak w ekonomii wiele elementów c nie daje się obserwować. Wooldridge podaje znany przykład problemu polegającego na badaniu wpływu wykształcenia na poziom płacy za pomocą modelu:

E(płaca | wykształcenie, doświadczenie, zdolności),

czyli modelu z wektorem c = (*doświadczenie, zdolności*). W tym wektorze zmienna *doświadczenie* jest obserwowalna (np. za pomocą liczby lat pracy), natomiast zmienna *zdolności* jest nieobserwowalna. Ekonomiści zajmujący się rynkiem pracy są zgodni co do tego, że aby określić wpływ wykształcenia na poziom płacy, należy pozostawić takie zmienne jak doświadczenie i zdolności na niezmienionym poziomie, czyli zastosować analizę *ceteris paribus*. Pozostaje kwestia, jak to zrobić, jeśli jakiejś z tych zmiennych nie da się obserwować. Ponadto jest z pewnością wiele innych zmiennych w wektorze c, których tu nie uwzględniliśmy. Pytanie, jakie się stawia w badaniach empirycznych, polega na tym, czy wzięto pod uwagę dostatecznie dużo zmiennych w wektorze c, aby można było dobrze ocenić wpływ zmiennej w na zmienną y. Jeszcze inny problem pojawia się wtedy, kiedy trudno zmierzyć wartości głównych zmiennych, to jest y oraz w. Czasami obserwujemy jedynie wartości równowagi dla y i w, to jest wówczas, gdy są one wzajemnie współzależne. Niezbędne jest wówczas posługiwanie się narzędziami właściwymi dla współzależnych modeli wielorównaniowych.

Powyższy wywód, który pochodzi od Wooldridge'a (2002), wykazuje, że typowe podejście do interpretacji parametru regresji z użyciem pojęcia *ceteris paribus* powinno być stosowane z dużą ostrożnością. Kwestia zasadności analiz *ceteris paribus* jest podejmowana przez ekonomię i ekonometrię właściwie od początków istnienia tych dyscyplin. Warto w tym względzie wskazać interesujący artykuł Bierensa i Swansona (2000), a także (w kontekście ekonometrii szeregów czasowych) artykuł Hendry'ego (2005).

Badanie przyczynowości w mikroekonometrii znajduje pewne rozwiązanie w analizie efektów oddziaływania.

1.4.4. Efekty oddziaływania

Efekty oddziaływania (*treatment effects*), podobnie jak opisane niżej endogeniczność i heterogeniczność, należą do ważnych tematów współczesnej mikroekonometrii. Chodzi o to, że dostępność mikrodanych pozwala na studiowanie, jak pewne oddziaływania (*treatment*) wpływają na wartości zmiennych endogenicznych, to jest na jednostki, które są tym oddziaływaniom poddane i oglądane poprzez te właśnie endogeniczne zmienne.

Teoria efektów oddziaływania wykorzystuje pojęcie **wyniku hipotetycznego** (*counter-factual*), który w rzeczywistości nie może się zdarzyć. Poniższe omówienie pochodzi z książki M. Gruszczyńskiego (2012).

Dla przykładu wyobraźmy sobie decyzję firmy o wprowadzeniu akcji do obrotu publicznego (*Initial Public Offering*: IPO), a także efekt tej decyzji w postaci jej wpływu na wynik finansowy firmy, mierzony na przykład w rok po IPO za pomocą wskaźnika ROE rentowności kapitału własnego. Przedsięwzięcie w postaci IPO jest w tym przypadku oddziaływaniem. Aby oszacować efekt tego oddziaływania dla próby firm należałoby znać dwie wielkości ROE dla każdej firmy: ROE z IPO oraz ROE bez IPO, to znaczy wynik z oddziaływaniem i wynik bez oddziaływania. Wtedy (przyczynowy) efekt oddziaływania jest po prostu różnicą tych dwóch wyników. Jednakże dana firma może występować w próbie tylko w jednej roli: albo faktycznie wprowadza akcje do obrotu publicznego, albo nie wprowadza i pozostaje firmą niepubliczną. Zatem dla danej firmy obserwujemy tylko jeden wynik, drugi jest hipotetyczny („kontrfaktyczny"). Przyjmijmy, że wprowadzanie akcji i-tej firmy do obrotu publicznego jest reprezentowane przez losową zmienną zerojedynkową $D_i\{0,1\}$, gdzie 1 oznacza, że firma przeprowadza IPO, natomiast 0 oznacza, że nie przeprowadza IPO. Wynik, który nas interesuje, to znaczy ROE, zapiszmy jako y. Pytanie brzmi: czy IPO ma wpływ na y? Potencjalny wynik dla i-tej firmy jest następujący:

$$\text{potencjalny wynik} = \begin{cases} y_{1i} & \text{jeżeli} \quad D_i = 1 \\ y_{0i} & \text{jeżeli} \quad D_i = 0 \end{cases}$$

