

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 327

Taksonomia 22

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

The Central and Eastern European Online Library www.ceeol.com,

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się

na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2014

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Eugeniusz Gatnar , Balance of payments statistics and external competitiveness of Poland.....	15
Andrzej Sokolowski, Magdalena Czaja , Efektywność metody k -średnich w zależności od separowalności grup.....	23
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw	30
Elżbieta Gołata , Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011.....	39
Aleksandra Łuczak, Feliks Wysocki , Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów	49
Marek Walesiak , Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej	60
Paweł Lula , Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu <i>RAKE</i>	69
Mariusz Kubus , Propozycja modyfikacji metody złagodzonego LASSO.....	77
Andrzej Bąk, Tomasz Bartłomowicz , Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie <i>DiscreteChoice</i> programu R.....	85
Justyna Brzezińska , Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012.....	95
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Zastosowanie pakietu <i>dcMNM</i> programu R w badaniach preferencji konsumentów wódki	104
Barbara Batóg, Jacek Batóg , Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010	113
Małgorzata Markowska, Danuta Strahl , Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego.....	121
Kamila Migdał-Najman, Krzysztof Najman , Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	131
Kamila Migdał-Najman, Krzysztof Najman , Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	139
Beata Basiura, Anna Czapkiewicz , Badanie jakości klasyfikacji szeregów czasowych	148
Michał Trzęsiok , Wybrane metody identyfikacji obserwacji oddalonych.....	157

Grażyna Dehnel, Tomasz Klimanek , Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej.....	167
Michał Bernard Pietrzak, Justyna Wilk , Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji.....	177
Maciej Beręsewicz , Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena.....	186
Marcin Szymkowiak, Tomasz Józefowski , Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji.....	195
Marcin Pelka , Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym.....	202
Małgorzata Machowska-Szewczyk , Ocena klas w rozmytej klasyfikacji obiektów symbolicznych.....	210
Justyna Wilk , Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych.....	220
Andrzej Dudek , Metody analizy skupień w klasyfikacji markerów map Google.....	229
Ewa Roszkowska , Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW.....	237
Marcin Szymkowiak, Marek Witkowski , Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych.....	248
Bartłomiej Jefmański , Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R.....	257
Karolina Bartos , Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych.....	266
Joanna Trzęsiok , Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych.....	275
Beata Bal-Domańska , Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja.....	285
Beata Bieszk-Stolorz, Iwona Markowicz , Wpływ zasiłku na proces poszukiwania pracy.....	294
Marta Dziechciarz-Duda, Klaudia Przybysz , Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni.....	303
Tomasz Klimanek , Problem pomiaru procesu dezagrarnizacji wsi polskiej w świetle wielowymiarowych metod statystycznych.....	313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Wybrane metody analizy danych wzdluznych.....	321
Artur Zaborski , Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych.....	330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy.....	338

Katarzyna Wawrzyniak , Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego	346
---	-----

Summaries

Eugeniusz Gatnar , Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski	22
Andrzej Sokółowski, Magdalena Czaja , Cluster separability and the effectiveness of k -means method	29
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Multisectoral analysis of latent transitions in bankruptcy prediction models.....	38
Elżbieta Golata , Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011	48
Aleksandra Łuczak, Feliks Wysocki , Determination of weights for features in problems of linear ordering of objects	59
Marek Walesiak , Reinforcing measurement scale for ordinal data in multivariate statistical analysis	68
Paweł Lula , Automatic identification of keywords and keyphrases in documents written in Polish.....	76
Mariusz Kubus , The proposition of modification of the relaxed LASSO method.....	84
Andrzej Bąk, Tomasz Bartłomowicz , Microeconomic multinomial logit models and their implementation in the <code>DiscreteChoice</code> R package .	94
Justyna Brzezińska , The analysis of unemployment data in Poland in 2004-2012 with application of log-linear models	103
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Application of the MMLM package of R software for vodka consumers preference analysis.....	112
Barbara Batóg, Jacek Batóg , Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity	120
Małgorzata Markowska, Danuta Strahl , Classification of the European regional space in terms of smart growth pillars using the reference limit system.....	130
Kamila Migdał Najman, Krzysztof Najman , Formal quality assessment of group structure mapping on the Kohonen's map	138
Kamila Migdał Najman, Krzysztof Najman , Graphical quality assessment of group structure mapping on the Kohonen's map	147
Beata Basiura, Anna Czapkiewicz , Validation of time series clustering	156
Michał Trzęsiok , Selected methods for outlier detection.....	166

