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

Summaries

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.

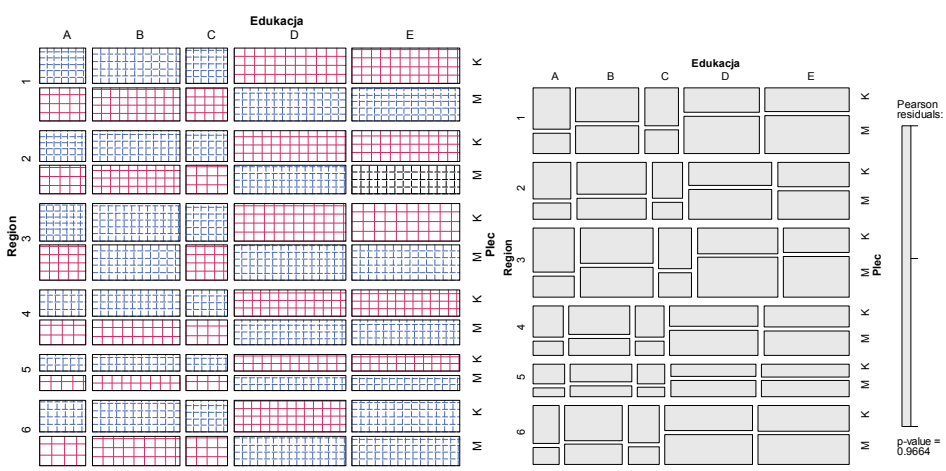
Uzyskane wyniki świadczą o silnej regule i zależności występującej pomiędzy zmiennymi w sposób określony w modelu. Po wyznaczeniu parametrów modelu za pomocą funkcji `param` także widoczna jest pewna prawidłowość i podobieństwo pomiędzy wynikami uzyskanymi dla poszczególnych lat, zarówno w znakach, jak 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.