PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Andrzej Bąk
Uniwersytet Ekonomiczny we Wrocławiu

ANALIZA DANYCH O PREFERENCJACH Z WYKORZYSTANIEM MIKROEKONOMETRYCZNYCH MODELI KATEGORII NIEUPORZĄDKOWANYCH I PROGRAMU R

Streszczenie: Celem artykułu jest wskazanie możliwości wykorzystania wybranych mikroekonometrycznych modeli zmiennych o wartościach nieuporządkowanych w badaniach preferencji konsumentów oraz prezentacja przykładów badań empirycznych. Omówiono wielomianowy, warunkowy oraz mieszany model logitowy i przedstawiono przykłady ich wykorzystania w analizie danych o preferencjach.

1. Wstęp

Wśród mikroekonometrycznych modeli kategorii nieuporządkowanych (zmiennych objaśnianych niemetrycznych o obserwacjach nieuporządkowanych) wyróżnia się: wielomianowy model logitowy, warunkowy model logitowy (McFaddena), mieszany model logitowy, zagnieżdżony model logitowy i uogólniony model wartości ekstremalnych oraz modele klas ukrytych. W analizie danych dyskretnych (w tym danych o preferencjach konsumentów), jeżeli zmienna objaśniana opisuje wybory, bardzo często są stosowane wielomianowe, warunkowe i mieszane modele logitowe oraz modele klas ukrytych. Zmienne objaśniające lub zmienne obserwowane w tych modelach opisują konsumentów i obiekty będące przedmiotem wyboru. Celem estymacji tych modeli jest identyfikacja czynników wpływających na określone zachowania rynkowe.

Program R nie oferuje pakietu bezpośrednio wspierającego analizę danych o preferencjach z wykorzystaniem wielomianowego, warunkowego i mieszanego modelu logitowego. Procedury obliczeniowe zawarte w różnych pakietach mogą być jednak wykorzystane w tego typu badaniach.

W artykule przedstawiono następujące zagadnienia:

– cechy wyróżniające mikroekonometrię,
– charakterystykę wybranych mikroekonometrycznych modeli kategorii nieuporządkowanych (wielomianowego, warunkowego oraz mieszanego modelu logitowego),

– sposób estymacji wielomianowego, warunkowego oraz mieszanego modelu logitowego w programie R,
– przykłady zastosowań wielomianowego, warunkowego oraz mieszanego modelu logitowego w badaniach preferencji.

Celem pracy jest wskazanie możliwości wykorzystania wybranych mikroekonometrycznych modeli zmiennych o wartościach nieuporządkowanych w badaniach preferencji konsumentów oraz prezentacja przykładów badań empirycznych.

2. Mikroekonometria

Teorie użyteczności mieszczą się w obrębie mikroekonomii, natomiast metody badania preferencji można zaklasyfikować jako narzędzia badawcze mikroekonometrii. W metodach tych wykorzystuje się dane o jednostkowych obiektach badania, szczególnie o konsumentach i produktach, które nazywa się w literaturze przedmiotu mikrodanymi dla podkreślenia ich szczegółowości. Metody badawcze stosowane w mikroekonometrii, w tym metody stosowane do pomiaru preferencji, umożliwiają „wydobycie" ukrytych w mikrodanych informacji, które mogą służyć wspomaganiu rynkowych procesów decyzyjnych i wyjaśnieniu zasad postępowania konsumentów.

Mikroekonometria jest dynamicznie rozwijającą się gałęzią ekonometrii, o czym świadczą m.in. wyróżnienia Nagrodą Nobla jej wybitnych reprezentantów[1], badających mikrodane za pomocą metod wyborów dyskretnych. Do głównych cech wyróżniających mikroekonometrię należą (por. [Winkelmann, Boes 2006; Gruszczyński 2002; Hozer 1993]):
– badanie zachowań ekonomicznych jednostek (konsumentów, gospodarstw domowych, firm),
– analiza danych (mikrodanych) na poziomie indywidualnym (jednostkowym),
– niski poziom agregacji mikrodanych (dane szczegółowe),
– możliwość zaobserwowania zjawisk lub zdarzeń niewidocznych w danych zagregowanych,
– nieliniowy rozkład obserwacji oraz wykorzystywanie nieliniowych modeli i metod estymacji parametrów,
– niejednorodność obserwacji (heterogeniczność badanych jednostek),
– duża liczba obserwacji (masowość mikrodanych),
– przekrojowy charakter mikrodanych.