Grażyna Dehnel, Tomasz Klimanek , Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics	176
Michał Bernard Pietrzak, Justyna Wilk , Economic distance in modeling spatial phenomena with the application of gravity model.....	185
Maciej Beręsewicz , An attempt to use different distance measures in the Generalized Petersen estimator	194
Marcin Szymkowiak, Tomasz Józefowski , Construction and practical using of SPREE estimators for two-dimensional contingency tables.....	201
Marcin Pelka , The ensemble conceptual clustering for symbolic data.....	209
Małgorzata Machowska-Szewczyk , Evaluation of clusters obtained by fuzzy classification methods for symbolic objects.....	219
Justyna Wilk , Problem of determining the number of clusters in taxonomic analysis of symbolic data	228
Andrzej Dudek , Clustering techniques for Google maps markers.....	236
Ewa Roszkowska , The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure	247
Marcin Szymkowiak, Marek Witkowski , The use of correspondence analysis in analysing the financial situation of cooperative banks.....	256
Bartłomiej Jefmański , The construction of fuzzy customer satisfaction indexes using R program.....	265
Karolina Bartos , Discovering patterns of consumer behaviour by market basket analysis of the transactional data.....	274
Joanna Trzęsiok , Cluster analysis of countries with respect to fertility rate and other demographic factors	284
Beata Bal-Domańska , An attempt to identify major regional clusters and their convergence	293
Beata Bieszk-Stolorz, Iwona Markowicz , The influence of benefit on the job finding process	302
Marta Dziechciarz-Duda, Klaudia Przybysz , Education and labor market needs. Classification of university graduates	312
Tomasz Klimanek , The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods.....	320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Selected methods for an analysis of longitudinal data.....	329
Artur Zaborski , The application of distance measures for ordinal data for aggregation individual preferences	337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market	345
Katarzyna Wawrzyniak , Classification of voivodeships according to the level of the realization of priorities of <i>the National Development Strategy 2007-2015</i> with using the values of centroid of the rows	355

Andrzej Bąk, Marcin Pelka, Aneta Rybicka

Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE PAKIETU *dcmnm* PROGRAMU R W BADANIACH PREFERENCJI KONSUMENTÓW WÓDKI

Streszczenie: Głównym celem artykułu jest zastosowanie modeli logitowych w analizie preferencji konsumentów wódki. W tym celu wykorzystano pakiet *dcmnm* (*discrete choice MultiNominal Models*) programu R. Artykuł prezentuje podstawowe pojęcia z zakresu modeli logitowych oraz rezultaty szacowania preferencji. Rezultaty te pozwalają na wskazanie najlepszych oraz najgorszych marek wódki i tych atrybutów, które mają największe znaczenie dla konsumentów.

Słowa kluczowe: badania preferencji, wielomianowe modele logitowe, metody wyborów dyskretnych, program R.

1. Wstęp

Preferencje pozwalają na wyjaśnienie, jak i dlaczego konsumenci dokonują swoich wyborów. W badaniach preferencji znajdują zastosowanie metody wyborów dyskretnych. Pozwalają one na odzwierciedlenie wyborów konsumentów dokonywanych pomiędzy różnymi ofertami (profilami). Modele w ramach metod wyborów dyskretnych mogą być różnego typu. Są to: wielomianowy model logitowy (*MultiNominal Logit Model* – MNL), warunkowy model logitowy (*Conditional Logit Model* – CLM) i mieszany model logitowy (*Mixed Logit Model* – MLM). Podstawą rozróżnienia tych modeli jest charakter zmiennych objaśniających uwzględnionych w modelu.

