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 naukowi
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

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 Sokołowski, Magdalena Czaja, Efektywność metody *k*-średnich w zależności od separowalności grup 23

Barbara Pawełek, Józef Pociecha, 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 *RAKE* 69

Mariusz Kubus, Propozycja modyfikacji metody złagodzonego LASSO 77

Andrzej Bąk, Tomasz Bartłomowicz, Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie `DiscreteChoice` programu R 85

Justyna Brzezińska, Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012 95

Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Zastosowanie pakietu `dcMNM` 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 Pełka, 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 dezagraryzacji wsi polskiej w świetle wielowymiarowych metod statystycznych ... 313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Wybrane metody analizy danych wzdłużnych ... 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 Sokołowski, Magdalena Czaja, Cluster separability and the effectiveness of *k*-means method 29
Barbara Pawełek, Józef Pociecha, Adam Sagan, Multisectoral analysis of latent transitions in bankruptcy prediction models..... 38
Elżbieta Gołata, 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 `DiscreteChoice` 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 Pełka, 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 Pełka, 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 *the National Development Strategy 2007-2015* with using the values of centroid of the rows 355

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Bąk, Tomasz Bartłomowicz
Uniwersytet Ekonomiczny we Wrocławiu

WIELOMIANOWE MODELE LOGITOWE WYBORÓW DYSKRETNYCH I ICH IMPLEMENTACJA W PAKIECIE `DiscreteChoice` PROGRAMU R

Streszczenie: Celem artykułu jest prezentacja pakietu `DiscreteChoice`, opracowanego dla środowiska R, który może być stosowany do analizy wyborów konsumentów ze zbioru dostępnych opcji. Pakiet zawiera implementację wielomianowego, warunkowego oraz mieszanego modelu, a także funkcje, które mogą być zastosowane w metodzie wyborów dyskretnych do konstrukcji eksperymentu m.in. budowy cząstkowego układu czynnikowego, kodowania zbioru opcji i estymacji modelu. W artykule przedstawione zostały przykłady zastosowania funkcji pakietu `DiscreteChoice` w analizie preferencji konsumentów.

Słowa kluczowe: mikroekonometria, preferencje, wielomianowe modele logitowe wyborów dyskretnych, program R.

1. Wstęp

W badaniach preferencji wyrażonych stosowane są najczęściej metody dekompozycyjne, których idea polega na rozkładzie (dekompozycji) preferencji empirycznych (wyrażonych na przykład w badaniu ankietowym) na użyteczności cząstkowe poziomów atrybutów opisujących badane profile produktów lub usług. Wśród metod dekompozycyjnych wyróżnia się dwie podstawowe grupy: metody *conjoint analysis* oraz metody wyborów dyskretnych.

Podstawę teoretyczną modeli wyborów dyskretnych stanowi teoria użyteczności losowej [Coombs, Dawes, Tversky 1977, s. 214], która umożliwia kwantyfikację użyteczności, a w konsekwencji identyfikację czynników, którymi kierują się konsumenci, wybierając określone produkty lub usługi. Czynniki te mogą być związane zarówno z charakterystykami konsumentów, jak i atrybutami produktów lub usług. Oznacza to, że w badaniu preferencji konsumentów zastosowanie znajdują modele kategorii nieuporządkowanych, wśród których najczęściej stosowane modele to: wielomianowy model logitowy, warunkowy model logitowy oraz mieszany model logitowy [Winkelmann, Boes 2006].

W literaturze przedmiotu rozróżnienie między tymi modelami nie jest jednoznacznie interpretowane. Celem artykułu jest omówienie podstawowych różnic między wyróżnionymi modelami logitowymi [Bąk 2012] oraz prezentacja pakietu `DiscreteChoice` opracowanego dla programu R [Bąk, Bartłomowicz 2013a], którego funkcje umożliwiają realizację procedury badawczej opartej na wyborach z wykorzystaniem modeli wielomianowych. Funkcje pakietu zilustrowano przykładem zastosowania w empirycznych badaniach preferencji wyrażonych konsumentów.

2. Wielomianowe modele logitowe wyborów dyskretnych

W badaniach marketingowych prowadzonych za pomocą metod wyborów dyskretnych znajdują zastosowanie modele kategorii nieuporządkowanych w postaci wielomianowego (*MultiNomial Logit Model* – MNLM) i warunkowego (*Conditional Logit Model* – CLM) modelu logitowego oraz mieszanego modelu logitowego (*Mixed Logit Model* – MLM), będącego połączeniem wymienionych modeli [Cameron, Trivedi 2009, s. 500].

Wielomianowy model logitowy jest uogólnieniem modelu logitowego dla danych binarnych i może być stosowany, gdy zmienna objaśniana przyjmuje w sposób dyskretny wartości ze zbioru liczącego więcej niż dwie kategorie. Model wywodzi się z teorii użyteczności losowej oraz tzw. aksjomatu wyboru Luce'a (modelu stałej użyteczności) [Coombs, Dawes, Tversky 1977, s. 217; Bierlaire 1997]. Wielomianowy model logitowy można przedstawić w postaci [So, Kuhfeld 1995; Long 1997, s. 151; Powers, Xie 2008, s. 243; Cameron, Trivedi 2009, s. 500; Gruszczyński (red.) 2010, s. 161]:

$$P_{ki} = \frac{\exp\left(x_k^T \beta_i\right)}{\sum_{l=1}^{n} \exp\left(x_k^T \beta_l\right)}, \tag{1}$$

gdzie: P_{ki} – prawdopodobieństwo wyboru i-tej kategorii przy k-tym stanie zmiennych objaśniających opisujących konsumentów; x_k^T – wektor reprezentujący k-ty wiersz macierzy **X** (wartości zmiennych objaśniających dla k-tego konsumenta); β_i – wektor parametrów związany z i-tą kategorią zmiennej objaśnianej.

W przypadku warunkowego modelu logitowego zaproponowanego przez McFaddena [McFadden 1974, s. 105-142] prawdopodobieństwo wyboru i-tego profilu ze zbioru liczącego n elementów jest szacowane na podstawie zależności [So, Kuhfeld 1995, s. 7; Long 1997, s. 178; Powers, Xie 2008, s. 256; Cameron, Trivedi 2009, s. 500; Gruszczyński (red.) 2010, s. 172-173]:

$$P_{ki} = \frac{\exp\left(z_{ki}^T \alpha\right)}{\sum_{l=1}^{n} \exp\left(z_{kl}^T \alpha\right)}, \tag{2}$$

gdzie: P_{ki} – prawdopodobieństwo wyboru *i*-tej kategorii przy *k*-tym stanie zmiennych objaśniających; z_{kl}^T – *k*-ty wektor macierzy **Z** (zmiennych objaśniających opisujących *i*-tą opcję wybraną przez *k*-tego konsumenta); α – wektor parametrów.