To znaczy, że y_{0i} oznacza wartość ROE dla i-tej firmy, przy założeniu, że nie przeprowadza ona IPO – bez względu na to, czy faktycznie tak się dzieje. Podobnie, y_{1i} oznacza wartość ROE dla i-tej firmy, przy założeniu, że przeprowadza ona IPO – bez względu na to, czy faktycznie tak się dzieje. Efekt oddziaływania dla i-tej firmy wynosi po prostu: $y_{1i} - y_{0i}$. Obserwowany w rzeczywistości wynik dla i-tej firmy jest następujący:

$$y_i = \begin{cases} y_{1i} & \text{jeżeli} \quad D_i = 1 \\ y_{0i} & \text{jeżeli} \quad D_i = 0 \end{cases}$$

$$= y_{0i} + (y_{1i} - y_{0i})D_i = D_i y_{1i} + (1 - D_i)y_{0i}$$

Średnia obserwowana różnica w wartościach ROE (dla firm z IPO i bez IPO) równa się:

$$ATE = E(y_i|D_i = 1) - E(y_i|D_i = 0).$$

Jest to średni efekt oddziaływania (*ATE: average treatment effect*). Ta wielkość oczywiście równa się $E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 0)$, co można zapisać jako:

$$E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 0) =$$

$$E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 1) + E(y_{0i}|D_i = 1) - E(y_{0i}|D_i = 0).$$

Po prawej stronie powyższego wzoru mamy sumę dwóch różnic. Pierwsza z nich nosi nazwę średniego efektu oddziaływania na jednostkę (*ATT: average treatment effect on the treated*):

$$ATT = E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 1) = E(y_{1i} - y_{0i}|D_i = 1)$$

i mówi o tym, jaki jest przyczynowy efekt IPO w firmach, które faktycznie przeprowadzają IPO. Jest to różnica pomiędzy wartością ROE dla firm, które przeprowadziły IPO, czyli $E(y_{1i}|D_i = 1)$ oraz wartością ROE dla tych samych firm, przyjmując (hipotetycznie), że IPO nie przeprowadziły, to znaczy $E(y_{0i}|D_i = 1)$. Wielkość ATT to jest właśnie to, co chcielibyśmy wiedzieć. Jednak to, co obserwujemy, czyli lewa strona równania, jest różne od ATT o wartość drugiego składnika po prawej stronie, który można nazwać **obciążeniem selekcji**. Jest to wielkość:

$$E(y_{0i}|D_i = 1) - E(y_{0i}|D_i = 0)$$

czyli różnica pomiędzy średnią wartością y_{0i} dla firm, które przeprowadziły oraz nie przeprowadziły IPO. Przypomnijmy, że y_{0i} to jest wartość ROE dla firmy nieprzeprowadzającej IPO. Obciążenie selekcji jest więc różnicą hipotetyczną, podobnie jak ATT. Nie wiadomo też, jaki jest znak tego obciążenia: czy firmy, które faktycznie przeprowadziły IPO mają hipotetyczne ROE (to jest takie bez przeprowadzenia IPO) większe czy mniejsze od ROE dla firm, które faktycznie nie przeprowadziły IPO. W każdym razie obciążenie może być na tyle duże, iż znacząco zakłóci wartość ATT. To, co faktycznie możemy obliczyć we wzorze:

$$ATE = ATT + \textit{obciążenie selekcji},$$

to lewa strona, czyli ATE.

Pytanie, czy można pozbyć się obciążenia selekcji? Można – przede wszystkim wtedy, gdy zmienna losowa D_i od początku jest niezależna od wyników y_i. W przypadku losowego doboru próby (lub zrandomizowanego eksperymentu) oddziaływanie i jego wynik są niezależne. (W naszym przykładzie chodziłoby o losowy wybór firm do IPO; w rzeczywistości nie jest to możliwe). Wówczas we wzorze:

$$ATE = E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 0)$$

można zastąpić średnią $E(y_{0i}|D_i = 0)$ przez $E(y_{0i}|D_i = 1)$, bowiem D_i jest niezależna od y_{0i}. W rezultacie otrzymujemy $ATE = ATT$. Oznacza to, że efekt losowo wybranego oddziaływania IPO na firmę, która przeprowadza IPO jest taki sam jak efekt losowo wybranego oddziaływania IPO na losowo wybraną firmę. Jednakże „oddziaływania", o jakich mowa, oznaczają dla firmy całkiem nielosowe decyzje (np. właśnie IPO) lub sytuacje (np. obiekt przejęcia przez inną firmę). Zatem firmy „z oddziaływaniem" są wyraźnie różne od firm „bez oddziaływania". Nie jest możliwe utrzymanie założenia o niezależności D_i od y_i. Jednak i teraz obciążenie selekcji może być równe zeru, jeśli można przyjąć pewne szczególne założenia. Jakie to założenia? O tym jest między innymi mowa w rozdziale 9 naszej książki.

