EKONOMETRIA ECONOMETRICS 3(37) • 2012

ISSN 1507-3866

Marcin Pełka, Aneta Rybicka
Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE REGRESJI KLAS UKRYTYCH W ANALIZIE DANYCH MIKROEKONOMETRYCZNYCH

Streszczenie: W artykule przedstawiono zastosowanie regresji klas ukrytych w metodach wyborów dyskretnych, gdzie głównym celem jest wskazanie, które ze zmiennych wpływają na nieznaną strukturę klas. Dzięki wykorzystaniu regresji klas ukrytych z zastosowaniem pakietu `flexmix` programu R odkryto nieznaną wcześniej strukturę dwóch klas konsumentów piwa jasnego. Klasa 1 to osoby częściej spożywające piwo, w niewielkich ilościach, a co za tym idzie – wydające na piwo niewiele. Natomiast klasa 2 to osoby kupujące piwo rzadziej, ale w większych ilościach i za większe kwoty pieniężne.

Słowa kluczowe: regresja klas ukrytych, analiza preferencji, metody wyborów dyskretnych.

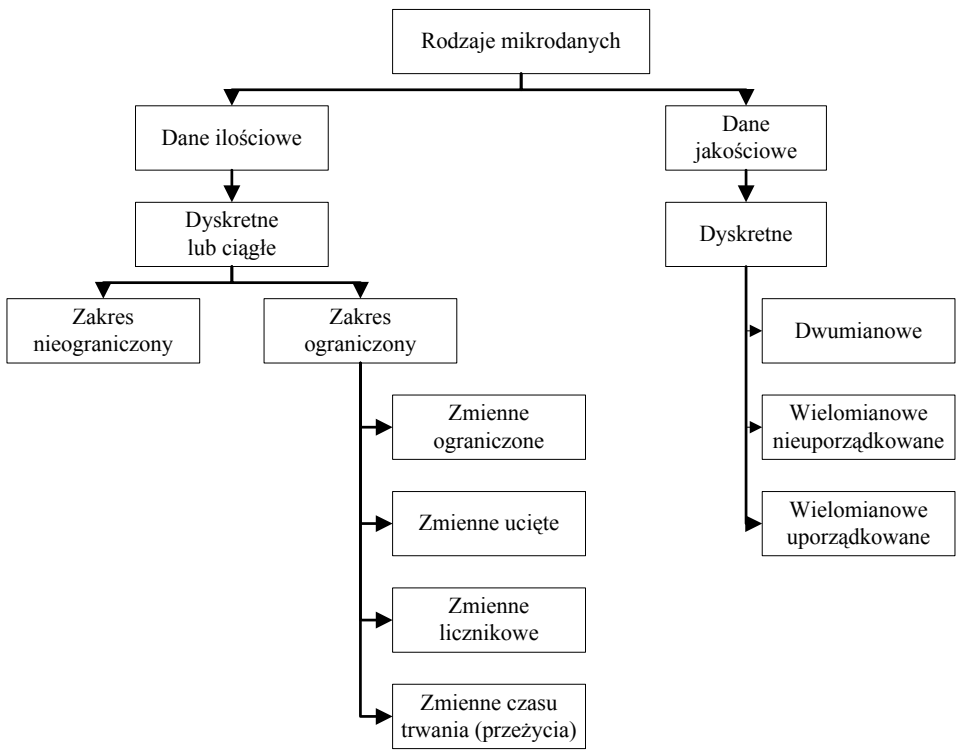
1. Wstęp

Termin „mikroekonometria" pojawia się w literaturze ekonometrycznej od połowy lat 80. XX wieku i jest odpowiedzią na potrzebę wyodrębnienia tej części ekonometrii, która obejmuje metody wykorzystywania mikrodanych w analizie zagadnień ekonomicznych [Gruszczyński 2010, s. 16].

Mikroekonometria wyróżnia się kilkoma cechami [Gatnar, Walesiak 2011, s. 112-113]:

- zajmuje się badaniem zachowań ekonomicznych jednostek,
- analizuje mikrodane na poziomie indywidualnym,
- mikrodane wykorzystywane w analizie to dane szczegółowe,
- istnieje możliwość zaobserwowania zjawisk lub zdarzeń, które są niewidoczne w danych zagregowanych (np. metody stosowane do pomiaru preferencji umożliwiają „wydobycie" ukrytych w mikrodanych informacji),
- nieliniowy rozkład obserwacji,
- wykorzystywanie nieliniowych modeli i metod estymacji parametrów,
- niejednorodność obserwacji (niejednorodność badanych jednostek),
- masowość mikrodanych,
- przekrojowy charakter mikrodanych.

Mikrodane to zbiór danych liczbowych o pojedynczych jednostkach [Gruszczyński 2010, s. 13]. Mikrodane są podstawą modelowania w mikroekonometrii i można klasyfikować je na różne sposoby. Jedną z możliwych klasyfikacji mikrodanych przedstawia rys. 1.