Zarówno wielomianowe, jak i warunkowe modele logitowe wykorzystywane są do analizy wyborów indywidualnych ze zbioru dostępnych alternatyw (kategorii, opcji, profilów), przy czym w modelu wielomianowym (1) szacuje się prawdopodobieństwo wyboru *i*-tej opcji z uwzględnieniem zmiennych objaśniających charakteryzujących konsumentów, podczas gdy w modelu warunkowym (2) jest to prawdopodobieństwo wyboru *i*-tej opcji ze względu na wartości zmiennych objaśniających charakteryzujących tę opcję. Wynika to z założenia – co stanowi jednocześnie główne rozróżnienie tych modeli – że w wielomianowym modelu logitowym analizuje się cechy konsumentów i ich wpływ na wybory określonych opcji. W modelu warunkowym natomiast analizuje się cechy produktów lub usług (opcji wyboru) i ich wpływ na dokonywane przez konsumentów wybory.

W mieszanym modelu logitowym (3) prawdopodobieństwo wyboru *i*-tej kategorii przy *k*-tym stanie zmiennych objaśniających opisać można za pomocą zależności:

$$P_{ki} = \frac{\exp\left(x_k^T \beta_i + z_{ki}^T \alpha\right)}{\sum_{l=1}^{n} \exp\left(x_k^T \beta_l + z_{kl}^T \alpha\right)}. \tag{3}$$

Model mieszany uwzględnia zarówno cechy konsumentów, jak i cechy opisujące opcje wyboru. Oznacza to, że macierze zmiennych objaśniających uwzględniają zarówno indywidualne charakterystyki konsumentów, jak i charakterystyki produktów lub usług.

3. Pakiet `DiscreteChoice` programu R

Pakiet `DiscreteChoice` [Bąk, Bartłomowicz 2013a] to autorskie oprogramowanie zawierające implementację wielomianowych modeli logitowych wyborów dyskretnych dla programu R [R Development Core Team 2013], udostępniane na podstawie licencji GNU GPL. Korzystanie z pakietu wymaga zainstalowania wersji bazowej programu R oraz pakietów `AlgDesign` i `rms`. Pakiet można pobrać i zainstalować ze strony internetowej Katedry Ekonometrii i Informatyki Uniwersytetu Ekonomicznego we Wrocławiu[1].

W aktualnej wersji pakietu (1.00) oferowanych jest 12 funkcji, które umożliwiają (por. tab. 1): wygenerowanie układu badania w postaci cząstkowego układu czynnikowego z zadeklarowaną liczbą bloków oraz profilów w każdym z bloków (funkcja MNMdesign), przekonwertowanie danych o indywidualnych wyborach

[1] URL: http://keii.ue.wroc.pl/DiscreteChoice.

konsumentów na zbiory danych dla modeli logitowych (funkcje: MNLdata, CLMdata, MLMdata), estymację parametrów modeli logitowych (funkcje: MNLmodel, CLMmodel, MLMmodel) z wykorzystaniem funkcji `optim` [zob. Jackman 2007] oraz szacowanie prawdopodobieństw wyborów profilów według zadanych modeli (funkcje: MNLprob, CLMprob, MLMprob). Ponadto w roli funkcji specjalnego przeznaczenia wymienić należy funkcję CLMattrsel, której zadaniem jest

Tabela 1. Funkcje pakietu `DiscreteChoice`

Funkcje pakietu	
`MNMdesign(vars, names, blocks=3, profiles=6)` – funkcja generująca cząstkowy układ czynnikowy z sugerowaną liczbą bloków oraz profilów w każdym z bloków (wliczając profil „żaden z powyższych")	
`MNLdata(y, x, z)` – funkcja konwertująca macierz danych opisującą indywidualne wyboryna zbiór danych do wielomianowego modelu logitowego	
`CLMdata(x, y)` – funkcja konwertująca macierz danych opisującą indywidualne wybory na zbiór danych do warunkowego modelu logitowego	
`MLMdata(y, x, z)` – funkcja konwertująca macierz danych opisującą indywidualne wybory na zbiór danych do mieszanego modelu logitowego	
`MNLmodel(y, x, z)` – funkcja wyznacza wartości wyrazów wolnych dla indywidualnych predyspozycji jednostki (charakterystyk konsumentów), wykorzystując wielomianowy model logitowy; wartości parametrów reprezentują efekt wpływający na prawdopodobieństwo wyboru opcji (profilów) w stosunku do opcji (profilu) odniesienia	
`CLMmodel(x, y)` – funkcja wyznacza wartości wyrazów wolnych dla charakterystyk alternatyw (atrybutów produktów lub usług), wykorzystując warunkowy model logitowy	
`MLMmodel(y, x, z)` – funkcja wyznacza wartości wyrazów wolnych dla indywidualnych predyspozycji jednostki (charakterystyk konsumentów) oraz charakterystyk alternatyw (atrybutów produktów lub usług), wykorzystując mieszany (hybrydowy) model logitowy	
`MNLprob(y, x, z)` – funkcja szacująca prawdopodobieństwo wyboru profilów według wielomianowego modelu logitowego	
`CLMprob(x, y)` – funkcja szacująca prawdopodobieństwo wyboru profilów według warunkowego modelu logitowego	
`MLMprob(y, x, z)` – funkcja szacująca prawdopodobieństwo wyboru profilów według mieszanego modelu logitowego	
`CLMattrsel(x, y)` – funkcja generująca zbiór zmiennych (bez zmiennych powtarzających się) do warunkowego modelu logitowego	
`CLMgraph(x, y)` – funkcja generująca wykres ilorazu hazardu na podstawie modelu warunkowego	
Argumenty funkcji	
`vars`	wektor liczby poziomów zmiennych
`names`	wektor nazw zmiennych
`blocks`	liczba bloków (parametr opcjonalny), wartość domyślna: blocks=3
`profiles`	liczba profilów w każdym z bloków (parametr opcjonalny), wartość domyślna: profiles=6
`y`	wektor lub macierz zmiennych opisujących indywidualne wybory (zmienna objaśniana)
`x`	macierz zmiennych objaśniających, np. charakterystyk opcji wyboru (alternatyw)
`z`	macierz parametrów związanych z kategoriami zmiennej objaśnianej

Źródło: opracowanie własne.

selekcja zbioru zmiennych dla warunkowego modelu logitowego (pominięcie zmiennych sztucznych współliniowych) oraz funkcję CLMgraph, która generuje wykres ilorazu hazardu dla modelu warunkowego.

Należy zauważyć, że prezentowany pakiet wspiera nierealizowane dotychczas w żadnym z pakietów programu R elementy procedury metody wyborów dyskretnych.

Szczegółowa charakterystyka oraz przykłady zastosowania wszystkich funkcji dostępne są w dokumentacji pakietu `DiscreteChoice`.

4. Badanie preferencji z wykorzystaniem pakietu `DiscreteChoice`

W przykładzie ilustrującym sposób wykorzystania wielomianowych modeli logitowych identyfikacja i analiza preferencji respondentów dotyczy usług gastronomicznych [Bąk, Bartłomowicz 2013b]. W badaniu wytypowano 4 atrybuty (wraz z odpowiadającymi im poziomami): cenę (do 10 zł, 10-20 zł, powyżej 20 zł), miejsce konsumpcji (bar, restaurację, stołówkę, punkt gastronomiczny), rodzaj konsumpcji (posiłek, deser, napój) oraz godzinę konsumpcji (poranną, popołudniową, wieczorną). Badanie przeprowadzono wśród mieszkańców Jeleniej Góry i okolic w roku 2010, co ostatecznie umożliwiło wykorzystanie danych[2] ze 136 prawidłowo wypełnionych kwestionariuszy ankietowych.