Głównym celem artykułu jest wykorzystanie wielomianowych modeli logitowych i pakietu *dcmnm* opracowanego dla programu R [Bąk 2013a] w analizie preferencji konsumentów wódki. W artykule przedstawiono podstawowe pojęcia z zakresu modeli logitowych oraz wyniki szacowania preferencji. Rezultaty te pozwalają na wskazanie najlepszych i najgorszych marek wódki oraz tych atrybutów, które mają największe znaczenie dla konsumentów.

2. Wielomianowe modele kategorii nieuporządkowanych i ich estymacja

Realizacje zmiennych w modelach mikroekonometrycznych są najczęściej wynikami pomiarów na skalach słabych (niemetrycznych). Zgromadzone obserwacje są zwykle liczbowymi wartościami dyskretnymi.

W mikroekonometrii występują następujące rodzaje zmiennych objaśnianych [zob. Gatnar, Walesiak (red.) 2011, s. 113; Gruszczyński (red.) 2012; Bąk 2013b]:

- a) zmienne dychotomiczne (dwukategorialne, np. binarne),
- b) zmienne politomiczne (wielokategorialne), wśród których wyróżnia się zmienne o kategoriach uporządkowanych i nieuporządkowanych,
- c) zmienne ograniczone, wśród których wyróżnia się zmienne cenzurowane i ucięte,
- d) zmienne licznikowe.

W mikroekonometrii stosuje się najczęściej następujące rodzaje modeli [zob. Gatnar, Walesiak (red.) 2011, s. 113; Gruszczyński 2012 (red.); Bąk 2013b]:

- a) modele dwumianowe:
 - liniowe modele prawdopodobieństwa,
 - modele logitowe i probitowe,
 - modele komplementarne log-log,
 - modele log-liniowe;
- b) modele wielomianowe:
 - modele kategorii nieuporządkowanych,
 - modele kategorii uporządkowanych;
- c) modele klas ukrytych;
- d) modele przeżycia (trwania);
- e) modele zmiennych ograniczonych.

Istotne miejsce wśród tych modeli zajmują: wielomianowy model logitowy, warunkowy model logitowy oraz mieszany model logitowy. Wielomianowy model logitowy jest uogólnieniem modelu logitowego dla danych binarnych. Model tego typu może być stosowany, gdy zmienna objaśniana przyjmuje w sposób dyskretny wartości ze zbioru liczącego więcej niż dwie kategorie [Gatnar, Walesiak (red.) 2011, s. 113]. Podstawami teoretycznymi tego modelu są teoria użyteczności losowej oraz aksjomat wyboru Luce'a [Coombs, Dawes i Tversky 1977, s. 217 i n.].

Wielomianowy model logitowy można przedstawić jako [Agresti 2002, s. 267-268; Gatnar, Walesiak (red.) 2011, s. 113-114; Bąk 2004, s. 118-120]:

$$P_{ki} = \frac{\exp(\mathbf{x}_k^T \boldsymbol{\beta}_i)}{\sum_{l=1}^n \exp(\mathbf{x}_k^T \boldsymbol{\beta}_l)}, \quad (1)$$

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, przy czym $\boldsymbol{\beta}_n = \mathbf{0}$.

Warunkowy model logitowy został zaproponowany przez McFaddena [1974] jako uogólnienie wielomianowego modelu logitowego. Podstawową różnicą jest charakter zmiennych objaśniających, tzn. macierzy \mathbf{X} , w równaniu (1). Jeżeli zmienne objaśniające charakteryzują konsumentów, to na ogół wykorzystuje się wielomianowy model logitowy. Jeżeli natomiast zmienne objaśniające opisują obiekty będące przedmiotem wyboru, to z reguły stosuje się warunkowy model logitowy [Gatnar, Walesiak (red.) 2011, s. 114].

Warunkowy model logitowy można przedstawić jako [Gatnar, Walesiak (red.) 2011, s. 114; So, Kuhfeld 1995; Bąk 2004, s. 120-122]:

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

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} we wzorze (2) zawiera charakterystyki produktów lub usług, względem których badane są preferencje respondentów.