Rozdział 9 przedstawia kompleksowo tematykę efektów oddziaływania. Warto pamiętać, że badania z tego zakresu były podstawą Nagrody Nobla dla J. Heckmana w roku 2000. Estymacja efektów oddziaływania należy do szybko rozwijających się gałęzi mikroekonomii empirycznej, określanych jako metody oceny polityki społecznej (por. Blundell i Costa Dias, 2008). Znana jest też ogólniejsza nazwa *evaluation econometrics* (metody ekonometrii „ewaluacyjnej").

1.4.5. Endogeniczność

Endogeniczność oznacza, że zmienna objaśniająca *X* w modelu wyjaśniającym zmienną *Y* jest skorelowana ze składnikiem losowym. To znaczy, że zmienna X nie objaśnia zmiennej Y z zewnątrz modelu, lecz objaśnia ją „w porozumieniu" (w korelacji) ze składnikiem losowym. Zmienna X jest w istocie „endogeniczna", to znaczy objaśniają ją czynniki zawarte w składniku losowym, podobne do tych, które objaśniają zmienną Y. Co więcej, składnik losowy nie jest wtedy naprawdę losowy, ponieważ jest częściowo przewidywalny na podstawie informacji zawartej w zmiennej X. Kwestia endogeniczności jest podstawą rozważań w wielorównaniowych modelach współzależnych. Problem występuje także w mikroekonometrii. W kryminologii znany jest wynik estymacji modelu regresji nasilenia przestępczości w powiatach względem wielu zmiennych, w tym liczby policjantów *per capita*, gdzie oszacowany parametr regresji jest dodatni („im więcej policji, tym większa przestępczość"). Wśród różnych przyczyn takiego wyniku można wymienić i taką, że liczba policjantów jest endogeniczna w równaniu: taka zmienna może być na przykład skorelowana z wielkością powiatu.

Dobry intuicyjny przykład dotyczący modeli mikroekonometrycznych pochodzi z książki Camerona i Trivediego (2009). Jest on zresztą wykorzystywany także w innych podręcznikach. Załóżmy, że zmienna x oznacza lata nauki, zmienna y oznacza zarobki danej osoby, a ϵ jest składnikiem losowym w modelu liniowym, w którym y jest zmienną objaśnianą, a x zmienną objaśniającą. Przy klasycznych założeniach MNK x nie jest skorelowana z ϵ. Wtedy zwykły MNK-estymator β dla parametru regresji y względem x jest estymatorem zgodnym.

Pamiętajmy, że ϵ zawiera wszystkie czynniki, które wpływają na y, a nie są uwzględnione w x. Jednym z nich są zdolności danej osoby. Są one jednak skorelowane ze zmienną x: wiemy, że większe zdolności idą w parze z większą liczbą lat nauki. Z tego wynika, że zwykły estymator MNK dla β nie jest zgodny: ten estymator zawiera bowiem zarówno efekt bezpośredni liczby lat nauki względem zmiennej y, jak i efekt pośredni – poprzez składnik losowy ϵ. Osoby z dużą liczbą lat nauki mają także prawdopodobnie duże zdolności, dużą wartość ϵ, a w konsekwencji dużą wartość y.

Jeśli oszacujemy model za pomocą MNK i otrzymamy oszacowanie β na poziomie 400 zł (co oznacza przyrost miesięcznych zarobków na 1 rok nauki), to nie wiemy, jaka część tej liczby wynika z wpływu samej nauki (x), a jaka – z wpływu zdolności (ϵ). Rozwiązaniem problemu byłoby wprowadzenie obok x zmiennej reprezentującej „zdolności". Na ogół nie jest to jednak możliwe. Wtedy jedynym wyjściem jest poszukanie odpowiedniego „instrumentu", to jest zmiennej z, która nie jest skorelowana z ϵ, a która jest bezpośrednio skorelowana z x. W tym przykładzie taką zmienną może być odległość od miejsca zamieszkania do wyższej uczelni. Taka zmienna z może mieć wpływ na liczbę lat nauki (x) i jednocześnie bezpośrednio nie wpływa na wartość y. Odpowiedni estymator metody zmiennych instrumentalnych dla parametru β jest estymatorem zgodnym. Testowanie endogeniczności i metoda zmiennych instrumentalnych są omówione w wielu podręcznikach ekonometrii, na przykład Wooldridge'a (2002, 2003).