Pomimo że w badaniu można było wykorzystać maksymalnie 108 różnych profilów usług gastronomicznych, ostatecznie na potrzeby badania wytypowano 3 zbiory danych po 6 profilów, w których na ostatniej (6., 12. oraz 18.) pozycji znalazły się profile odniesienia (indeksy kodowania o jeden większe niż liczba poziomów danej zmiennej):

```
> X<-MNMdesign(vars=c(3, 4, 3, 3), names=c(”price”, ”place”, ”kind”,
”time”), profiles=6, blocks=3)
> print(X)
  price place kind time
1     2     4    2    1
2     1     1    3    1
3     1     3    2    2
4     3     3    3    2
5     2     3    1    3
6     4     5    4    4
7     1     4    1    2
8     3     1    2    2
9     2     2    2    2
10    2     2    3    2
11    2     1    3    3
12    4     5    4    4
13    2     3    3    1
```

[2] Dane zostały zebrane przez M. Więcław.

```
14     2      1     1     2
15     3      4     3     2
16     3      4     2     3
17     1      2     3     3
18     4      5     4     4
```

Dla przykładu, blok pierwszy (wiersze 1-6 macierzy X) to zakodowany zbiór profilów (wraz z profilem odniesienia) prezentowany przez tab. 2.

Tabela 2. Przykładowy zbiór profilów do wyboru

Cena (*price*)	Miejsce konsumpcji (*place*)	Rodzaj konsumpcji (*kind*)	Godzina konsumpcji (*time*)	Wybór (*choice*)
10-20 zł	bar	deser	poranna	1
do 10 zł	restauracja	napój	poranna	2
do 10 zł	stołówka	deser	popołudniowa	3
powyżej 20 zł	stołówka	napój	popołudniowa	4
10-20 zł	stołówka	posiłek	wieczorna	5
żaden z profilów				6

Źródło: opracowanie własne.

Respondenci za każdym razem wybierali 1 z 6 profilów. Zmienną specyficzną dla respondenta jest wiek (`age`), wybraną opcję wyboru określa cena (`price`), natomiast wybrany profil reprezentuje zmienna (`choice`).

W wyniku zastosowania funkcji MNLmodel, w wielomianowym modelu logitowym otrzymuje się oszacowania parametrów dla charakterystycznej dla respondentów zmiennej specyficznej wiek (`age`):

```
> z1<-as.data.frame(Z[, 2])
> n<-names(Z)
> colnames(z1)<-n[2]
> MNLmodel(Y3, X3, z1)
$estimate
                coef        se           t  Pr(>|t|) exp(coef)
intercept1 -3.8384611 2.1502095 -1.78515680 0.0742359 0.0215267
intercept2  1.0990329 0.7011029  1.56757717 0.1169798 3.0012621
intercept3  1.0298781 0.9619891  1.07057149 0.2843621 2.8007244
intercept4 -0.0127671 1.0149001 -0.01257966 0.9899631 0.9873141
intercept5  1.9049344 0.8047372  2.36715092 0.0179256 6.7189668
age1        0.5098369 0.5252525  0.97065107 0.3317221 1.6650196
age2       -0.0460678 0.2040738 -0.22574088 0.8214029 0.9549772
age3       -0.5154486 0.3236660 -1.59253243 0.1112651 0.5972326
age4       -0.2220330 0.3141375 -0.70680196 0.4796895 0.8008889
age5       -0.5731541 0.2643481 -2.16817938 0.0301450 0.5637445

$logLik
[1] -199.8646
$McFaddenR2
[1] 0.026917
```

Uzyskany zestaw wyników (5 wyrazów wolnych (`intercept`) oraz 5 parametrów dla zmiennej specyficznej (`age`)) reprezentuje efekt wpływający na prawdopodobieństwo wyboru opcji (profilu) 1–5 w stosunku do profilu odniesienia oznaczonego numerem 6. Porównanie uzyskanych wartości wskazuje, że profil 1 jest wybierany głównie przez młodszych respondentów (w wieku do 25 lat).

W wyniku zastosowania funkcji MNLprob uzyskuje się oszacowania prawdopodobieństwa wyboru poszczególnych profilów przez respondentów (na poziomie indywidualnym oraz w przekroju wszystkich respondentów):

```
> z1<-as.data.frame(Z[,2])
> n<-names(Z)
> colnames(z1)<-n[2]
> MNLprob(Y3, X3, z1)

             p1        p2         p3         p4         p5        p6
[1,] 0.01650361 0.4341366 0.09909333 0.08423948 0.19993692 0.1660900
[2,] 0.03242261 0.4891803 0.06982926 0.07960448 0.13299180 0.1959715
[3,] 0.03242261 0.4891803 0.06982926 0.07960448 0.13299180 0.1959715
[4,] 0.06014667 0.5204827 0.04646493 0.07103201 0.08353173 0.2183420
[5,] 0.06014667 0.5204827 0.04646493 0.07103201 0.08353173 0.2183420
[6,] 0.06014667 0.5204827 0.04646493 0.07103201 0.08353173 0.2183420

> apply(Pmnl, 2, "mean")
        p1         p2         p3         p4         p5         p6
0.02205912 0.41911326 0.10293990 0.08088135 0.21323584 0.16177054
```

Należy zauważyć, że spośród profilów p1-p6 prawdopodobieństwo wyboru profilu 2 jest największe (cena – 10-20 zł, miejsce – restauracja, rodzaj – posiłek, godzina konsumpcji – popołudniowa). Prawdopodobieństwo wyboru profilu 1 jest natomiast najmniejsze (cena – powyżej 20 zł, miejsce – stołówka, rodzaj – napój, godzina konsumpcji – poranna). Podobne oszacowania uzyskuje się w wyniku zastosowania funkcji CLMprob (w odniesieniu do wszystkich 18 profilów):

```
> Pclm<-CLMprob(X,Y)
> print(Pclm)
$probave
         p1          p2          p3          p4          p5          p6
0.009243354 0.016961742 0.018691963 0.013300332 0.033819154 0.071208816
         p7          p8          p9         p10         p11         p12
0.108384655 0.056830334 0.051394611 0.059358559 0.108690721 0.071208816
        p13         p14         p15         p16         p17         p18
0.003265141 0.155407830 0.043486643 0.040435557 0.067102958 0.071208816

$probavesort
        p14         p11          p7          p6         p12         p18
0.155407830 0.108690721 0.108384655 0.071208816 0.071208816 0.071208816
        p17         p10          p8          p9         p15         p16
0.067102958 0.059358559 0.056830334 0.051394611 0.043486643 0.040435557
         p5          p3          p2          p4          p1         p13
0.033819154 0.018691963 0.016961742 0.013300332 0.009243354 0.003265141
```

Spośród profilów p1-p18 prawdopodobieństwo wyboru profilu 14 jest największe (cena – 10-20 zł, miejsce – restauracja, rodzaj – posiłek, godzina konsumpcji – popołudniowa). Najmniejszym prawdopodobieństwem wyboru charakteryzuje się 13 profil (cena – 10-20 zł, miejsce – stołówka, rodzaj – napój, godzina konsumpcji – poranna).