W modelach wykorzystywanych w mikroekonometrii występują najczęściej następujące typy zmiennych objaśnianych:

a) zmienne dychotomiczne (dwukategorialne, np. binarne);
b) zmienne politomiczne (wielokategorialne, np. wybór z wielu opcji):

[1] Nagrodę Banku Szwecji w dziedzinie ekonomii za rok 2000 otrzymali James J. Heckman i Daniel L. McFadden, zajmujący się mikroekonometrią, analizą mikrodanych i metodami wyborów dyskretnych.

– zmienne o kategoriach uporządkowanych,
– zmienne o kategoriach nieuporządkowanych;

c) zmienne ograniczone:
– zmienne cenzurowane,
– zmienne ucięte;

d) zmienne licznikowe (wartości reprezentowane przez nieujemne liczby całkowite).

Realizacje zmiennych (mikrodane) są najczęściej wynikami pomiarów na skalach niemetrycznych (zgromadzone obserwacje są zwykle liczbowymi wartościami dyskretnymi lub symbolami).

Do najczęściej stosowanych w badaniach empirycznych modeli mikroekonometrycznych należą:

a) modele dwumianowe:
– modele liniowe prawdopodobieństwa,
– modele logitowe i probitowe ,
– modele komplementarne log-log,
– modele log-liniowe (tablice kontyngencji);

b) modele wielomianowe:
– kategorii nieuporządkowanych,
– kategorii uporządkowanych;

c) modele klas ukrytych;

d) modele przeżycia (trwania);

e) modele zmiennych ograniczonych.

3. Wybrane modele kategorii nieuporządkowanych

W badaniach preferencji szczególnie ważną rolę odgrywają wielomianowe i warunkowe modele logitowe oraz ich połączenie w postaci tzw. mieszanych modeli logitowych. Modele te mieszczą się w grupie wielomianowych modeli kategorii nieuporządkowanych.

Wielomianowy model logitowy jest uogólnieniem modelu logitowego dla danych binarnych (regresji logistycznej) i może być stosowany wówczas, kiedy zmienna objaśniana przyjmuje w sposób dyskretny wartości ze zbioru liczącego więcej niż dwie kategorie. Model ten wywodzi się z teorii użyteczności losowej oraz tzw. aksjomatu wyboru Luce'a (modelu stałej użyteczności) (zob. [Coombs, Dawes i Tversky 1977, s. 217 i nast.; Bierlaire 1997]).

Wielomianowy model logitowy można przedstawić w postaci (zob. np. [Long 1997; So, Kuhfeld 2005]):

$$P_{ki} = \frac{\exp(\mathbf{x}_k^T \boldsymbol{\beta}_i)}{\sum_{l=1}^{n} \exp(\mathbf{x}_k^T \boldsymbol{\beta}_l)}, \text{ przy czym } \boldsymbol{\beta}_n = \mathbf{0}, \tag{1}$$

gdzie: P_{ki} – prawdopodobieństwo wyboru *i*-tej kategorii przy *k*-tym stanie zmiennych objaśniających; $\mathbf{x}_k^T$ – wektor reprezentujący *k*-ty wiersz macierzy $\mathbf{X}$ (zmiennych objaśniających); $\boldsymbol{\beta}_i$ – wektor parametrów związany z *i*-tą kategorią zmiennej objaśnianej.

Macierz $\mathbf{X}$ w modelu (1) zawiera charakterystyki respondentów, których preferencje dotyczące produktów lub usług są przedmiotem badań. Charakterystyki respondentów są stałe względem tych produktów lub usług.

Warunkowy model logitowy został zaproponowany przez McFaddena (zob. [McFadden 1974]) jako uogólnienie wielomianowego modelu logitowego. Podstawowym kryterium rozróżniania[2] tych modeli jest charakter zmiennych objaśniających, tzn. macierzy $\mathbf{X}$ w równaniu (1). Jeżeli zmienne objaśniające charakteryzują konsumentów, to na ogół wykorzystuje się wielomianowy model logitowy. Jeśli natomiast zmienne objaśniające opisują obiekty będące przedmiotem wyboru, to z reguły stosuje się warunkowy model logitowy (por. [*Categorial Analysis*… 1999])[3].