1.4.6. Heterogeniczność

Korzyści z dezagregacji danych, czyli korzyści z mikrodanych, są w jakimś sensie okupione potrzebą uwzględniania niejednorodności danych w analizach. Niejednorodność (heterogeniczność), a właściwie nieobserwowana niejednorodność, odgrywa ważną rolę w mikroekonometrii. Wiele zmiennych, które mają własność niejednorodności, można obserwować, np. płeć, wykształcenie, czynniki socjodemograficzne. Inne, na przykład motywacja, zdolności, inteligencja itd. są nieobserwowalne bądź nie dają się dobrze obserwować.

Najprostszym rozwiązaniem jest zignorowanie takiej niejednorodności i włączenie jej do składnika losowego. To jednak zwiększa niewyjaśnioną w modelu część zmienności zmiennej endogenicznej (objaśnianej). W istocie ignorowanie wyraźnych różnic indywidualnych prowadzi do pomylenia z innymi czynnikami, które też są źródłem wyraźnych różnic indywidualnych. To pomylenie występuje wtedy, kiedy nie da się

statystycznie wydzielić indywidualnego udziału każdej ze zmiennych objaśniających w zmienności zmiennej objaśnianej. Załóżmy, że np. wykształcenie (x_1) ma być źródłem zmienności zarobków (y), a z kolei inna zmienna: zdolności (x_2), która też jest źródłem zmienności y, nie występuje w modelu. Wtedy ta część zmienności, która pochodzi od drugiej zmiennej, jest nieprawidłowo przypisana pierwszej zmiennej. Ich względna ważność w wyjaśnianiu y ulega zatem pomieszaniu. Powstające obciążenie (*confounding bias*) ma w tym przypadku źródło w nieprawidłowym pominięciu zmiennych objaśniających w modelu. Może także być wynikiem wstawienia w ich miejsce zmiennych zastępczych (szczegóły podają Cameron i Trivedi, 2005, s. 9).

Jest kilka podejść do modelowania niejednorodności w mikroekonometrii. Całkowite zignorowanie tego zjawiska, czyli nieobserwowalnych różnic indywidualnych, często spotykane w prostych zastosowaniach (takich, jak opisywane w tej książce), jest uzasadnione jedynie przy przyjęciu mocnych założeń. Należy do nich nieskorelowanie niejednorodności nieobserwowalnej z obserwowalną, a także brak zależności międzyokresowych dla zmiennej, którą się opisuje (zmiennej endogenicznej).

Heterogeniczność można traktować jako efekt stały (*fixed effect*) i ujmować ją za pomocą parametru przy zmiennej zerojedynkowej odnoszącej się do danej jednostki. Taka możliwość istnieje, jeśli dysponujemy danymi panelowymi, czyli licznymi obserwacjami dla każdej jednostki. Jednostka otrzymuje wtedy własną zmienną zerojedynkową, czyli w istocie własny wyraz wolny w modelu.

Z kolei nieobserwowana niejednorodność w modelu z efektami losowymi (*random effects*) może być ujęta na różne sposoby. Jedno z założeń głosi np., że wyraz wolny w modelu regresji zmienia się losowo względem przekroju. Według innego z założeń składnik losowy równania regresji jest sumą kilku składników, w tym składnika losowego odnoszącego się do pojedynczej jednostki. Celem jest estymacja parametrów rozkładu tego składnika losowego. Np. w analizie popytu ten składnik losowy interpretuje się jako losową zmienność preferencji. Modele z efektami losowymi można szacować na podstawie danych panelowych lub danych przekrojowych.

Powyższy opis tematu niejednorodności pochodzi od Camerona i Trivediego (2005). Browning i Carro (2006) wskazują, że we współczesnych badaniach mikroekonometrycznych heterogeniczność nie jest traktowana z należytą starannością. Na ogół heterogeniczności jest więcej niż zakłada się w badaniach, a ponadto nieprawidłowe ujęcie heterogeniczności może prowadzić do błędów przy szacowaniu efektów, które interesują badaczy.

Przykład z mikroekonometrii finansowej (endogeniczność i heterogeniczność) (Gruszczyński 2012).