W przypadku modelu mieszanego zastosowanie funkcji MLMprob pozwala potwierdzić wyniki uzyskane za pomocą funkcji CLMprob:

```
> z1<-as.data.frame(Z[, 1])
> n<-names(Z)
> colnames(z1)<-n[1]
> x1<-as.data.frame(X3[, 1])
> n<-names(X3)
> colnames(x1)<-n[1]
> MLMprob(Y3, x1, z1)
             p1        p2         p3         p4        p5        p6
[1,] 0.030128257 0.4435556 0.11584168 0.09468429 0.1684209 0.1473693
[2,] 0.030128257 0.4435556 0.11584168 0.09468429 0.1684209 0.1473693
[3,] 0.001680195 0.3641731 0.07310887 0.04884232 0.3170734 0.1951221
[4,] 0.001680195 0.3641731 0.07310887 0.04884232 0.3170734 0.1951221
[5,] 0.030128257 0.4435556 0.11584168 0.09468429 0.1684209 0.1473693
[6,] 0.030128257 0.4435556 0.11584168 0.09468429 0.1684209 0.1473693
> apply(Pmlm,2,"mean")
        p1         p2         p3         p4         p5         p6
0.02155200 0.41962410 0.10295899 0.08086428 0.21323522 0.16176541
> Pmlm<-as.matrix(apply(Pmlm, 2, "mean"))
> Pmlm[order(Pmlm, decreasing=TRUE), ]
        p2         p5         p6         p3         p4         p1
0.41962410 0.21323522 0.16176541 0.10295899 0.08086428 0.02155200
```

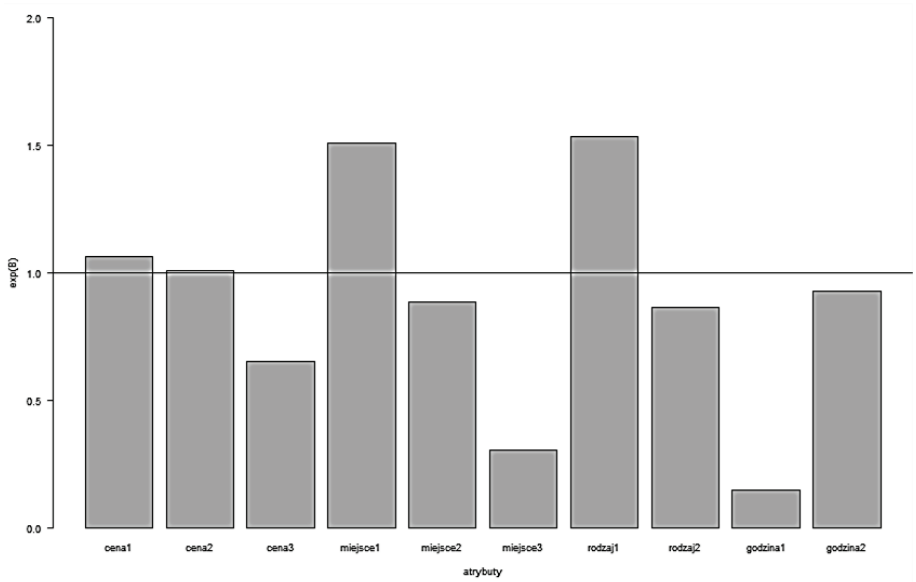
Profil 2 jest preferowany w największym stopniu (cena – 10-20 zł, miejsce – restauracja, rodzaj – posiłek, godzina konsumpcji – popołudniowa). Prawdopodobieństwo wyboru profilu 1 (z poranną godziną konsumpcji) jest najniższe.

Warto w tym miejscu zauważyć, że analizę preferencji respondentów w odniesieniu do wybranych opcji (atrybutów) umożliwia model warunkowy oraz funkcja CLMmodel:

```
> clm<-CLMmodel(X, Y)
> print(clm)
$estimate
               coef        se           t  Pr(>|t|) exp(coef)
cena1     0.0625235 0.2709762  0.23073429 0.8175212 1.0645195
cena2     0.0112104 0.2982141  0.03759178 0.9700131 1.0112735
cena3    -0.4218441 0.2426206 -1.73869861 0.0820878 0.6558363
miejsce1  0.4116794 0.1848646  2.22692392 0.0259524 1.5093505
miejsce2 -0.1219120 0.2678896 -0.45508299 0.6490495 0.8852263
miejsce3 -1.1846648 0.2621713 -4.51866699 0.0000062 0.3058487
rodzaj1   0.4288651 0.2170864  1.97555029 0.0482057 1.5355139
rodzaj2  -0.1440630 0.1637459 -0.87979607 0.3789698 0.8658332
godzina1 -1.9088592 0.3150913 -6.05811458 0.0000000 0.1482494
godzina2 -0.0713187 0.1431154 -0.49833002 0.6182515 0.9311651
```

```
$logLik
[1] -652.0696
$McFaddenR2
[1] 0.0696202
$LR
  df     Chisq   Pr(>Chisq)
[1,] 5 97.5886 1.702566e-19
```

Wartości `exp(coef)` oznaczają współczynniki ilorazu hazardu wybranych atrybutów. Analiza współczynników wskazuje, że respondenci preferują niską lub średnią cenę, restauracje, posiłki (tudzież popołudniową godzinę konsumpcji), co potwierdza wykres ilorazu hazardu uzyskany z zastosowaniem funkcji CLMgraph:



Rys. 1. Wpływ opcji wyboru (atrybutów) na wybór profilów

Źródło: opracowanie własne z wykorzystaniem programu R.

5. Podsumowanie

Mikroekonometryczne wielomianowe modele logitowe znajdują zastosowanie w analizie preferencji wyrażonych z wykorzystaniem podejścia opartego na wyborach dyskretnych. Różne typy modeli wielomianowych mogą być szacowane za pomocą metody największej wiarygodności z zastosowaniem iteracyjnego algorytmu optymalizacyjnego. Estymacja różnych typów wielomianowych modeli logitowych wymaga różnej struktury danych empirycznych. Pakiet `DiscreteChoice` zawiera funkcje przekształcające dane empiryczne do postaci wymaganych struktur oraz funkcje szacujące parametry różnych typów wielomianowych modeli logitowych wyborów dyskretnych.

Literatura

Bąk A. (2012), *Modele kategorii nieuporządkowanych w badaniach preferencji*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 17, Prace Naukowe UE we Wrocławiu nr 242, Wrocław, s. 86-95.

Bąk A., Bartłomowicz T., (2013a), *Discrete choice multinomial models – package* DiscreteChoice, [URL:] http://keii.ue.wroc.pl/DiscreteChoice.

Bąk A., Bartłomowicz T., (2013b), *Mikroekonometryczne modele wielomianowe i ich zastosowanie w analizie preferencji z wykorzystaniem programu R*, Prace Naukowe UE we Wrocławiu nr 278, Wrocław, s. 169-179.