Mieszany model logitowy stanowi połączenie cech jednostek (osób) i cech opcji (alternatyw) w jednym modelu. Mieszany model logitowy można przedstawić jako [So, Kuhfeld 1995]:

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

gdzie: oznaczenia jak we wzorach (1) i (2).

Do estymacji parametrów warunkowego modelu logitowego wykorzystuje się funkcję największej wiarygodności w postaci:

$$\log L(\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = \sum_{k=1}^p \sum_{i=1}^n d_{ki} \log P_{ki}, \quad (4)$$

gdzie: $d_{ki} = 1$, jeśli wybrano opcję i ($d_{ki} = 0$ w przeciwnym przypadku); k – numer konsumenta (osoby); i – numer opcji wyboru; \mathbf{y} – zmienna zależna (objaśniana); \mathbf{X} – zmienne objaśniające (np. charakterystyki konsumentów).

W programie R do optymalizacji funkcji największej wiarygodności (4) wykorzystuje się funkcję `optim` z pakietu `stats`. Wymaga ona podania wartości początkowych parametrów do optymalizacji oraz funkcji, która będzie poddawana optymalizacji, a także zmiennych zależnych oraz objaśniających. Na potrzeby pakietu dcMNM [Bąk 2013a] programu R opracowano funkcję największej wiarygodności na podstawie pracy [Jackman 2007].

3. Zastosowanie modeli wielomianowych w badaniu preferencji konsumentów wódki

W 2012 r. przygotowano i przeprowadzono badanie ankietowe dotyczące preferencji konsumentów wódki (w badaniu wzięli udział mieszkańcy Jeleniej Góry i okolic). Głównym celem badania była identyfikacja i analiza preferencji respondentów w odniesieniu do przedmiotu badania.

W badaniu ankietowym uwzględniono pięć atrybutów opisujących wódkę:

a) kraj – z pięcioma poziomami: Polska, Rosja, Szwecja, Finlandia, Niemcy,

b) objętość – z trzema poziomami: 0,5 l, 0,7-0,75 l, 1 litr i więcej,

c) smak – z czterema poziomami: czysta, owocowa, ziołowa, wytrawna,

d) zawartość alkoholu – z trzema poziomami: 40%, 45%, 50% i więcej,

e) cena – z czterema poziomami: poniżej 20 zł, 20-40 zł, 40-100 zł, powyżej 100 zł.

Pełny układ czynnikowy liczy 720 profilów ($5 \times 3 \times 4 \times 3 \times 4$). Za pomocą funkcji z pakietu `AlgDesign` programu R wygenerowano cząstkowy układ czynnikowy z podziałem na bloki liczące 360 profilów. Liczba bloków wynosiła 4 (przykładowy zbiór z bloku 1 prezentuje tab. 1). Liczba zbiorów w każdym bloku – 15. Liczba opcji w każdym zbiorze – 6 (5 profilów plus opcja rezygnacji z wyboru). Łącznie w bloku 1 zebrano 157 ankiet, w bloku 2 – 123 ankiet, w bloku 3 – 140 ankiet, a w bloku 4 – 134 ankiet. Łącznie zgromadzono 49 860 obserwacji. W trakcie badania ankietowego (próbę miała charakter przypadkowy) zebrano 544 kwestionariusze poprawnie wypełnionych ankiet.

Tabela 1. Przykładowy zbiór profilów wyboru z bloku pierwszego

Kraj produkcji	Objętość	Smak	Zawartość alkoholu	Cena	Wybieram opcję
Niemcy	1 l i więcej	ziołowa	40%	poniżej 20 zł	1
Niemcy	0,5 l	czysta	45%	poniżej 20 zł	2
Szwecja	0,7-0,75 l	czysta	45%	poniżej 20 zł	3
Finlandia	1 l i więcej	czysta	45%	poniżej 20 zł	4
Szwecja	1 l i więcej	owocowa	45%	poniżej 20 zł	5
Żaden z powyższych					6

Źródło: opracowanie własne.