W badaniach na temat związku między poziomem nadzoru właścicielskiego ($X = CG$) i rozmaitymi kategoriami, na przykład wynikami firmy ($Y = Wynik$) można mówić o wzajemnej przyczynowości (por. Brown, Beekes i Verhoeven 2011). W istocie, zarówno CG (*corporate governance*) jak i $Wynik$ są potencjalnie skorelowane ze zmiennymi, które reprezentuje składnik losowy. W modelu, w którym $Wynik$ jest

zmienną objaśnianą, a zmienną objaśniającą jest CG, składnik losowy reprezentuje zmienne, które wpływają na $Wynik$ i nie zostały do modelu włączone. Na wyniki firmy ($Y = Wynik$) wpływa wiele elementów, nie tylko poziom nadzoru ($X = CG$). Ważna jest koniunktura gospodarcza, sytuacja na rynkach finansowych, innowacyjność produktów firmy, zdolność konkurowania w sprzedaży, zdolność dostosowania się do zmieniających się warunków rynkowych itd. Niektóre z wymienionych zmiennych wpływają także na zmienną $X = CG$, bowiem lepsza, „zdolniejsza" firma ma na ogół lepszy poziom nadzoru. Te „zdolności" są na ogół niemierzalne. Dla przykładu, w modelu liniowym:

$$Wynik_i = \alpha + \beta CG_i + \gamma Z_i + \epsilon_i$$

$Wynik_i$ oznacza wartość wskaźnika ceny do wartości księgowej (P/BV) dla i-tej firmy, CG_i oznacza wartość indeksu nadzoru właścicielskiego dla tej firmy, Z_i to symbol oznaczający odpowiednie zmienne kontrolne typu wielkość firmy mierzona kapitalizacją rynkową itd. Endogeniczność CG_i oznacza, że ta zmienna jest skorelowana z ϵ_i. Podobną sytuację skorelowania Y ze składnikiem losowym mamy w przypadku nieobserwowanej heterogeniczności. Jeśli wyraźnie dopuszczamy, żeby wartość oczekiwana zmiennej objaśnianej Y była zależna od zmiennych nieobserwowanych (losowych), wówczas takie zmienne nazywa się nieobserwowaną heterogenicznością (zmiennej Y). W tym przykładzie chodzi o heterogeniczność związaną z oddziaływaniem poszczególnych firm. Prawidłowe wzięcie pod uwagę tego efektu musiałoby wiązać się z wykorzystaniem panelu. Wówczas mielibyśmy:

$$Wynik_{i,t} = \alpha + \beta CG_{i,t} + \gamma Z_i + \delta_i + \epsilon_{i,t}$$

gdzie t oznacza numer okresu, natomiast δ_i jest nieobserwowanym efektem (stałym w czasie). Taki model nosi nazwę modelu z efektami stałymi (lub z efektami nieobserwowanymi), a parametr δ_i nazywa się właśnie nieobserwowaną heterogenicznością. Tutaj wyraźnie zakładamy, że wartość oczekiwana zmiennej $Wynik_i$ zależy od δ_i, czyli od „nieobserwowanej heterogeniczności" wynikającej z wpływu poszczególnych firm. W obu powyższych modelach zmienna CG jest skorelowana ze składnikiem losowym. W pierwszym zmienna CG_i jest skorelowana bezpośrednio z ϵ_i, natomiast w drugim zmienna $CG_{i,t}$ jest skorelowana z $\delta_i + \epsilon_{i,t}$. Rozwiązaniem problemu jest użycie metody zmiennych instrumentalnych (MZI), która zapewnia zgodność estymatorów. Chodzi o dobranie zmiennych (instrumentów), które nie są skorelowane z ϵ, a które są skorelowane z CG.

Metoda zmiennych instrumentalnych polega na odrębnej estymacji równania, w którym zmienna „podejrzewana" o endogeniczność (tutaj: CG) jest objaśniana za pomocą instrumentów oraz pozostałych zmiennych egzogenicznych. Oszacowaną w tym pierwszym równaniu wartość CG („wartość teoretyczną") należy wykorzystać jako zmienną objaśniającą w miejsce CG w oryginalnym równaniu i równanie to oszacować. Ta procedura jest tożsama z podwójną MNK (2MNK) znaną dla modeli o równaniach współzależnych. Dobór instrumentów nie jest łatwy. Mają to być zmienne skorelowane

z naszą zmienną CG i jednocześnie nieskorelowane ze składnikiem losowym. Znaczy to, że nie mogą to być zmienne pominięte w równaniu objaśniającym zmienną $Wynik$, czyli nie powinny być skorelowane z tą zmienną. W praktyce takie zmienne trudno znaleźć w „czystej" postaci. Zmienne proponowane jako instrumenty mogą słabo objaśniać zmienną endogeniczną (*weak instruments*), a wtedy sytuacja jest nawet gorsza, niż gdyby pozostać przy MNK bez uciekania się do MZI. Zasadność stosowania MZI można testować za pomocą testu Hausmana lub testu Durbina-Wu-Hausmana.