W warunkowym modelu logitowym prawdopodobieństwo wyboru *i*-tego profilu ze zbioru liczącego *n* elementów jest szacowane na podstawie zależności (por. [Louviere, Woodworth 1983, s. 352-355; Haaijer i Wedel 2000, s. 335; Kuhfeld 1996, s. 7; Long 1997, s. 151-183]):

$$P_{ki} = \frac{\exp(\mathbf{z}_{ki}^T \boldsymbol{\alpha})}{\sum_{l=1}^{n} \exp(\mathbf{z}_{kl}^T \boldsymbol{\alpha})}, \tag{2}$$

gdzie: $\mathbf{z}_{ki}^T$ – *k*-ty wektor macierzy $\mathbf{Z}$ (zmiennych objaśniających) opisujący *i*-tą opcję; $\boldsymbol{\alpha}$ – wektor parametrów (wartość α_j jest związana z *j*-tą zmienną objaśniającą).

Macierz $\mathbf{Z}$ w modelu (2) zawiera charakterystyki produktów lub usług, względem których badane są preferencje respondentów. Wartości zmiennych objaśniających opisujących produkty lub usługi są specyficzne w przekroju opcji wyboru oferowanych respondentom (np. w badaniu ankietowym).

Mieszany model logitowy (3) jest połączeniem modeli (1) i (2), a więc uwzględnia charakterystyki zarówno respondentów, jak i opcji wyboru (produktów lub usług) (zob. np. [Long 1997; So, Kuhfeld 2005]):

---

[2] W literaturze przedmiotu spotkać można przykłady zarówno respektowania (zob. np. [Long 1997; *Categorial Analysis*… 1999]), jak i ignorowania tego rozróżnienia (zob. np. [Lehmann, Gupta, Steckel 1998]). W sensie formalnym modele te są ekwiwalentne, ponieważ wielomianowy model logitowy można przekształcić do postaci warunkowego modelu logitowego i odwrotnie (zob. [Long 1997, s. 180-182]).

[3] Podstawą tego stwierdzenia nie są oczywiście reguły formalne, lecz przykłady praktycznych zastosowań obu modeli.

$$P_{ki} = \frac{\exp(\mathbf{x}_k^T \boldsymbol{\beta}_i + \mathbf{z}_{ki}^T \boldsymbol{\alpha})}{\sum_{l=1}^{n} \exp(\mathbf{x}_k^T \boldsymbol{\beta}_l + \mathbf{z}_{kl}^T \boldsymbol{\alpha})}. \quad (3)$$

Omawiane logitowe modele kategorii nieuporządkowanych (wielomianowy, warunkowy i mieszany) są stosowane w badaniach preferencji w tych sytuacjach, kiedy pomiar preferencji jest przeprowadzany na skali nominalnej (wybór jednej opcji z koszyka ofert). Zastosowanie takiego pomiaru jest bardzo wygodne z punktu widzenia respondenta, ale wymaga zgromadzenia dużej liczby obserwacji (zob. [Kuhfeld 1996]).

Szacowanie parametrów wielomianowych, warunkowych i mieszanych modeli logitowych w programie R można przeprowadzić z wykorzystaniem funkcji `optim` z pakietu `stats`. Funkcję `optim` można wykorzystać do maksymalizacji funkcji największej wiarygodności (zob. [Jackman 2007]). Wybrane argumenty funkcji `optim` są następujące:

```
optim(par, fn, gr=NULL, method=c("Nelder-Mead",
"BFGS", "CG", "L-BFGS-B", "SANN"))
```

`par` wartości początkowe parametrów, które są optymalizowane,
`fn` funkcja,
`gr` gradient (wektor pochodnych cząstkowych),
`method` metoda optymalizacji.

4. Zastosowanie modeli kategorii nieuporządkowanych

W styczniu i lutym 2008 r. w Karpaczu, Szklarskiej Porębie i Dziwiszowie przeprowadzono badanie ankietowe preferencji osób korzystających z wyciągów i stoków narciarskich w Kotlinie Jeleniogórskiej[4]. Spośród 250 rozprowadzonych kwestionariuszy ankiet poprawnie zostało wypełnionych 211, tj. 84%.