Rys. 1. Rodzaje mikrodanych

Źródło: [Gruszczyński 2010, s. 15].

Podstawą klasyfikacji mikrodanych jest model, który objaśnia zmienną o określonej postaci (tab. 1) [Gruszczyński 2010, s. 17-18].

W najszerszym ujęciu model mikroekonometryczny to model typu regresyjnego oparty na mikrodanych.

Należałoby wspomnieć, że za ważniejszych przedstawicieli mikroekonometrii uważa się laureatów Nagrody Nobla z dziedziny ekonomii w roku 2000: Jamesa Heckmana i Daniela McFaddena. James Heckman został wyróżniony nagrodą za „rozwój teorii i metod analizy prób selektywnych", natomiast McFadden za „rozwój teorii i metod analizy wyboru dyskretnego".

Ważnym elementem mieszczącym się w obrębie mikroekonometrii są metody badania preferencji. Preferencje są wykorzystywane w celu kwantyfikacji użyteczno-

ści, której bezpośrednio nie można zmierzyć. Teorie użyteczności mieszczą się natomiast w obrębie mikroekonomii.

Tabela 1. Systematyka modeli zmiennych jakościowych i ograniczonych

1. Modele dwumianowe	2. Modele wielomianowe
Liniowy model prawdopodobieństwa (LMP) Probitowy Logitowy Logarytmiczno-logitowy Gompitowy (krzywej Gompertza) Komplementarny log-log Burritowy (rozkładu Burra) Ucięty LMP Krzywej Urbana	Modele kategorii uporządkowanych Uporządkowany model logitowy i probitowy Uogólniony model uporządkowany Modele danych sekwencyjnych Modele kategorii nieuporządkowanych Wielomianowy model logitowy i probitowy Warunkowy model logitowy (McFadden) Zagnieżdżony model logitowy Mieszany model logitowy
3. Modele zmiennych ograniczonych i uciętych	4. Modele licznikowe i czasu trwania
Regresja ucięta Modele tobitowe (klasyczne) Dwugraniczny model tobitowy Model doboru próby (Heckman) Modele efektów oddziaływania	Regresja Poissonowska Model rozkładu ujemnego dwumianowego Model czasu trwania Model licznikowy ucięty Model płotkowy

Źródło: [Gruszczyński 2010, s. 18].

W artykule zaprezentowano zastosowanie regresji klas ukrytych w badaniu preferencji konsumentów piwa. Do obliczeń wykorzystano pakiet `flexmix` programu R.

2. Regresja klas ukrytych

Modele klas ukrytych zostały zaproponowane przez Lazerfelda i Henryego [1968]. W ostatnich latach zaproponowano modyfikacje modeli klas ukrytych, które pozwalają analizować zmienne obserwowalne różnych typów [Magidson, Vermunt 2004, s. 175; Hagenaars, McCutheon 2002, s. xi, 4-6; McCutheon 1987, s. 7-8]. Modele klas ukrytych nazywane są także modelami ze zmiennymi ukrytymi i są one przykładem podejścia modelowego w analizie skupień. Idea ta wykorzystuje znane w statystyce podejście oparte na mieszankach rozkładów (por. np. [Gatnar, Walesiak 2011, s. 204]).

Ze względu na różne rozkłady zmiennych obserwowalnych i zmiennych ukrytych możemy mówić o różnych modelach zmiennych ukrytych [Vermunt, Magidson 2003, s. 1].

Tak jak podają Bartholomew i Knott [2002, s. 3], wyróżniamy cztery główne rodzaje modeli (zob. tab. 2):

– analiza czynnikowa (*Factor Analysis* – FA),
– analiza z ukrytymi charakterystykami (*Latent Trait Analysis* – LTA),
– analiza z ukrytymi profilami (*Latent Profile Analysis* – LPA),
– analiza z ukrytymi zmiennymi (*Latent Class Analysis* – LCA).

Tabela 2. Klasyfikacja modeli zmiennych ukrytych

Zmienna obserwowalna	Zmienna ukryta	
	ciągła	skokowa
Ciągła	*Factor Analysis*	*Latent Profile Analysis*
Skokowa	*Latent Trait Analysis*	*Latent Class Analysis*

Źródło: [Vermunt, Magidson 2003, s. 1].

Istnieją trzy główne obszary analizy z wykorzystaniem modeli klas ukrytych, które obejmują: umieszczanie analizowanych przypadków w segmentach, redukcję zmiennych, konstrukcję skali oraz predykcje zmiennej zależnej [Magidson, Vermunt 2002, s. 2].

Można więc wyróżnić trzy główne rodzaje modeli klas ukrytych [Magidson, Vermunt 2002, s. 2]:
- modele klas ukrytych z wykorzystaniem segmentów (*Latent Class Cluster Models*),
- modele klas ukrytych z wykorzystaniem czynników (*Latent Class Factor Models*),
- modele klas ukrytych w regresji i w modelach wyboru (*Latent Class Regressioon and Choice Models*).