1.4.7. Skokowość, nieliniowość, zawartość informacyjna zbiorów mikrodanych

Cameron i Trivedi (2005) wskazują, że ze względu na niski poziom agregacji mikrodanych modele liniowe nie są najwłaściwsze. Jeśli dane są zagregowane, to zależności pomiędzy zmiennymi reprezentującymi takie dane są z reguły „gładsze" niż zależności pomiędzy mikrodanymi. Agregowanie danych prowadzi do wygładzania. Na przykład przeciętne tygodniowe wydatki na kulturę w rodzinie są na ogół dość wyraźną funkcją przeciętnych dochodów. Jeśli jednak bierzemy pod uwagę dane indywidualne, to zauważamy, że wydatki te są często równe zeru, a związek z dochodem nie jest już tak wyraźny dla mikrodanych jak dla danych zagregowanych.

Dane jednostkowe charakteryzują się bardzo dużą zmiennością – pokazują różnorodność populacji, z której się wywodzą. Mamy więc w tych danych „dziury, węzły i narożniki", jak to ujmuje Pudney (1989). „Dziury" to nieuczestnictwo jednostek w opisywanym przez dane procesie. Jeśli na przykład interesująca nas zmienna oznacza tygodniowe spożycie czerwonego wina, to jej wartość równa zeru jest właśnie taką „dziurą". Z kolei „węzły" to zmieniające się zachowanie jednostki. Zmienna „liczba godzin uczestnictwa w rynku pracy w danym roku" pokazuje dla wielu osób, w trakcie ich kariery, przestawianie się z wartości dodatnich na zera i z powrotem. „Narożniki" opisują nieuczestnictwo w konkretnym momencie czasu. Czyli skokowość i nieliniowość obserwowanych efektów to istotna cecha mikroekonometrii.

Kolejna cecha zbiorów mikrodanych to ich możliwe wewnętrzne „skłócenie" (*noise*). O wartości jakiejś zmiennej dla pojedynczej jednostki mogą bowiem decydować czynniki bardzo indywidualne (idiosynkratyczne), które są nieobserwowalne i które są wówczas reprezentowane przez składnik losowy równania wyjaśniającego tę zmienną. Wtedy ta losowa część równania ma duży udział w wyjaśnianiu zmienności. W tym sensie składnik losowy odgrywa przy mikrodanych większą rolę niż przy makrodanych. Z tego właśnie powodu w równaniach szacowanych dla danych indywidualnych (przekrojowych) wartości współczynnika determinacji R-kwadrat są niskie.

1.4.8. Zbieranie danych

Na początku rozdziału wskazaliśmy, że mikrodane są na ogół danymi obserwacyjnymi, rzadziej pochodzą z tak zwanych eksperymentów społecznych. Źródłem danych

obserwacyjnych są rozmaite bazy danych lub badania ankietowe, sondaże. Bardzo liczne bazy danych dotyczące całych populacji (np. dane ze spisu powszechnego, bazy REGON czy PESEL) mogą być podstawą do generowania prób danych obserwacyjnych. Za Cameronem i Trivedim (2005) przedstawimy hasłowo tematykę zbierania prób danych obserwacyjnych. Ta tematyka jest centralna dla ustalania, na ile próba reprezentuje populację. Literatura z badań reprezentacyjnych jest bardzo obfita. Tutaj ten problem jedynie sygnalizujemy.

Próby losowe i złożone

W przypadku **prostej próby losowej** (*simple random sample*) każdy element populacji („generalnej") ma jednakowe prawdopodobieństwo dostania się do próby. W tym przypadku indywidualne jednostki badania są wybierane do próby bezpośrednio. Przy dużych populacjach losowanie to nie jest jednak wygodne. Nie jest także prawdą, że takie losowanie jest najlepsze z uwagi na dokładność otrzymywanych wyników. Prosta próba losowa nie bierze bowiem pod uwagę cech jednostki losowania, a te właśnie mogą stanowić podstawę dobrego doboru próby. Przywołajmy tu formalnie pojęcie charakterystyki jednostki, czyli jej fotografię w wektorze zmiennych, z których jedna jest zmienną zależną y (zmienną wynikową: *response, outcome variable*), a pozostałe są zmiennymi egzogenicznymi x. Funkcja gęstości ich łącznego rozkładu ma postać $f_J(y, x)$. Można ją zapisać jako iloczyn funkcji gęstości rozkładu warunkowego $f_C(y|x, \theta)$ oraz rozkładu brzegowego $f_M(x)$, to jest:

$$f_J(y, x) = f_C(y|x, \theta) f_M(x).$$

Próba losowa prosta oznacza losowanie z kombinacji (y, x) w jednakowy sposób z całej populacji, bez brania pod uwagę tych charakterystyk, które akurat są ważne dla podjętego badania.