Bierlaire M. (1997), *Discrete Choice Models*, URL: http://web.mit.edu/mbi/www/ michel.html, Cambridge, Massachusetts Institute of Technology.

Cameron A.C., Trivedi P.K. (2009), *Microeconometrics. Methods and Applications*, Cambridge University Press, New York.

Coombs C.H., Dawes R.M., Tversky A. (1977), *Wprowadzenie do psychologii matematycznej*, PWN, Warszawa.

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

Jackman S. (2007), *Models for Unordered Outcomes*, Political Science 150C/350C, [URL:] http://jackman.stanford.edu/classes/350C/07/unordered.pdf (12.03.2012).

Long J.S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, SAGE Publications, Thousand Oaks – London – New Delhi.

McFadden D. (1974), *Conditional Logit Analysis of Qualitative Choice Behavio*r, [w:] P. Zarembka (red.), *Frontiers in Econometrics*, Academic Press, New York – San Francisco – London.

Powers D.A., Xie Y. (2008), *Statistical Methods for Categorical Data Analysis*, 2nd ed. Emerald, Bingley.

R Development Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, URL: http://cran.r-project.org/.

So Y., Kuhfeld W.F. (1995), *Multinomial Logit Models*, [URL:] http://support.sas.com/ techsup/technote/ mr2010g.pdf (12.03.2012).

Winkelmann R., Boes S. (2006), *Analysis of Microdata*, Springer, Berlin.

MICROECONOMIC MULTINOMIAL LOGIT MODELS AND THEIR IMPLEMENTATION IN THE `DiscreteChoice` R PACKAGE

Summary: The main aim of the paper is to present the `DiscreteChoice` package developed for R program, which can be used to analyze consumers' preferences. The package contains an implementation of the discrete choice method, and some extensions are prepared to build fractional factorial design, to code the variables and to estimate of discrete choice model. Functions of `DiscreteChoice` package and their application in marketing research are presented in the article.

Keywords: microeconometrics**,** preferences, multinomial logit models of discrete choice method, R program.

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Justyna Brzezińska
Uniwersytet Ekonomiczny w Katowicach

WYKORZYSTANIE MODELI LOGARYTMICZNO-LINIOWYCH DO ANALIZY BEZROBOCIA W POLSCE W LATACH 2004-2012[1]

Streszczenie: Analiza logarytmiczno-liniowa pozwala na szczegółową ocenę zależności pomiędzy dowolną liczbą zmiennych niemetrycznych. W analizie tej wyróżnia się wiele rodzajów zależności, a jakość dopasowania modelu do danych ocenia się za pomocą współczynnika chi-kwadrat, ilorazu wiarygodności oraz kryteriów informacyjnych. W ciągu kilku lat bezrobocie w Polsce stało się jednym z poważniejszych problemów ekonomiczno-społecznych. Można zaobserwować duże jego zróżnicowanie pomiędzy różnymi regionami wśród osób z wyższym wykształceniem, a także względem płci. W niniejszym artykule modele logarytmiczno-liniowe wykorzystano do analizy struktury bezrobocia w Polsce w latach 2004-2012 na podstawie tablic zmiennych w czasie. Badanie przeprowadzono na podstawie danych pochodzących z Głównego Urzędu Statystycznego. Obliczenia przeprowadzone zostaną w programie R.

Słowa kluczowe: analiza logarytmiczno-liniowa, tablice kontyngencji, bezrobocie.

1. Wstęp

Analiza logarytmiczno-liniowa, należąca do wielowymiarowej analizy danych, jest metodą wykorzystywaną do badania zależności pomiędzy zmiennymi niemetrycznymi zapisanymi w wielowymiarowej tablicy kontyngencji. W metodzie tej nie rozróżnia się zmiennej zależnej oraz niezależnej, gdyż wszystkie zmienne traktowane są jako zmienne niezależne.

Modelowaniu poddane są liczebności w poszczególnych komórkach tablicy kontyngencji, które pełnią rolę zmiennej zależnej. Liczebności te traktowane są jako realizacja pewnej zmiennej losowej. Model logarytmiczno-liniowy zdefiniowany jest jako wyrażenie liczebności oczekiwanych (m_{hj}) w postaci funkcji para-

[1] Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2012/05/N/HS4/00174.

metrów reprezentujących charakterystyki zmiennych dyskretnych oraz zachodzących pomiędzy nimi relacji (interakcji). Budowanych jest wiele modeli, przy czym każdy z nich może zawierać różną liczbę parametrów wpływu oraz interakcji. Modele te budowane są według zasady hierarchiczności, następnie oceniane są za pomocą mierników oceny jakości dopasowania (chi-kwadrat, iloraz wiarygodności, kryteria informacyjne *AIC* oraz *BIC*, współczynnik determinacji). Celem analizy logarytmiczno-liniowej jest wybór modelu o jak najmniejszej liczbie parametrów, który jednocześnie jest modelem dobrze dopasowanym do danych. Dopasowanie modelu do danych rozumiane jest jako różnica pomiędzy wartościami empirycznymi a teoretycznymi. Im różnica między tymi wartościami jest mniejsza, tym dopasowanie modelu do danych jest lepsze.

Atutem analizy logarytmiczno-liniowej jest fakt, iż pozwala ona na analizę zmiennych w tablicach kontyngencji o dowolnym wymiarze, a także jako jedna z nielicznych metod analizy danych jakościowych, uwzględnia interakcje zachodzące między badanymi zmiennymi. W metodzie tej, w zależności od interakcji zawartych w równaniu modelu, możliwe jest wyróżnienie kilku rodzajów niezależności (np. model niezależności całkowitej, model niezależności częściowej, model niezależności łącznej oraz zależności homogenicznej).

Analiza logarytmiczno-liniowa jest metodą, którą wykorzystuje się do analizy danych przekrojowych, tj. takich, które dotyczą wybranego momentu czasowego. W niniejszym artykule metoda ta wykorzystana została do analizy bezrobocia w Polsce w latach 2004-2012, dzięki czemu możliwe jest zaobserwowanie zmiany struktury zachodzącej pomiędzy zmiennymi zależności. Celem artykułu jest opis modeli logarytmiczno-liniowych w analizie tablic kontyngencji oraz analiza struktury zależności zmiennych nominalnych dla wielu tablic kontyngencji zmiennych w czasie na przykładzie danych dotyczących bezrobocia w Polsce.

Dane pochodzą z Banku Danych Lokalnych Głównego Urzędu Statystycznego (www.stat.gov.pl). Niniejszy artykuł stanowi prezentację wykorzystania analizy logarytmiczno-liniowej w badaniu różnych tablic kontyngencji dla tych samych zmiennych zapisanych w różnych momentach czasu (w różnych tablicach kontyngencji). Analizie poddano kilka trójwymiarowych tablic kontyngencji (jedna tablica dla każdego roku), a następnie dla każdej z nich przeprowadzono pełną analizę logarytmiczno-liniową oraz wybrano model najlepszy. Badanie to pozwala na zaobserwowanie zależności występujących pomiędzy badanymi zmiennymi w różnych momentach czasowych.