Model klas ukrytych w regresji, znany również jako model segmentacji klas ukrytych, charakteryzuje się tym, że [Magidson, Vermunt 2002, s. 5]:
- jest wykorzystywany do predykcji zależnej zmiennej będącej funkcją predyktorów,
- zawiera zmienną ukrytą o R kategoriach, z których każda reprezentuje homogeniczną populację (klasę, segment),
- dla każdego z ukrytych segmentów można wyestymować inny model regresji,
- klasyfikuje cechy w segmenty i symultanicznie szacuje dla każdego z nich modele regresji.

Zalety tego podejścia to m.in. [Magidson, Vermunt 2002, s. 5-6]:
- osłabienie tradycyjnych założeń, mówiących o tym, że każdy model dla wszystkich cech zakłada $R = 1$, co pozwala na oszacowanie osobnego modelu regresji dla każdego z segmentów,
- diagnostyczne statystyki pozwalają na określenie wartości dla R,
- w przypadku, gdy $R > 1$, model może zostać rozszerzony o dodatkowe zmienne objaśniające, by przeprowadzona analiza była dokładniejsza i by przyporządkowanie do segmentu było bardziej klarowne.

Zakładając, że mamy do czynienia ze skończoną liczbą mieszanek modeli o R elementach, możemy zapisać (zob. np. [Leish 2004, s. 2; Gatnar, Walesiak 2011, s. 224-226; McLachlan, Peel 2000, s. 7-8, 22-28]):

$$h(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\varphi}) = \sum_{r=1}^{R} \pi_r f(\mathbf{y} \mid \mathbf{x}, \theta_r), \tag{1}$$

gdzie: $\pi_r \geq 0,\ \sum_{r=1}^{R} \pi_r = 1,$

$\mathbf{y}$ – zmienna zależna o warunkowym rozkładzie h,

$\mathbf{x}$ – wektor zmiennych niezależnych (obserwowalnych),

π_r – prawdopodobieństwa bezwarunkowe, wyrażające przynależności do poszczególnych klas ukrytych,

θ_r – wektor nieznanych parametrów w *r*-tej klasie,

$\boldsymbol{\varphi} = (\pi_1, \ldots, \pi_r, \theta_1', \ldots, \theta_r')$ – wektor wszystkich parametrów.

Prawdopodobieństwo *a posteriori*, że obserwacja (x, y) należy do *j*-tej klasy, zdefiniowane jest następująco [Leish 2004, s. 3]:

$$P(j \mid x, y, \phi) = \frac{\pi_j f(y \mid x, \theta_j)}{\sum_r \pi_r f(y \mid x, \theta_r)}. \tag{2}$$

Prawdopodobieństwa *a posteriori* są wykorzystane do segmentacji obiektów. Obiekt przydzielany jest do klasy (segmentu) o największym prawdopodobieństwie.

Typowe zastosowania w marketingu odpowiednika klas ukrytych w regresji i modelach wyboru to: [Magidson, Vermunt 2002, s. 6]:

- studia, analiza satysfakcji klienta: identyfikacje poszczególnych determinant satysfakcji klienta, które są odpowiednie dla każdego segmentu,
- wspólne studia: identyfikacja atrybutów produktów, które należą do różnych segmentów rynku,
- bardziej ogólnie: identyfikacja ukrytych segmentów, które mogą wyjaśnić nieobserwowalną heterogeniczność wśród danych.

Innym zastosowaniem regresji klas ukrytych jest wskazanie czynników (zmiennych), które różnicują nieznaną (ukrytą) strukturę klas. Przykład takiego zastosowania jest prezentowany w niniejszym artykule.

3. Przykład empiryczny

Piwo jest określane jako niskoprocentowy napój alkoholowy o złotej lub ciemnej barwie, chmielowym zapachu, gorzkim smaku i obfitej pianie, wytwarzany z jęczmienia, chmielu, pszenicy, słodu, wody i drożdży. Gatunki tego trunku można podzielić na kilka podstawowych grup. Najpopularniejszy podział odnosi się do zabarwienia piwa. Według niego, rozróżniamy piwa jasne i piwa ciemne. Innym kryterium

jest zawartość ekstraktu w piwie. Jeśli napój zawiera go do 10%, należy do lekkich, 10-15% – to piwo pełne, a piwa zawierające ponad 15% ekstraktu to piwa mocne. Można jeszcze dokonać podziału piwa na alkoholowe i bezalkoholowe. Jest to jednak podział umowny, ponieważ nie istnieją piwa całkowicie pozbawione alkoholu. Nie przekracza on jednak 1-1,2% całkowitej objętości, co nie ma istotnego wpływu na koncentrację alkoholu we krwi.