Ze względu na fakt, że próba losowa prosta jest trudna do realizacji oraz na to, że jest to swego rodzaju losowanie „ślepe" z punktu widzenia celu losowania, korzysta się częściej z losowania złożonego, np. z **wielostopniowego zespołowego losowania warstwowego** (*stratified multistage cluster sampling*). Jednostki populacji w losowaniu wielostopniowym są oglądane z takich perspektyw jak warstwy (rozłączne podpopulacje składające się na całą populację), jednostki losowania pierwszego stopnia (rozłączne podzbiory warstw), jednostki losowania drugiego stopnia (części jednostek pierwszego stopnia) i tak dalej, aż do indywidualnych jednostek badania. Warstwami mogą być np. powiaty, jednostkami pierwszego stopnia – gminy, a indywidualnymi jednostkami badania – gospodarstwa domowe. Na ogół losuje się z wszystkich warstw (powiatów). Ale już nie wszystkie jednostki pierwszego stopnia (gminy) są losowane z warstwy (powiatu). Przy tym mogą być losowane z różnymi prawdopodobieństwami wyboru – w zależności od charakterystyki próby, która nas interesuje. W losowaniu dwustopniowym z wylosowanych jednostek pierwszego stopnia wybiera się losowo indywidualne jednostki badania (gospodarstwa).

W efekcie przy tym schemacie losowania różne gospodarstwa domowe mają różne prawdopodobieństwa wyboru do próby. To oznacza, że próba nie jest reprezentantką

populacji. Wówczas można posłużyć się odpowiednimi **wagami** dla danego losowania, odwrotnie proporcjonalnymi do prawdopodobieństw wyboru do próby. Na podstawie próby można wówczas otrzymać dobre oceny charakterystyk całej populacji.

Szersza ekspozycja tematyki schematów losowania nie jest tu zamierzona. Przydatne praktycznie rozważania na temat technik sondaży przedstawia Szreder (2004). Informacje o samej metodzie reprezentacyjnej (*sampling*) można znaleźć w pracach Brachy (1996, 1998).

Próby obciążone

W wielu praktycznych zastosowaniach, np. w mikroekonometrii finansowej, losowanie opisane wyżej nie jest możliwe. Mamy raczej do czynienia z próbami obciążonymi, to jest takimi, dla których po prostu rozkład prawdopodobieństwa jest różny od rozkładu w populacji (generalnej). Ta różnica wynika ze sposobu zbierania danych, z tego, na ile różni się on od próby czysto losowej.

Istnieją różne odstępstwa od czysto losowego schematu losowania. Przy zbieraniu danych obserwacyjnych często stosuje się **losowanie egzogeniczne** (*exogenous sampling*). Dostępną próbę dzieli się na segmenty odpowiadające wyłącznie zmiennym egzogenicznym. Cameron i Trivedi (2005) uważają, że jest to w istocie losowanie wtórne (*subsampling*), dotyczy bowiem już wylosowanej (dostępnej) próby. Bardzo często dzielimy dostępną próbę według płci, statusu itd. W przypadku losowania egzogenicznego zakłada się, że rozkład prawdopodobieństwa zmiennych egzogenicznych x jest niezależny od zmiennej y i nie zawiera informacji o parametrach populacji θ. Można wówczas zignorować rozkład brzegowy zmiennych egzogenicznych i opierać swoje wnioskowanie po prostu na rozkładzie warunkowym $f_C(y|x,\theta)$. Tego rodzaju założenie może być jednak nieprawdziwe. Obserwowany rozkład zmiennej zależnej może zależeć od wyróżnionej egzogenicznej zmiennej segmentującej populację. Wtedy taka zmienna może być skorelowana ze zmienną wynikową (zależną), co spowoduje odstępstwo od zasady losowania egzogenicznego.

Próbę obciążoną daje też **losowanie względem zmiennej zależnej** (*response-based sampling*). Ma to miejsce wtedy, kiedy prawdopodobieństwo dostania się danego przypadku do próby zależy od podjętej w tym przypadku decyzji czy też od efektu, który się ujawnił. Dana obserwacja pojawia się w próbie wówczas, gdy jednostka podejmuje określoną decyzję, np. gdy korzysta z publicznego transportu, a my pytamy ją o determinanty wyboru tego lub innego środka transportu. Ci, którzy nie korzystają akurat z transportu publicznego, nie dostają się do próby.