W celu identyfikacji preferencji przedstawiono respondentom do oceny profile reprezentujące różne konfiguracje miejscowości, stoków, bazy technicznej i zaplecza hotelowo-noclegowego. Charakterystykę profilów zawiera tab. 1.

Tabela 1. Charakterystyka profilów reprezentujących infrastrukturę narciarską

Atrybut	Nazwy poziomów	Liczba poziomów
Miejscowość	Dziwiszów, Karpacz, Szklarska Poręba	3
Długość stoku	długi, krótki	2
Baza techniczna	instruktor, wypożyczalnia, oświetlenie	3
Zaplecze hotelowe	hotele, tanie noclegi, bary i restauracje	3

Źródło: opracowanie własne.

[4] Badanie przeprowadził Damian Jaskółowski na potrzeby pracy magisterskiej.

Pełny układ czynnikowy liczy 54 profile. W wyniku zastosowania cząstkowego ortogonalnego układu czynnikowego uzyskano zredukowany zbiór 9 profilów, które zostały przedstawione respondentom (narciarzom korzystającym z wyciągów i stoków) do oceny (tab. 2).

Tabela 2. Zredukowany zbiór profilów

Numer profilu	Miejscowość	Długość stoku*	Baza techniczna	Zaplecze hotelowe
1	Dziwiszów	krótki	instruktor	hotele
2	Dziwiszów	długi	wypożyczalnia	tanie noclegi
3	Szklarska Poręba	długi	instruktor	tanie noclegi
4	Szklarska Poręba	długi	oświetlenie	hotele
5	Szklarska Poręba	krótki	wypożyczalnia	bary i restauracje
6	Karpacz	długi	instruktor	bary i restauracje
7	Karpacz	długi	wypożyczalnia	hotele
8	Dziwiszów	długi	oświetlenie	bary i restauracje
9	Karpacz	krótki	oświetlenie	tanie noclegi

*Długość trasy zjazdowej (stoku) do 1000 m – krótki, 1000-4500 m – długi.

Źródło: opracowanie własne.

Zgromadzono 1899 obserwacji (211 · 9), które wykorzystano do oszacowania warunkowego modelu logitowego (2), przy założeniu, że ostatnie poziomy atrybutów są poziomami odniesienia. Wyniki estymacji zestawiono w tab. 3 i 4. Parametry modelu wskazują, że najważniejszą charakterystyką profilów dla badanych respondentów jest długość stoku, a w dalszej kolejności miejscowość, w której stok jest położony. Te dwie cechy są jednak w oczywisty sposób ze sobą powiązane. Znacznie mniej istotne są pozostałe charakterystyki, tj. baza techniczna i zaplecze hotelowe.

Tabela 3. Parametry modelu

Zmienna	Parametr	exp(parametr)	Błąd standardowy
miejscowosc1	0,459	1,582	0,268
miejscowosc2	1,061	2,888	0,223
stok1	1,343	3,829	0,252
baza1	–0,385	0,680	0,217
baza2	–0,632	0,532	0,253
zaplecze1	–0,548	0,578	0,215
zaplecze2	–0,598	0,550	0,223

Źródło: opracowanie własne.

Tabela 4. Częstości wybierania i wyrazy wolne profilów

Numer profilu	Częstość wybierania	Wyraz wolny	Prawdopodobieństwo
1	5	–1.23000	0,024
2	14	–0.19900	0,059
3	51	1.09000	0,249
4	31	0.59600	0,139
5	17	–0.00456	0,081
6	54	1.15000	0,248
7	19	0.10700	0,098
8	16	–0.06520	0,083
9	4	–1.45000	0,019

Źródło: opracowanie własne.

Informacje z tab. 4 wskazują, iż profil 6 charakteryzuje się najwyższą, a profil 9 najniższą użytecznością wśród badanych respondentów. W celu identyfikacji czynników wpływających na te użyteczności oszacowano wielomianowe modele logitowe, w których wzięto pod uwagę wybrane charakterystyki respondentów, oraz mieszane modele logitowe, w których uwzględniono charakterystyki zarówno respondentów, jak i profilów. W tabeli 5 zestawiono wyniki estymacji wielomianowego modelu logitowego ze zmienną „wiek" i mieszanego modelu logitowego ze zmiennymi „wiek" i „stok" (uwzględniono profile nr 6 i 9).