Badanie piwa przeprowadzone na przełomie lipca i sierpnia 2002 r. na terenie miast Jelenia Góra i Lwówek Śląski jest przykładem[1] wykorzystania metod wyborów dyskretnych w badaniach preferencji konsumentów piwa jasnego. W niniejszym artykule wykorzystana zostanie jedynie część informacji o preferencjach konsumentów – przy czym pod uwagę wzięto jedynie bloki zawierające najlepiej i najgorzej ocenione profile piwa.

Do zbadania preferencji konsumentów piwa, na podstawie literatury przedmiotu oraz informacji zasięgniętych z działu marketingu Prywatnego Browaru w Lwówku Śląskim, wyznaczono zbiór atrybutów uwzględnianych przez konsumenta piwa: kraj pochodzenia (produkcji), cenę, zawartość alkoholu, opakowanie oraz pojemności. Każdy wariant został opisany przez 5 atrybutów decydujących o wyborze (konsumpcji) danego piwa, a każdy atrybut opisano za pomocą odpowiadających mu poziomów (tab. 3).

Tabela 3. Atrybuty charakteryzujące piwo jasne i ich poziomy

Atrybuty	Poziomy
Kraj	Polska, Niemcy, Czechy, Dania, Holandia
Cena	do 2 zł, 2-4 zł, powyżej 4 zł
Zawartość alkoholu	do 1%, 1,8-5%, powyżej 5%
Opakowanie	butelka, puszka, kufel
Pojemność	0,33 l, 0,5 l, powyżej 0,5 l

Źródło: [Bąk, Rybicka 2005, s. 309].

Pełny eksperyment czynnikowy zawierałby $3^{4\times5}$ zbiorów (5 profilów opisanych 4 atrybutami zawierającymi po 3 poziomy). Zatem liczba zbiorów w pełnym eksperymencie wyniosłaby 3 486 784 401. Wstępny zbiór danych (tzw. obserwacje kandydujące) zawierałby 19 683 zbiorów. Minimalny rozmiar eksperymentu zawierałby $20 \times (3–1) + 1 = 41$ zbiorów, zaś rozmiar eksperymentu wykorzystanego w badaniu wynosił 45 zbiorów. Efektywność układu czynnikowego wyniosła D = 85,96. Redukcję kompletnego układu czynnikowego przeprowadzono za pomocą iteracyjnego algorytmu Fedorova, który pozwala na znalezienie optymalnego nieortogonalnego układu czynnikowego wytypowanego wstępnie zestawu danych [Bąk, Rybicka 2005, s. 309].

[1] Szerzej badanie to opisują: A. Bąk, A. Rybicka [2005, s. 305-312].

W badaniu wykorzystano 3 bloki. Liczba zbiorów w każdym z bloków wyniosła 15. Liczba profilów w pełnym eksperymencie równałaby się 405 (5×3×3×3×3). Każdy ze zbiorów zawierał 6 profilów (5 + opcja „żaden z tych profilów") (rys. 2).

1. Wybierz preferowany profil piwa jasnego lub zrezygnuj z wyboru (zaznacz jedną z 6 opcji):

Kraj	Cena	Alkohol	Opakowanie	Pojemność	*Wybieram opcję*
Polska	2-4 zł	pow. 5,0%	kufel	pow. 0,5 l	*1*
Niemcy	pow. 4 zł	do 1,0%	puszka	0,5 l	*2*
Czechy	pow. 4 zł	do 1,0%	puszka	0,33 l	*3*
Dania	2-4 zł	1,8–5,0%	butelka	0,5 l	*4*
Holandia	do 2 zł	do 1,0%	kufel	0,33 l	*5*
Żaden z tych profilów					*6*

Rys. 2. Zbiór profilów charakteryzujących piwo jasne

Źródło: opracowanie własne.

Tabela 4. Estymacja parametrów modelu

Atrybut		DF	Parametr	Błąd	χ^2	Pr > χ^2	Iloraz hazardu
Kraj	Polska	1	–0,19009	0,08048	5,5779	0,0182	0,827
	Niemcy	1	–0,84030	0,08626	94,8857	<,0001	0,432
	Czechy	1	–0,66955	0,08359	64,1529	<,0001	0,512
	Dania	1	–0,96916	0,09008	115,7588	<,0001	0,379
	Holandia	1	–1,37733	0,09538	208,5386	<,0001	0,252
	brak	0	0	.	.	.	.
Cena	brak	0	0	.	.	.	.
	2-4 zł	1	0,35644	0,05740	38,5609	<,0001	1,428
	do 2 zł	1	0,50510	0,05521	83,7052	<,0001	1,657
	powyżej 4 zł	0	0	.	.	.	.
Alkohol	brak	0	0	.	.	.	.
	1,8-5%	1	–0,02562	0,05012	0,2614	0,6092	0,975
	do 1%	1	–0,85350	0,05995	202,7177	<,0001	0,426
	powyżej 5%	0	0	.	.	.	.
Opakowanie	brak	0	0	.	.	.	.
	butelka	1	0,18360	0,05470	11,2653	0,0008	1,202
	kufel	1	0,25960	0,05625	21,3002	<,0001	1,296
	puszka	0	0	.	.	.	.
Pojemność	brak	0	0	.	.	.	.
	0,33 l	1	–0,31304	0,05465	32,8142	<,0001	0,731
	0,5 l	1	0,00462	0,05182	0,0080	0,9289	1,005
	pow. 0,5 l	0	0	.	.	.	.