Tego rodzaju losowanie stosuje się dlatego, że kosztuje mniej w porównaniu z losowaniem prostym. Otrzymanie w losowej próbie odpowiedniej reprezentacji dla rzadko zdarzających się wyników (wyborów) może być bardzo kosztowne. Lepiej jest skonstruować próbę z tych jednostek, które takich wyborów dokonały. Praktycznym wyrazem tego rodzaju próby jest to, że nie można na podstawie warunkowej funkcji gęstości $f_C(y|x,\theta)$ otrzymać zgodnego estymatora parametrów θ. Wówczas należy wziąć pod uwagę sam schemat losowania.

Próby nielosowe typu ***choice-based sample*** są w mikroekonometrii codziennością. Można je nazwać **próbami dobieranymi**. Na ogół są to **próby niezbilansowane** (*disproportionate samples*), czyli takie, dla których udziały jednostek poszczególnych kategorii w próbie są niejednakowe[2]. W przypadku modeli zmiennej dwumianowej może to być na przykład wybór do próby wszystkich dostępnych firm upadłych oraz tylko niektórych firm nieupadłych. Wówczas jednostki pierwszego rodzaju (bankruci) są reprezentowane w próbie w 100%, a jednostki drugiego rodzaju (nie-bankruci) są reprezentowane w próbie na przykład na poziomie 5% wszystkich jednostek drugiego rodzaju. W rozdziale 3 omawiany jest przypadek estymacji dwumianowego modelu logitowego dla takiej próby.

Z kolei sprawa obciążenia ze względu na dobór próby (*sample selection*) jest omawiana w rozdziale 6 tej książki. **Modele doboru próby** (selekcji próby) obejmują szeroką klasę przypadków. Można np. mówić o **samoselekcji** (lub o autoselekcji), kiedy jednostka „sama włącza się do próby" poprzez decyzję uczestnictwa w jakimś działaniu (np. bycia zatrudnionym czy bycia członkiem związku zawodowego) lub o **selekcji próby**, gdy ci, którzy w działaniu uczestniczą, są celowo reprezentowani częściej niż inni. Mechanizmy selekcji mogą być wbudowane w sam proces zbierania danych bądź mogą być zależne od samych jednostek losowania (powodujących na przykład braki odpowiedzi w sondażach).

W mikroekonometrii często zajmujemy się próbami określanymi mianem *matched samples*, to jest **próbami dopasowanymi**[3]. Tego rodzaju próby są opisane w rozdziale 9. W tym przypadku do próby dostają się jedynie pary obserwacji, dopasowane w określony sposób przez badacza. Parę mogą stanowić na przykład firmy o jednakowych charakterystykach (branża, przychody, zatrudnienie itd.) różniące się tym, że wobec jednej prowadzi się postępowanie upadłościowe, a wobec drugiej nie. Parę może stanowić ta sama jednostka „przed" i „po" odpowiednim na nią oddziaływaniu (*within-subject matching*), np. przed i po wprowadzeniu nowych stawek audytorskich. Zazwyczaj parą są dwie różne jednostki (*between-subject matching*). Rozróżnia się jeszcze **próby semi-dopasowane** (*semi-matched samples*) oraz **próby w pełni dopasowane** (*fully matched samples*). Te pierwsze występują wtedy, gdy dobór charakterystyk dla obu elementów każdej pary w próbie jest niejednoznaczny. Próby dobierane (*choice-based*) oraz próby dopasowane (*matched*) często występują jednocześnie. Możemy zatem mieć do czynienia z próbami dobieranymi i niedopasowanymi (*choice-based non-matched*), dobieranymi i dopasowanymi (*choice-based*

[2] Termin „zbilansowanie próby" używany jest także w innych znaczeniach. W planowaniu doświadczeń mianem *balanced sampling* określa się pewien dobór próby w sytuacji niejednakowych prawdopodobieństw wyboru. Jest to dobór, który zapewnia, że wartości globalne zmiennej Y (objaśnianej) oraz zmiennych X (objaśniających) mogą być na podstawie tej próby oszacowane w sposób nieobciążony (por. Deville i Tillé 2004). Ponadto, termin „próba zbilansowana" spotyka się także w ekonometrii panelowej, gdzie chodzi o to, że wszystkie jednostki przekrojowe mają być obserwowane dokładnie w tych samych okresach.

[3] W polskiej literaturze mikroekonometrycznej spotyka się także określenie „próby łączone".