2. Modele logarytmiczno-liniowe

Model pełny w przypadku trójwymiarowej tablicy kontyngencji $H \times J \times K$ ($h = 1,\ 2,...,H, j = 1, 2,...,J, k = 1, 2,...,K$) zdefiniowany jest następująco:

$$\ln\left(m_{hjk}\right) = \lambda + \lambda_h^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hj}^{XY} + \lambda_{hk}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{hjk}^{XYZ}, \qquad (1)$$

gdzie: λ to średnia arytmetyczna zlogarytmowanych liczebności cząstkowych z tablicy kontyngencji; λ_h^X, λ_j^Y, λ_k^Z odzwierciedlają wpływy poszczególnych zmiennych X, Y, Z; λ_{hj}^{XY}, λ_{hk}^{XZ}, λ_{jk}^{YZ} są interakcjami zmiennych XY, XZ, YZ; λ_{hjk}^{XYZ} jest interakcją rzędu drugiego zmiennych XYZ.

Dla modelu (1) spełniony jest warunek:

$$\sum_{h=1}^{H}\lambda_h^X=\sum_{j=1}^{J}\lambda_j^Y=\sum_{k=1}^{K}\lambda_k^Z=0,$$

$$\sum_{h=1}^{H}\lambda_{hj}^{XY}=\sum_{j=1}^{J}\lambda_{hj}^{XY}=\sum_{h=1}^{H}\lambda_{hk}^{XZ}=\sum_{k=1}^{K}\lambda_{hk}^{XZ}=\sum_{j=1}^{J}\lambda_{jk}^{YZ}=\sum_{k=1}^{K}\lambda_{jk}^{YZ}=0, \tag{2}$$

$$\sum_{h=1}^{H}\lambda_{hjk}^{XYZ}=\sum_{j=1}^{J}\lambda_{hjk}^{XYZ}=\sum_{k=1}^{K}\lambda_{hjk}^{XYZ}=0.$$

Model pełny ze względów praktycznych jest jednak modelem bezużytecznym, gdyż zawiera wszystkie możliwe interakcje. Celem badacza jest wybór modelu o postaci zredukowanej według zasady hierarchiczności w taki sposób, by wybrany model miał mniej parametrów niż model pełny.

Otrzymywane w modelu liczebności oczekiwane oraz podlegające interpretacji ilorazy szans silnie zależą od wyboru postaci modelu. Na ogół badacz nie posiada wiedzy *a priori* dotyczącej właściwego wyboru postaci modelu. Należy wtedy zbudować wiele modeli różniących się złożonością, a następnie dokonać oceny jakości ich dopasowania i wybrać model najlepszy. Pomiar ten odbywa się przez porównanie liczebności empirycznych n_{hjk} z liczebnościami oczekiwanymi m_{hjk}.

Wybór modelu odbywa się zazwyczaj dwuetapowo. W pierwszym etapie eliminowane są wszystkie modele, dla których iloraz wiarygodności wskazuje konieczność odrzucenia hipotezy głoszącej, że liczebności teoretyczne nie różnią się istotnie od liczebności empirycznych. Iloraz wiarygodności G^2 zdefiniowany jest jako [Christensen 1997; Agresti 2002; Zelterman 2006]:

$$G^2=2\sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{k=1}^{K}n_{hjk}\ln\left(\frac{n_{hjk}}{m_{hjk}}\right). \tag{3}$$

Współczynnik ten wykorzystuje się do porównywania modeli sąsiednich, budowanych wedle zasady hierarchiczności. Badana jest wówczas różnica ilorazów wiarygodności, która porównywana jest z liczbą odpowiadających jej stopni swobody. Pożądany jest przypadek braku podstaw do odrzucenia hipotezy zerowej o braku różnic między liczebnościami empirycznymi a teoretycznymi. W takich sytuacjach wzrasta ryzyko błędu II rodzaju i przy testowaniu tej hipotezy przyjmuje się poziom istotności z przedziału między 0,1 a 0,35 [Knoke, Burke 1980].

Kolejną statystyką służącą do porównania większej liczby modeli jest kryterium informacyjne Akaike *AIC* [Akaike 1973] (*Akaike Information Criteria*):

$$AIC = G^2 - 2df \,, \tag{4}$$

gdzie df oznacza liczbę stopni swobody.

Kryterium Beyesowskie BIC (*Bayesian Information Criteria*) [Schwarz 1978; Raftery 1986] jest drugim kryterium postaci:

$$BIC = G^2 - df \cdot \ln n \,, \tag{5}$$

gdzie n oznacza liczebność tablicy kontyngencji.

Minimalna wartość kryteriów informacyjnych pozwala na wybór najlepszego modelu logarytmiczno-liniowego. Ich istotą jest wskazanie nie modelu prawdziwego, lecz modelu, który zapewnia najwięcej informacji o badanym zjawisku. Mierniki te służą także do wyboru najlepszego modelu spośród kilku badanych, dzięki czemu badacz dysponuje obiektywnymi kryteriami wyboru modelu.

Kolejnym miernikiem pozwalającym na ocenę jakości dopasowania modelu do danych są współczynniki determinacji, zdefiniowane następująco [Christensen 1997]:

$$R^2 = \frac{G^2(M_0) - G^2(M)}{G^2(M_0)} \,, \tag{6}$$

lub w postaci skorygowanej jako:

$$\tilde{R}^2 = 1 - \frac{G^2(M)/(q-r)}{G^2(M_0)/(q-r_0)} = 1 - \frac{q-r_0}{q-r}\left(1-R^2\right), \tag{7}$$

gdzie: $q-r_0$ i $q-r$ to liczba stopni swobody odpowiadająca modelom M_0 i M, R^2 – współczynnik determinacji ocenianego modelu. Ze względu na uwzględnienie liczby stopni swobody każdego z badanych modeli, wartość skorygowanego współczynnika determinacji (7) jest nienormowana i może osiągać wartości ujemne.

Wybrany model jest najczęściej kompromisem między jego złożonością a jakością dopasowania do danych.

3. Wykorzystanie modeli logarytmiczno-liniowych w analizie bezrobocia w latach 2004-2012

Analiza logarytmiczno-liniowa w programie **R** dostępna jest w pakiecie `MASS` (funkcja `loglm`) oraz w pakiecie `stats` (funkcja `glm`). Zbiór danych pochodzący z Głównego Urzędu Statystycznego wykorzystany do zaprezentowania analizy logarytmiczno-liniowej dotyczy liczby osób bezrobotnych w Polsce w latach 2004-

-2012. Dla każdego roku zbudowano tablice o wymiarach $6 \times 5 \times 2$ dla trzech zmiennych nominalnych:

- *Region* [R] (Centralny, Południowy, Wschodni, Północno-zachodni, Południowo-zachodni, Północny),
- *Wykształcenie* [W] (Wyższe, Policealne i średnie zawodowe, Ogólnokształcące, Zasadnicze zawodowe, Gimnazjalne i poniżej),
- *Płeć* [P] (Kobieta, Mężczyzna).

W tabeli 1 zaprezentowano liczebności poszczególnych tablic wraz ze stopą bezrobocia w danym roku.