Wielomianowy model logitowy
profile: p6 i p9 cecha respondenta: wiek

	Estimate	Std. Error	z-stat
p6	0.6030	0.4600	1.310
p9	-0.7350	1.5400	-0.476
p6wiek	0.0100	0.0143	0.703
p9wiek	-0.0349	0.0559	-0.624

Mieszany model logitowy
profile: p6 i p9 cecha respondenta: wiek
cecha profilu: stok

	Estimate	Std. Error	z-stat
p6	0.4220	0.4610	0.9150
p9	-0.0661	1.5600	-0.0424
p6wiek	0.0100	0.0143	0.7040
p9wiek	-0.0344	0.0559	-0.6150
stok	-0.8680	0.2300	-3.7700

Rys. 1. Wyniki estymacji wielomianowego i mieszanego modelu logitowego

Źródło: opracowanie własne z wykorzystaniem programu R.

W badanej próbie średnia arytmetyczna wieku respondentów wynosi 29 lat. Wyniki estymacji zamieszczone na rys. 1 informują, że narciarze „starsi" preferują profil nr 6 (długi stok) i negatywnie oceniają profil nr 9 (krótki stok). Głównym czynnikiem na to wpływającym jest długość stoku (w tych modelach ujemna wartość parametru dla zmiennej „stok" oznacza, że negatywnie oceniane są stoki o krótszej długości). Potwierdzają to wyniki estymacji mieszanego modelu logitowego, w którym uwzględniono tę cechę profilów (długość stoku) i otrzymano ujemną wartość parametru, wskazującą na negatywną ocenę krótkich stoków.

Literatura

Agresti A., *Categorical Data Analysis*, Second Edition, Wiley, New York 2002.
Bierlaire M., *Discrete Choice Models*, http://web.mit.edu/mbi/www/ michel.html, Cambridge, Massachusetts Institute of Technology, 1997.
Categorial Analysis – Part 1, http://pytheas.ucs.indiana.edu/~statmath/ stat/all/cat/printable.htm. Indiana University, 1999.
Coombs C.H., Dawes R.M., Tversky A., *Wprowadzenie do psychologii matematycznej*, PWN, Warszawa 1977.
Gruszczyński M., *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa 2002.
Haaijer R., Wedel M., *Conjoint Choice Experiments: General Characteristics nd Alternative Model Specifications*, [w:] A. Gustafsson, A. Herrmann, F. Huber (red.), *Conjoint Measurement: Methods and Applications*, Springer, Berlin 2000.
Hozer J., *Mikroekonometria. Analizy, diagnozy, prognozy*, PWE, Warszawa 1993.
Jackman S., *Models for Unordered Outcomes*, Political Science 150C/350C, http://jack-man.stanford.edu/classes/, 2007.
Kuhfeld W.F., *Multinomial Logit, Discrete Choice Modeling. An Introduction to Designing Choice Experiments, Collecting, Processing, and Analyzing Choice Data with the SAS® System*, http://ftp.sas. com/techsup/download/technote/ts273.pdf. Cary, SAS Institute Inc, 1996.
Lehmann D.R., Gupta S., Steckel J.H., *Marketing Research*, Reading, Massachusetts, Addison-Wesley, 1998.
Long J.S., *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks--London-New Delhi, SAGE Publications, 1997.
Louviere J.J., Woodworth G., *Design and Analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data*, „Journal of Marketing Research" 1983, November 20.
McFadden D., *Conditional Logit Analysis of Qualitative Choice Behavior*, [w:] *Frontiers in Econometrics*, P. Zarembka (red.), Academic Press, New York-San Francisco-London 1974.
So Y., Kuhfeld W.F., *Multinomial Logit Models*, http://support.sas.com/ resources/papers/tnote/ tnote_marketresearch.html, SAS Institute, Cary 2005.
Winkelmann R., Boes S., *Analysis of Microdata*, Springer-Verlag, Berlin, Heidelberg 2006.

PREFERENCES DATA ANALYSIS USING MICROECONOMETRIC UNORDERED RESPONSE MODELS AND R PROGRAM

Summary: The main aim of the paper is presentation, with empirical examples, possibilities to use models with unordered polytomous dependent variable in the consumer preferences analysis. There are presented multinomial, conditional and mixed logit model and empirical examples of their using in consumer preferences analysis.