Źródło: [Bąk, Rybicka 2005, s. 310].

Liczba profilów wykorzystanych w badaniu równała się zatem 270 (225 profilów charakteryzujących piwo jasne oraz 45 profilów opcji rezygnacji z wyboru). Rozprowadzono 300 ankiet, uzyskano zaś 235 prawidłowo wypełnionych (wykorzystanych w badaniu). Liczba ankiet w poszczególnych blokach przedstawiała się następująco: blok 1 – 75, blok 2 – 78, blok 3 – 82 ankiety. Łącznie zgromadzono 21 150 obserwacji (15 zbiorów × 6 profilów × 235 respondentów).

W wyniku estymacji warunkowego modelu logitowego uzyskano oszacowania parametrów, które zostały przedstawione w tab. 4.

Oszacowano użyteczności cząstkowe poziomów atrybutów. Użyteczności te zostały wykorzystane do obliczenia użyteczności całkowitych każdego z prezentowanych profilów. Następnym krokiem było obliczenie prawdopodobieństwa wyboru każdego profilu z całego zestawu. Tabela 5 przedstawia 5 profilów o największych prawdopodobieństwach wyboru i 5 profilów o najmniejszych prawdopodobieństwach wyboru spośród 270 ocenianych profilów.

W związku z tym, iż wartości oszacowanych współczynników regresji logistycznej są trudne do interpretacji, możliwości interpretacyjne stwarza przekształcenie oszacowanego równania regresji logistycznej w tzw. iloraz hazardu (*hazard ratio*). Iloraz hazardu przedstawia relatywną możliwość wystąpienia zdarzenia w wyniku działania czynnika opisanego przez zmienną niezależną (zakładając kontrolowanie, czyli stabilność, pozostałych zmiennych uwzględnionych w równaniu). Iloraz ten informuje o stymulującym lub destymulującym wpływie oszacowanej wartości na prawdopodobieństwo wyboru danego profilu.

Tabela 5. Profile o największym i najmniejszym prawdopodobieństwie wyboru

Kraj	Cena	Alkohol	Opakowanie	Pojemność	Prawdopodobieństwo
Największe prawdopodobieństwo wyboru					
Polska	do 2 zł	powyżej 5%	kufel	powyżej 0,5 l	0,0104
Polska	do 2 zł	1,8-5%	kufel	powyżej 0,5 l	0,0102
Polska	do 2 zł	powyżej 5%	butelka	0,5 l	0,0097
Polska	do 2 zł	1,8-5%	butelka	powyżej 0,5 l	0,0094
Polska	2-4 zł	powyżej 5%	kufel	powyżej 0,5 l	0,0090
Najmniejsze prawdopodobieństwo wyboru					
Holandia	powyżej 4 zł	do 1%	butelka	powyżej 0,5 l	0,0008
Dania	pow. 4 zł	do 1%	puszka	0,33 l	0,0007
Holandia	2-4 zł	do 1%	puszka	0,33 l	0,0007
Holandia	pow. 4 zł	do 1%	puszka	powyżej 0,5 l	0,0006
Holandia	pow. 4 zł	do 1%	butelka	0,33 l	0,0006

Źródło: [Bąk, Rybicka 2005, s. 311].

Z przeprowadzonego badania wynika, iż:

- o wyborze marki (gatunku) piwa jasnego decydują w kolejności: cena, opakowanie, objętość, zawartość alkoholu oraz kraj;

- wpływ stymulujący na prawdopodobieństwo wyboru marki (gatunku) piwa jasnego mają: cena do 2 zł, cena od 2 do 4 zł, opakowanie kufel, opakowanie butelka, objętość 0,5 l.

Część trzecia kwestionariusza ankietowego (tzw. metryczka) zawierała pytania dotyczące podstawowych cech socjologicznych badanych respondentów, które w żaden sposób nie identyfikują indywidualnie poszczególnych osób. Informacje te pozwoliły na analizowanie badanej grupy w zakresie przekrojów społecznych lub segmentów.

W dalszej części, która ma na celu zastosowanie regresji klas ukrytych, wykorzystano informacje (łącznie 153 obserwacje z bloku 1 i 2, które zawierały najlepsze i najgorsze piwa) otrzymane z metod wyborów dyskretnych na temat najlepszych i najgorszych profilów piwa. Z wykorzystaniem regresji klas ukrytych zbadano, które ze zmiennych mają wpływ na wybór profilów (zmienna objaśniana).