Tabela 1. Stopa bezrobocia oraz liczebność trójwymiarowych tablic kontyngencji w latach 2004-2012

Rok	2004	2005	2006	2007	2008	2009	2010	2011	2012
Stopa bezrobocia	19%	17,6%	14,8%	11,2%	9,5%	12,1%	12,4%	12,5%	13,4%
Liczebność w tys. osób	2999,601	2773	2309,410	1746,573	147,752	1892,680	1954,706	1982,676	2136,815

Źródło: Główny Urząd Statystyczny (www.stat.gov.pl).

W pierwszym etapie analizy zbudowano wszystkie modele zawierające trzy zmienne, tj. model pełny $[RWP]$, model zależności homogenicznej $[RW][RP][WP]$, modele zależności warunkowej $[RW][RP]$, $[RW][WP]$, $[RP][WP]$, modele niezależności częściowej $[RP][E]$, $[RW][P]$, $[WP][R]$ oraz model niezależności całkowitej $[R][W][P]$. W pierwszym etapie analizy okazało się, że wartość prawdopodobieństwa testowego p przekracza ustalony poziom $0{,}1$ w przypadku modeli: $[WP][R]$, $[RP][WP]$, $[RW][WP]$, $[RW][RP][WP]$ oraz $[RWP]$. Dla tych modeli różnice między wartościami empirycznymi i teoretycznymi są nieistotne o modele te należy uznać za akceptowalne. W drugim etapie analizy oceniono je za pomocą mierników 3-6. Oceny modeli dla danych z 2012 r. przedstawia tabela 2.

Istotny i interesujący w analizie dotyczącej bezrobocia w latach 2004-2012 jest fakt, iż wyniki uzyskane dla lat 2004-2012 są bardzo zbliżone. W pierwszym etapie analizy dla każdego roku na podstawie prawdopodobieństwa testowego p wskazywane są te same modele jako akceptowalne. Bardzo podobne wyniki uzyskuje się z podzielenia statystyki G^2 przez odpowiadającą modelowi liczbę stopni swobody *df*. Prawie identyczne okazują się także kryteria informacyjne (4 i 5) oraz współczynniki determinacji (6 i 7). Jako najlepszy wybrany zostaje model, dla którego kryteria informacyjne osiągają wartość najmniejszą. Dla każdego roku jest to model niezależności częściowej $[WP][R]$, który można zapisać w postaci równania:

$$\ln\left(m_{hjk}\right) = \lambda + \lambda_h^R + \lambda_j^W + \lambda_k^P + \lambda_{jk}^{WP}. \tag{8}$$

Tabela 2. Oceny modeli z trzema zmiennymi dla trójwymiarowej tablicy kontyngencji z 2012 r.

Model	*df*	G^2	*p*	R^2	$\tilde{R}^2$	*AIC*	*BIC*
$[P][R][W]$	49	123,213	0,000	0,000	0,0000	25,213	–252,473
$[WP][R]$	**45**	**29,684**	**0,962**	**0,759**	**0,7377**	**–60,316**	**–315,334**
$[RW][P]$	29	102,012	0,000	0,172	–0,3989	44,012	–120,333
$[RP][W]$	44	118,756	0,000	0,036	–0,0734	30,756	–218,595
$[RP][WP]$	**40**	**25,227**	**0,967**	**0,795**	**0,7492**	**–54,773**	**–281,456**
$[RW][WP]$	**25**	**8,483**	**0,999**	**0,931**	**0,8651**	**–41,517**	**–183,194**
$[RW][RP]$	24	97,555	0,000	0,208	–0,6165	49,555	–86,455
$[PR][PW][RW]$	**20**	**0,892**	**1,000**	**0,993**	**0,9823**	**–39,108**	**–152,450**
$[PRW]$	**0**	**0,000**	**1,000**	**1,000**	**1,0000**	**0,000**	**0,000**

Źródło: opracowanie własne w programie **R**.

Istotny jest również fakt, że współczynniki korelacji między wartościami empirycznymi a teoretycznymi dla modelu niezależności częściowej $[WP][R]$ w poszczególnych latach, które również świadczą o jakości dopasowania modelu do danych (im mniejsze odchylenia, tym lepsze dopasowanie modelu), osiągają zbliżone wartości. Dla roku 2012 współczynnik ten wynosi 0,968, co świadczy o niewielkich odchyleniach między wartościami empirycznymi a teoretycznymi wyznaczonymi dla danego modelu.

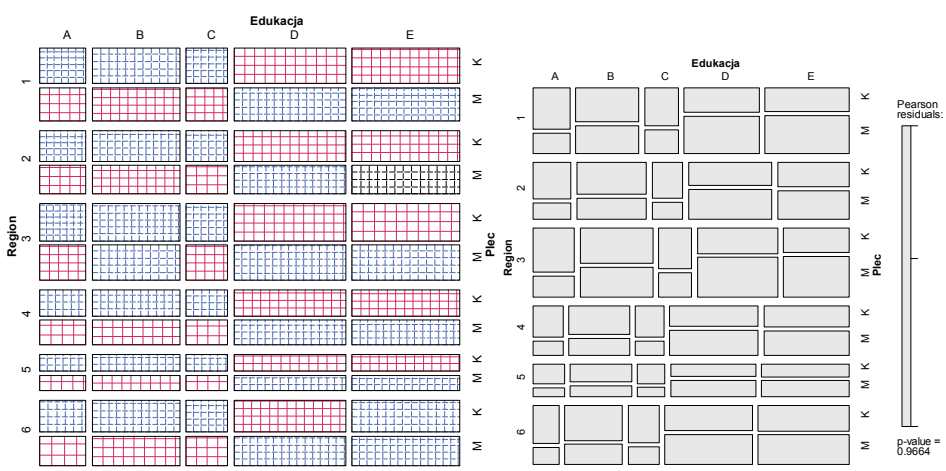
Uzyskane wyniki świadczą o silnej regule i zależności występującej pomiędzy zmiennymi w sposób określony w modelu. Po wyznaczeniu parametrów modelu za pomocą funkcji `param` także widoczna jest pewna prawidłowość i podobieństwo pomiędzy wynikami uzyskanymi dla poszczególnych lat, zarówno w znakach, jak i wartościach parametrów. Znaki parametrów dla interakcji $[WP]$ dla poziomu wykształcenia: wyższe, policealne i średnie zawodowe, ogólnokształcące są dodatnie, a dla poziomu zasadniczego zawodowe oraz gimnazjalnego i poniżej parametry te są ujemne, zarówno w grupie mężczyzn, jak i kobiet. Oznacza to, że w komórkach dla wykształcenia o wyższych kategoriach, dla których parametry są dodatnie, liczebność tej komórki jest większa względem liczebności średniej. Dla niższych kategorii, dla których parametry interakcji mają znaki ujemne, liczebności te są mniejsze niż liczebność przeciętna.