W prezentowanej analizie uwzględniono pierwszy blok danych – tj. 90 profili (15 zbiorów po 6 opcji), które oceniło 157 respondentów, co daje łącznie 14 130 obserwacji. Do estymacji warunkowego modelu logitowego zastosowano skrypt korzystający z pakietu `dcMNM`:

```
library(dcMNM)
x<-read.csv2("wodka_X1.csv",header=TRUE)
y<-read.csv2("wodka_Y1.csv",header=TRUE)
head(CLMdata(x,y))
AS<-CLMattrsel(x,y)
print(AS$vif)
print(head(AS$W))
clm<-CLMmodel(x,y)
print(clm)
CLMgraph(x,y)
Pclm<-CLMprob(x,y)
print(Pclm)
```

Polecenie `head(CLMdata(x,y))` wyświetla fragment zbioru danych o strukturze wymaganej do estymacji warunkowego modelu logitowego.

choice	kraj1	kraj2	kraj3	kraj4	kraj5	objetoscl	objetosc2	objetosc3	smak1
0	0	0	0	0	1	0	0	1	0
0	0	0	0	0	1	1	0	0	1
0	0	0	1	0	0	0	1	0	1
0	0	0	0	1	0	0	0	1	1
0	0	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	0
smak2	smak3	smak4	alkohol1	alkohol2	alkohol3	cena1	cena2	cena3	cena4
0	1	0	1	0	0	1	0	0	0
0	0	0	0	1	0	1	0	0	0
0	0	0	0	1	0	1	0	0	0
0	0	0	0	1	0	1	0	0	0
1	0	0	0	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0

Atrybuty opisujące opcje wyboru są reprezentowane za pomocą zmiennych sztucznych otrzymanych w wyniku kodowania zero-jedynkowego (są to zmienne objaśniające w modelu). Zmienną objaśnianą, reprezentującą wybory respondentów, jest `choice`. Polecenie `AS<-CLMattrsel(x,y)` umożliwia usunięcie współliniowości występującej między zmiennymi sztucznymi na podstawie czynnika wzrostu wariancji *VIF* (*Variance Inflation Factor*) [zob. np. Maddala 2006; Greene 2008]. W rezultacie w dalszej analizie zostały uwzględnione zmienne sztuczne, dla których czynnik wzrostu wariancji *VIF* nie przekroczył wartości 10¹:

¹ Przyjmuje się, że wartość czynnika wzrostu wariancji *VIF* większa lub równa 10 wskazuje na istotną współliniowość zmiennych [zob. np. Dobosz 2004, s. 207; Orme, Combs-Orme 2009, s. 27].

```

print(AS$vif)
  kraj1   kraj2   kraj3   kraj4   kraj5   objetosc1 objetosc2
2.766260 1.600265 3.610946 2.817511 4.276752 1.959654 1.984138
  smak1   smak2   smak3   alkohol1 alkohol2   cena1   cena2
2.060787 1.788195 1.900510 1.623302 2.703496 1.565596 3.643379
  cena3
1.919827
print(head(AS$W))
  choice kraj1 kraj2 kraj3 kraj4 kraj5 objetosc1 objetosc2 smak1 smak2
1      0      0      0      0      0      1      0      0      0      0
2      0      0      0      0      0      1      1      0      1      0
3      0      0      0      1      0      0      0      1      1      0
4      0      0      0      0      1      0      0      0      1      0
5      0      0      0      1      0      0      0      0      0      1
6      1      0      0      0      0      0      0      0      0      0
  smak3 alkohol1 alkohol2 cena1 cena2 cena3
1      1      1      0      1      0      0
2      0      0      1      1      0      0
3      0      0      1      1      0      0
4      0      0      1      1      0      0
5      0      0      1      1      0      0
6      0      0      0      0      0      0