Pod uwagę wzięto następujące zmienne opisujące konsumpcję piwa oraz dane z metryczki ankiety (zmienne objaśniające):

1) x_1 – częstość konsumpcji piwa (zmienna porządkowa),
2) x_2 – okazje konsumpcji piwa (zmienna nominalna wielostanowa),
3) x_3 – miejsce konsumpcji piwa (zmienna nominalna wielostanowa),
4) x_4 – kwota miesięcznych wydatków na konsumpcję piwa (zmienna ilorazowa),
5) x_5 – miejsce zakupu piwa (zmienna nominalna wielostanowa),
6) x_6 – liczba zakupywanych jednorazowo butelek (puszek) piwa (zmienna ilorazowa),
7) x_7 – płeć respondenta (zmienna nominalna dwumianowa),
8) x_8 – wiek (zmienna ilorazowa),
9) x_9 – status edukacyjny (zmienna nominalna wielostanowa),
10) x_{10} – wykształcenie (zmienna porządkowa),
11) x_{11} – kategoria zawodowa (zmienna nominalna wielostanowa),
12) x_{12} – miesięczny dochód na jedną osobę w rodzinie (zmienna ilorazowa),
13) x_{13} – źródło dochodów (zmienna nominalna wielostanowa),
14) x_{14} – stan cywilny (zmienna nominalna wielostanowa),
15) x_{15} – liczba posiadanych dzieci (zmienna ilorazowa),
16) x_{16} – miejsce zamieszkania (zmienna nominalna wielostanowa).

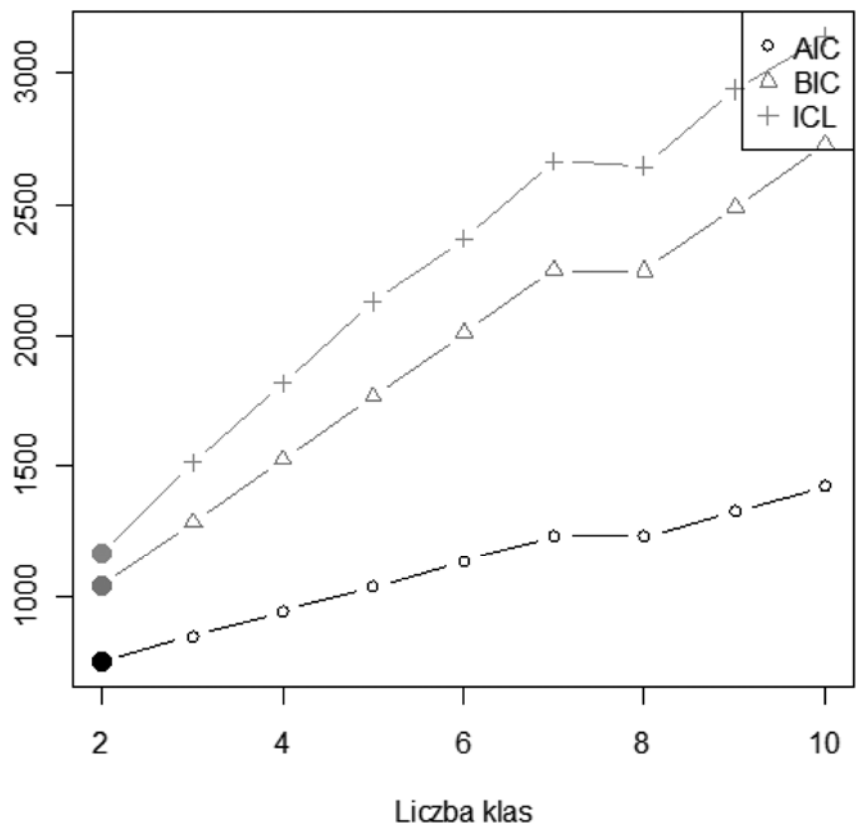
Dla tych danych przygotowano model z wykorzystaniem rozkładu Poissona. Można podać kilka przykładów zastosowania modelu Poissona w badaniach społeczno-ekonomicznych. Rozkład zmiennej zależnej dla funkcji regresji tego typu można zapisać jako (zob. [Gatnar, Walesiak 2011, s. 230]):

$$Poisson\ \left(y_i \mid \lambda_{is}\right) = \frac{\lambda_{is}^{y_i}}{y_i!}\exp\left(-\lambda_{is}\right), \tag{3}$$

gdzie: λ_{is} – to jedyny parametr rozkładu równy wartości oczekiwanej zmiennej Y, $\lambda_{is} = E(Y) = \exp(\boldsymbol{\beta}'\mathbf{X})$.

W rozkładzie tym zakłada się, że wartość przeciętna jest równa wariancji, a zdarzenia są wzajemnie niezależne (zob. [Gatnar i Walesiak 2011, s. 230]).