Do oceny jakości dopasowania modelu do danych, szczególnie w przypadku znacznej liczby zmiennych, można posłużyć się wykresem mozaikowym [Friendly 1994, 1995, 2000]. Wykresy mozaikowe składają się z prostokątnych płytek (*tile*, *bin*, *box*, *rectangle*), których pole jest proporcjonalne do liczebności empirycznej n_{hj}, szerokość proporcjonalna jest do liczebności brzegowej $n_{h\bullet}$, a wysokość do proporcji $\frac{n_{hj}}{n_{h\bullet}}$. Budowa tego wykresu oparta jest na standaryzowanych resztach Pearsona, zdefiniowanych jako:

$$d_{hj} = \frac{n_{hj} - \hat{m}_{hj}}{\sqrt{\hat{m}_{hj}}}.$$

Jeśli reszta jest dodatnia, dany prostokąt oznaczony jest kolorem niebieskim, jeśli ujemna – kolorem czerwonym. Przedziały, w których znajdują się reszty, oznaczone są coraz ciemniejszym kolorem w miarę wzrostu wartości d_{hj} ($|d_{hj}| > 0, 2, 4, \ldots$). W programie **R** wykres mozaikowy uzyskuje się dzięki funkcji `mosaic()`.

Kolejnym wykresem przeznaczonym do wizualizacji danych w wielowymiarowych tablicach kontyngencji jest wykres sitkowy (*sieve plot*), zwany także wykresem parkietowym (*parquet diagram*). Na wykresie tym powierzchnia każdego prostokąta jest proporcjonalna do liczebności oczekiwanych m_{hj}, przy czym liczebność empiryczna odpowiada liczbie kwadratów w danym prostokącie [Friendly 2000]. Szerokość każdego prostokąta jest proporcjonalna do liczebności brzegowych kolumn $n_{\bullet j}$, a jego wysokość do liczebności brzegowych wierszy $n_{h\bullet}$. Odchylenia liczebności empirycznych od teoretycznych ($n_{hj} - m_{hj}$) przedstawione są w postaci kolorowych linii. Jeśli różnica ta jest ujemna, wówczas linia tworząca kwadraty w odpowiednim prostokącie jest czerwoną linią ciągłą. Jeśli różnica ta jest dodatnia, wówczas linia w danym prostokącie jest przerywana niebieska. Niezależność pomiędzy zmiennymi występuje wówczas, gdy zagęszczenie i struktura kwadratów jest jednorodna. W przypadku niejednorodności można przypuszczać, że zmienne są zależne [Friendly 2002]. W programie **R** wykres sitkowy otrzymywany jest dzięki funkcji `sieve()`.



Rys. 1. Wykres sitkowy: (a) i mozaikowy (b) dla trójwymiarowej tablicy kontyngencji.

Źródło: opracowanie własne w programie **R**.

Niewielkie odchylenia liczebności empirycznych od teoretycznych na wykresie mozaikowym (rys. 1a) świadczą o dobrym dopasowaniu modelu do danych. Strukturę poszczególnych komórek trójwymiarowej tablicy kontyngencji przedstawia wykres sitkowy (rys. 1b).

Interpretacja parametrów modelu jest trudniejsza w przypadku większej liczby zmiennych. Wówczas interpretuje się jedynie końcowe równanie modelu, które poprzez uwzględnione parametry i interakcje określa rodzaj zachodzącej pomiędzy zmiennymi zależności. Modele te jednak opisują w szczegółowy sposób charakter powiązań pomiędzy zmiennymi w tablicy kontyngencji, zarówno w przypadku zmiennych nominalnych, jak i porządkowych.

4. Zakończenie

Analiza logarytmiczno-liniowa jest metodą pozwalającą na badanie zależności zachodzących pomiędzy zmiennymi zapisanymi w wielowymiarowych tablicach kontyngencji. Metoda ta wykorzystywana jest zazwyczaj dla danych przekrojowych, dotyczących wielu zmiennych w tablicy kontyngencji badanej w danym momencie czasu. Zaletą tej metody jest fakt, iż może być ona stosowana dla tablic kontyngencji o dowolnych wymiarach, a także dla zmiennych nominalnych oraz porządkowych.

W niniejszym artykule zaprezentowano jej wykorzystanie do analizy bezrobocia w latach 2004-2012. Analizie poddano te same zmienne (*Region*, *Wykształcenie*, *Płeć*); dla każdego roku zbudowano trójwymiarową tablicę kontyngencji i przeprowadzono analizę, wybierając model najlepszy. Wybrany model dla każdego roku ma to samo równanie, co wskazuje, że istotna jest interakcja między zmienną *Wykształcenie* oraz *Płeć*. Współczynniki oceny jakości modelu dla każdego roku także mają zbliżone wartości. Analiza parametrów pozwala na wyciągnięcie interesujących wniosków. Znaki parametrów w przypadku interakcji $[WP]$ dla wykształcenia wyższego, policealnego i średniego zawodowego oraz ogólnokształcącego mają znaki dodatnie, a dla zasadniczego zawodowego oraz gimnazjalnego i poniżej parametry te są ujemne, zarówno w grupie mężczyzn, jak i kobiet. Oznacza to, że w widoczna jest taka sama struktura zależności pomiędzy badanymi zmiennymi, co potwierdzone jest wyborem te samej postaci modelu w każdym roku.

Analiza logarytmiczno-liniowa może także zostać wykorzystana w analizie zmiennych porządkowych oraz analizie klas ukrytych. Jej istotna przewaga nad innymi metodami analizy danych jakościowych polega na tym, iż możliwa jest wizualizacja wyników, znacznie ułatwiająca ich interpretację.

Literatura

Agresti A. (2002), *Categorical Data Analysis*, John Wiley & Sons, Hoboken, New Jersey.
Akaike H. (1973), *Information theory and an extension of the maximum likelihood principle*, Proceedings of the 2nd International Symposium on Information, Petrow B.N., Czaki F., Akademiai Kiado, Budapest.
Christensen R. (1997), *Log-linear Models and Logistic Regression*, Springer-Verlag, New York.
Friendly M. (1994), *Mosaic displays for multi-way contingency tables*, „Journals of the American Statistical Association" 49, s. 153-160.
Friendly M. (1995), *Conceptual and visual models for categorical data*, „The Amercian Statistician" 49, s. 153-160.
Friendly M. (2000), *Visualizing Categorical Data*, SAS Institute.
Knoke D., Burke P.J. (1980), *Log-linear Models*, Sage University Paper Series on Quantitative Applications in the Social Science, series no. 07-020, Beverly Hills and London Sage.
Raftery A.E. (1986), *Choosing models for cross-classification*, „American Sociological Review" 51, 1, s. 145-146.
Schwarz G. (1978), *Estimating the dimensions of a model*, „Annals of Statistics" 6, s. 461-464.
Zelterman D. (2006), *Models for Discrete Data*, Oxford University Press.

THE ANALYSIS OF UNEMPLOYMENT DATA IN POLAND IN 2004-2012 WITH APPLICATION OF LOG-LINEAR MODELS

Summary: Log-linear analysis allows to analyze the relationship between two or more categorical (e.g. nominal or ordinal) variables. There are several types of association. For testing the goodness of fit the Pearson chi-square statistic, likelihood ratio and information criteria are used. With the rising unemployment rate in recent years, unemployment is one of the most important socio-economic and social problems in Poland. The comparative log-linear analysis of unemployment will be presented on the data from the Central Statistical Office. Log-linear models are available in **R** software.

Keywords: log-linear analysis, contingency table, unemployment.