```

Polecenie `clm<-CLMmodel(x, y)` szacuje parametry modelu (`coef`) oraz oblicza błędy standardowe (`se`), wartości statystyki (`t`), wartości *p-value* (`Pr(>|t|)`) i ilorazy hazardu (`exp(coef)`):

```

print(clm)
$estimate
      coef      se      t      Pr(>|t|) exp(coef)
kraj1 -0.5654864 0.1422622 -3.9749589 0.0000704 0.5680838
kraj2 -0.5680860 0.1690173 -3.3611116 0.0007763 0.5666089
kraj3 -0.5921168 0.1399808 -4.2299858 0.0000234 0.5531551
kraj4 -0.4759962 0.1270159 -3.7475324 0.0001786 0.6212658
kraj5 -0.5286539 0.1353328 -3.9063250 0.0000937 0.5893978
objetosc1 0.1319658 0.0605365 2.1799377 0.0292621 1.1410693
objetosc2 -0.0967864 0.0667818 -1.4492931 0.1472557 0.9077499
smak1 -0.0189597 0.0754349 -0.2513386 0.8015523 0.9812189
smak2 -0.0419719 0.0915944 -0.4582365 0.6467825 0.9588967
smak3 -0.0737781 0.0801963 -0.9199689 0.3575890 0.9288778
alkohol1 -0.1033741 0.0978124 -1.0568609 0.2905751 0.9017896
alkohol2 -0.1040915 0.0847939 -1.2275824 0.2196038 0.9011428
cena1 0.3853024 0.1651003 2.3337474 0.0196090 1.4700588
cena2 0.2416567 0.1154108 2.0938829 0.0362704 1.2733570
cena3 0.0238967 0.1305776 0.1830077 0.8547920 1.0241845

$logLik
[1] -4166.118

$McFaddenR2
[1] 0.0007013

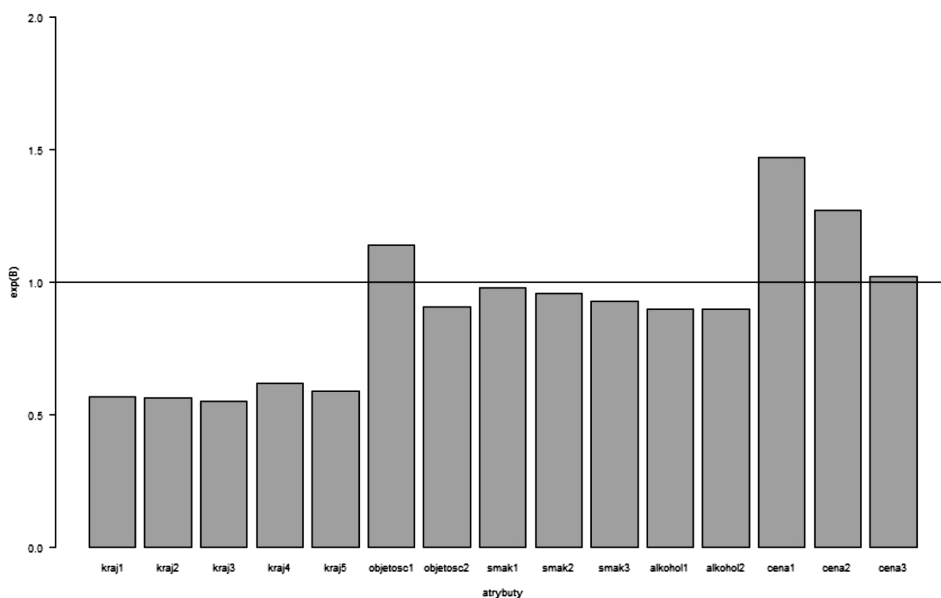
$LR
df Chisq Pr(>Chisq)
[1,] 10 5.847191 0.827934

```

Dopasowanie modelu do danych empirycznych jest oceniane na podstawie wartości współczynnika determinacji pseudo R^2 McFaddena (McFaddenR2) [McFadden 1974].

W celu interpretacji wyników estymacji parametrów wykorzystano ilorazy hazardu przedstawione graficznie za pomocą polecenia `CLMgraph(x, y)` na rys. 1. Wartości ilorazów hazardu informują o stymulującym (wartości większe od 1) lub destymulującym (wartości mniejsze od 1) wpływie atrybutu na prawdopodobieństwo wyboru określonej opcji.

Analiza wyników zilustrowanych na rys. 1 pozwala stwierdzić, że stymulująco na prawdopodobieństwo wyboru wódki wpływają następujące poziomy atrybutów: objętość 0,5 l, cena do 20 zł, cena 20-40 zł oraz cena 40-100 zł. Najbardziej destymulująco na wybór wpływa kraj pochodzenia (wszystkie poziomy atrybutu są traktowane tak samo), a nieco mniej destymulująco – smak wódki, zawartość alkoholu powyżej 40%.