Oszacowano modele regresji dla różnej liczby klas $(2;10)$ oraz obliczono wartości kryteriów informacyjnych, których wartości przedstawiono na rys. 3.



Rys. 3. Wartości kryteriów informacyjnych dla różnej liczby klas

Źródło: obliczenia własne z wykorzystaniem funkcji `stepFlexmix` z pakietu `flexmix` programu R.

Wartości kryterium zarówno AIC, jak i BIC sugerują, że w zbiorze danych mamy do czynienia z dwiema klasami o następujących liczebnościach (klasa 1 – 107 obiektów oraz klasa 2 – 46 obiektów). Model osiągnął konwergencję po 82 iteracjach.

Następnie obliczono ilorazy hazardu dla wszystkich zmiennych, które zawarto w tab. 6.

Tabela 6. Wpływ zmiennych na wybór profilów

Zmienna (atrybut)	Realizacje zmiennej	**Klasa I**		**Klasa II**	
		parametr	iloraz hazardu	parametr	iloraz hazardu
1	2	3	4	5	6
Częstość konsumpcji	–	0,16	1,00	– 0,27	0,81
Okazje konsumpcji	spotkania	– 0,22	1,21	0,54	0,99
	uroczystości	– 0,48	1,02	0,45	0,92
	zdrowie	0,47	1,91	1,03	1,81
Miejsce zakupu	dom	–0,30	0,69	–0,34	0,45
	plener	–0,15	0,56	–1,23	0,53
	sklep	–1,16	0,20	–1,29	0,79
Kwota miesięcznych wydatków	–	–0,14	0,86	0,07	1,04
Miejsce zakupu piwa	inne	0,70	1,98	0,99	1,83
	mały sklep	–0,22	1,33	0,57	1,87
	restauracja	–0,62	0,34	–1,28	0,41
Ilość zakupywanego piwa	–	0,14	0,89	–0,17	1,90
Płeć	M	–0,37	1,05	0,27	0,62
Wiek	–	0,00	0,99	0,00	1,00
Status edukacyjny	inne	–1,19	0,20	–1,69	0,72
	student	0,09	0,92	–0,03	1,43
	student podyplomowy	0,52	1,47	–	–
	uczeń	–0,17	1,53	0,65	0,66
Wykształcenie	–	–0,05	1,08	–0,15	0,72
Kategoria zawodowa	stanowisko kierownicze	0,22	1,14	–0,27	0,70
	pracownik fizyczny	0,10	0,92	–0,21	1,56
	pracownik umysłowy	0,47	0,91	–0,31	0,55
	pracodawca	–0,75	1,83	0,82	0,33
	urzędnik	0,32	0,86	–0,68	0,88
	wolny zawód	–0,02	1,91	0,90	1,38
Miesięczny dochód	–	–0,01	0,93	0,05	1,45

1	2	3	4	5	6
Źródło dochodów	inne	0,84	1,14	1,59	1,42
	praca dorywcza	0,61	1,43	1,52	1,03
	pracodawca państwowy	0,36	1,11	1,56	1,58
	pracodawca prywatny	0,71	1,84	1,49	1,51
	renta	1,04	1,26	0,68	1,97
	rodzina	0,74	1,07	0,67	1,44
	stypendium	–0,10	1,54	1,09	1,72
	własna działalność gospodarcza	1,21	1,63	0,92	1,82
	zasiłek	0,60	1,14	0,79	1,42
Stan cywilny	małżeństwo	–0,05	0,27	1,39	1,24
	rozwiedziony	–0,09	0,68	1,86	1,44
	wdowieństwo	0,06	0,43	2,38	1,79
	wolny	0,24	0,93	1,31	1,71
	wolny związek	0,40	1,52	0,91	1,92
Liczba dzieci	–	0,18	1,22	–0,20	1,01
Miejsce zamieszkania	miasto do 50 tys.	0,24	1,25	0,22	0,79
	miasto do 10 tys.	0,27	1,28	0,36	1,08
	miasto do 100 tys.	0,04	1,08	–0,56	0,37
	wieś	0,27	1,52	0,23	0,28

W tabeli zawarto zmienne istotne na poziomie 0,05.

Źródło: obliczenia własne.

Klasa 1 to osoby częściej spożywające piwo niż osoby z klasy 2. Głównym czynnikiem, który zwiększa prawdopodobieństwo przynależności osoby do tej klasy, jest konsumpcja piwa przy okazji spotkań, uroczystości i z powodów zdrowotnych. Osoby z tej klasy wydają mniej na konsumpcję piwa, niż ma to miejsce w przypadku osób z klasy 2, oraz zwykle kupują niewielkie ilości piwa. Najbardziej prawdopodobna jest przynależność mężczyzn niż kobiet, studentów podyplomowych oraz uczniów. Zwiększanie poziomu wykształcenia respondenta zwiększa prawdopodobieństwo przynależności osób do tej klasy. Podobnie stanowisko kierownicze, status pracodawcy, miejsce zamieszkania w mieście do 50 tys. mieszkańców lub na wieś lub miasto do 100 tys. mieszkańców zwiększają prawdopodobieństwo przynależności respondenta do tej klasy.

Klasa 2 to osoby rzadziej spożywające piwo, przy czym powody spożywania piwa są podobne jak w przypadku osób z klasy 1. Zwiększające się kwoty miesięcznych wydatków na konsumpcję piwa oraz ilość zakupywanego piwa zwiększają prawdopodobieństwo przynależności respondentów do tej klasy. W klasie tej znajdu-

ją się głównie studenci, mniej jest natomiast osób o innym statusie edukacyjnym. Większe prawdopodobieństwo znalezienia się w tej klasie mają małżeństwa, osoby rozwiedzione, owdowiałe oraz osoby niepozostające w żadnym związku.

4. Podsumowanie

Regresja klas ukrytych, a w pewnym uogólnieniu – analiza klas ukrytych, jest z pewnością użytecznym narzędziem w analizie danych mikroekonometrycznych – zwłaszcza w przypadku analizy danych związanych z preferencjami konsumentów. W artykule przedstawiono zastosowanie regresji klas ukrytych w metodach wyborów dyskretnych, gdzie głównym celem jest wskazanie, które ze zmiennych wpływają na nieznaną strukturę klas.

Dzięki wykorzystaniu regresji klas ukrytych z zastosowaniem pakietu `flexmix` programu R odkryto nieznaną wcześniej strukturę dwóch klas. W klasie pierwszej znalazły się osoby częściej spożywające piwo przy różnych okazjach. Jednakże osoby z tej klasy wydają jednorazowo na zakup piwa mniejsze kwoty pieniędzy. W klasie drugiej znalazły się osoby spożywające piwo rzadko, jednakże zwiększające się kwoty wydatków na piwo oraz ilość zakupywanego piwa zwiększają prawdopodobieństwo przynależności obiektów do tej klasy.

Można ogólnie podsumować charakterystyki klas w ten sposób, że klasa 1 to osoby częściej spożywające piwo, w niewielkich ilościach, a co za tym idzie – wydające na piwo niewiele (stali konsumenci). Natomiast klasa 2 to osoby kupujące piwo rzadziej, ale w większych ilościach i za większe kwoty pieniężne (konsumenci sporadyczni, „imprezowicze").

Literatura

Bartholomew D.J., Knott M., *Latent Variable Models and Factor Analysis*, Arnold Kendall's Library of Statistics, London 2002.

Bąk A., Rybicka A., *Application of Discrete Choice Methods in Consumer Preference Analysis*, [w:] D. Baier, K.-D. Wernecke (red.), *Innovations in Classification, Data Science, and Information Systems*, Proc. 27th Annual GfKL Conference, University of Cottbus, March 12-14, 2003, Springer-Verlag, Heidelberg-Berlin 2005.

Gatnar E., Walesiak M. (red.), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa 2011.

Gruszczyński M. (red.), *Mikroekonometria. Modele i metody analizy danych indywidualnych,* Wolters Kluwer Polska Sp. z o. o., Warszawa 2010.

Hagenaars J.A., McCutheon A.L., *Applied Latent Class Analysis*, Cambridge University Press, Cambridge 2002.

Lazerfeld P.F., Henry N.W., *Latent Structure Analysis*, Houghton Miffilin, Boston 1968.

Leish F., *FlexMix: A general framework for finite mixture models and latent class regression in R*, "Journal of Statistical Software" 2004, vol. 8, Issue 11.

Magidson J., Vermunt J. K., *A Nontechnical Introduction to Latent Class Models*, Statistical Innovations White Paper #1, www.statisticalinnovations.com, 2002.

Magidson J., Vermunt J.K., *Latent Class Models*, [w:] *The Sage Handbook of Quantitative Methodology and Social Sciences*, D. Kaplan (red.), Sage Publications, California 2004.
McCutheon A.L., *Latent Class Analysis*, Sage Publications, California 1987.
McLachlan G., Peel D., *Finite mixture models*, Wiley & Sons, New York 2000.
Vermunt J. K., Magidson J., *Latent Variable*, Encyclopedia of Social Science Research Methods, Sage Publications, www.statisticalinnovations.com, 2003.

APPLICATION OF LATENT CLASS REGRESSION IN THE ANALYSIS OF MICROECONOMETRIC DATA

Summary: The paper presents a possibility of application of latent class regression in discrete choice analysis. The main aim is to present which variables have significant influence on unknown class structure. The application of latent class regression, with application of `flexmix` package of R software, allowed to discover two, prior unknown, clusters of light beer consumers. Cluster 1 – people that drink beer more often, but not too much at once and they spend less money. Cluster 2 – people that drink beer from time to time but they buy much more beer and they spend more money.

Keywords: latent class regression, preference analysis, discrete choice methods.