Rys. 1. Wykres ilorazów hazardu dla parametrów różnych od zera

Źródło: opracowanie własne z wykorzystaniem programu R.

4. Podsumowanie

Mikroekonometryczne wielomianowe modele logitowe znajdują zastosowanie w analizie preferencji wyrażonych opartych na metodzie wyborów dyskretnych. W metodzie tej różne typy modeli wielomianowych mogą być szacowane z wyko-

rzystaniem tej samej procedury wykorzystującej iteracyjne algorytmy optymalizacyjne. Jednakże estymacja różnych typów modeli wielomianowych wymaga różnej struktury danych empirycznych. Pakiet dcMNM umożliwia reorganizację danych empirycznych do postaci struktur wymaganych dla różnych typów modeli oraz estymację parametrów tych modeli i szacowanie prawdopodobieństw wyboru poszczególnych profili.

Kierunkiem dalszych badań powinny być oszacowania modelu wielomianowego i mieszanego i interpretacja wyników oraz integracja pakietu dcMNM z rozwiązaniami typowymi dla stron WWW w celu gromadzenia danych empirycznych.

Literatura

- Agresti A. (2002), *Categorical Data Analysis*, John Wiley & Sons, Hoboken, New Jersey.
- Bąk A. (2004), *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*, Wyd. Akademii Ekonomicznej we Wrocławiu, Wrocław.
- Bąk A. (2013a), *Discrete choice multinomial models – package dcMNM*, [URL:] <http://keii.ue.wroc.pl/dcMNM>.
- Bąk A. (2013b), *Mikroekonometryczne metody badania preferencji konsumentów z wykorzystaniem programu R*, Wydawnictwo C.H. Beck, Warszawa.
- Cameron A.C., Trivedi P.K. (2005), *Microeconometrics. Methods and applications*, Cambridge University Press, New York.
- Coombs C.H., Dawes R.M., Tversky A. (1977), *Wprowadzenie do psychologii matematycznej*, PWN, Warszawa.
- Dobosz M. (2004), *Wspomagana komputerowo statystyczna analiza wyników badań*, wydanie drugie uaktualnione, Akademicka Oficyna Wydawnicza EXIT, Warszawa.
- Gatnar E., Walesiak M. (red.) (2011), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa.
- Greene W.H. (2008), *Econometric Analysis*, Sixth Edition, Pearson Prentice Hall, Upper Saddle River.
- Gruszczyński M. (red.) (2012), *Mikroekonometria. Modele i metody analizy danych indywidualnych*, wydanie drugie rozszerzone, Wolters Kluwer, Warszawa.
- Jackman S. (2007), *Models for unordered outcomes*, Political Science 150C/350C [URL:] <http://jackman.stan-ford.edu/classes> (2.05.2009).
- Maddala G.S. (2006), *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.
- McFadden D. (1974), *Conditional logit analysis of qualitative choice behavior*, [w:] P. Zarembka (red.), *Frontiers in Econometrics*, Academic Press, New York – San Francisco – London.
- Orme J.G., Combs-Orme T. (2009), *Multiple Regression with Discrete Dependent Variables*, Oxford University Press, Oxford – New York.
- So Y., Kuhfeld W.F. (1995), *Multinomial Logit Models*, [URL:] <http://support.sas.com/techsup/technote/mr2010g.pdf> (12.03.2012).

APPLICATION OF THE MMLM PACKAGE OF R SOFTWARE FOR VODKA CONSUMERS PREFERENCE ANALYSIS

Summary: The main aim of the paper is to apply logit models in preference analysis of vodka consumers. In order to obtain such a goal `dcMNM` (*discrete choice MultiNominal Models*) package of R software was applied. The paper presents basic terms of logit models and the results of preference analysis. The estimates allowed to identify the best and the worst vodka brands as well as the attributes that have the biggest influence on preferences.

Keywords: preference analysis, multinomial logit model, discrete choice models, R program.