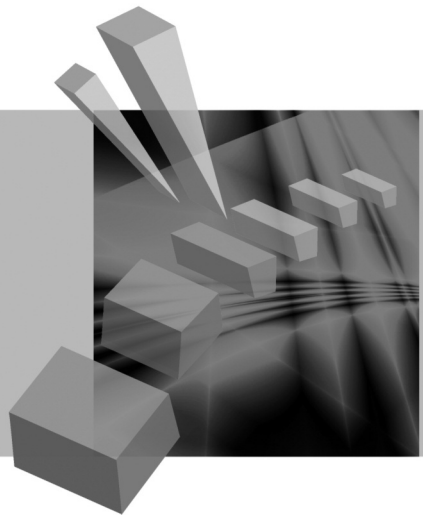
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w 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

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Justyna Brzezińska
Uniwersytet Ekonomiczny w Katowicach

ANALIZA LOGARYTMICZNO-LINIOWA W BADANIU PRZYCZYN UMIERALNOŚCI W KRAJACH UE

Streszczenie: Analiza logarytmiczno-liniowa jest metodą badania niezależności zmiennych niemetrycznych, która pozwala na analizę dowolnej liczby zmiennych przy jednoczesnym uwzględnieniu interakcji zachodzących pomiędzy nimi. Modele, które budowane są hierarchicznie, oceniane są za pomocą statystyki chi-kwadrat, ilorazu wiarygodności oraz kryterium informacyjnego AIC [Akaike 1973] oraz BIC [Raftery 1986]. W niniejszym artykule zaprezentowane zostanie także kryterium Aitkina. Spośród wszystkich modeli wybrany zostaje model o najmniejszej złożoności, który jest dobrze dopasowany do danych. W programie **R** analiza logarytmiczno-liniowa dostępna jest dzięki funkcji `loglm` w pakiecie `MASS`. W niniejszym artykule analiza ta zaprezentowana zostanie do analizy przyczyn umieralności w krajach UE.

Słowa kluczowe: analiza logarytmiczno-liniowa, modele logarytmiczno-liniowe, analiza niezależności zmiennych niemetrycznych.

1. Wstęp

Analiza logarytmiczno-liniowa jest metodą statystyki wielowymiarowej wykorzystywaną do modelowania liczebności w poszczególnych komórkach tablicy kontyngencji oraz do zaprezentowania struktury powiązań pomiędzy zmiennymi dyskretnymi o rozkładzie Poissona. W analizie tej model logarytmiczno-liniowy zdefiniowany jest jako wyrażenie liczebności oczekiwanych (m_{hj}) w postaci funkcji parametrów reprezentujących charakterystyki zmiennych dyskretnych oraz zachodzących pomiędzy nimi relacji (interakcji). Modele budowane są według zasady hierarchiczności, następnie każdy z nich jest oceniany za pomocą mierników oceny jakości dopasowania (iloraz wiarygodności, kryteria informacyjne AIC oraz BIC, metoda Aitkina). 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, przy czym im różnica jest mniejsza, tym dopasowanie modelu do danych lepsze. Modelem doskonale dopasowanym do danych (tj. którego liczebności empiryczne są równe liczebnościom teoretycznym) jest

model pełny uwzględniający wszystkie zmienne i możliwe między nimi interakcje. W praktyce jednak model ten jest bezużyteczny, natomiast modelem pożądanym jest model zredukowany, tj. zawierający jak najmniejszą liczbę zmiennych.

Atutem analizy logarytmiczno-liniowej jest fakt, iż pozwala ona na uwzględnienie dowolnej liczny zmiennych niemetrycznych, a także interakcji między nimi, dzięki czemu możliwa jest analiza różnych rodzajów niezależności (niezależność całkowita, warunkowa, częściowa, homogeniczna). Staje się to niezwykle użyteczne wówczas, gdy liczba wymiarów tablicy kontyngencji jest duża, a klasyczne metody analizy nie potrafią poradzić sobie z dużą liczbą wymiarów.

Celem niniejszej pracy jest prezentacja analizy logarytmiczno-liniowej, która wykorzystana została w badaniu demograficznym do analizy niezależności pomiędzy przyczynami umieralności w krajach Unii Europejskiej. Przykład ten pokazuje, iż jest to tylko niewielki obszar zastosowania modeli, które są praktycznym i skutecznym narzędziem analizy zależności zmiennych jakościowych w dowolnym obszarze zainteresowań. Metoda ta jest szczególnie użyteczna, gdy analizie poddanych jest wiele zmiennych mierzonych na skali nominalnej, dzięki czemu możliwe jest wyodrębnienie interakcji wyższych rzędów wpływających na rodzaj zależności. Weryfikację wyników umożliwiają inne metody statystycznej analizy wielowymiarowej, takie jak analiza korespondencji oraz metody taksonomiczne.

2. Analiza logarytmiczno-liniowa

Niech X i Y będą zmiennymi dyskretnymi o liczebnościach empirycznych n_{hj} w h-tym wierszu i j-tej kolumnie $(h = 1, \ldots H,\ j = 1, \ldots, J)$. Model addytywny dla dwóch zmiennych, uwzględniający wszystkie możliwe efekty pojedyncze oraz interakcje pomiędzy zmiennymi jest modelem o równaniu:

$$\ln\left(m_{hj}\right) = \lambda + \lambda_h^X + \lambda_j^Y + \lambda_{hj}^{XY}, \tag{1}$$

gdzie: λ – średnia arytmetyczna zlogarytmowanych liczebności cząstkowych z tablicy kontyngencji,

$\lambda_h^X,\ \lambda_j^Y$ – efekty zmiennej X i Y,

λ_{hj}^{XY} – efekt interakcji pomiędzy zmiennymi XY.

Model pełny (1) zawiera wszystkie możliwe interakcje pomiędzy zmiennymi i jest modelem, który w pełni odtwarza liczebności zaobserwowane, tzn. $n_{hj} = m_{hj}$. Ze względu na jego złożoność jest to model bezużyteczny, stąd konieczność zbudowania i oceny modeli prostszych. Każdemu modelowi odpowiada liczba stopni swobody *df*, zależna od liczby parametrów do oszacowania, definiowana jako [Everitt 1977; Agresti 2002]:

df = liczba komórek w tablicy kontyngencji – liczba wolnych parametrów. (2)

Liczba wolnych parametrów wynika z nałożonych na równanie modelu ograniczeń:

$$\sum_{h=1}^{H}\lambda_h^X = \sum_{j=1}^{J}\lambda_j^Y = \sum_{h=1}^{H}\lambda_{hj}^{XY} = \sum_{j=1}^{J}\lambda_{hj}^{XY} = 0. \tag{3}$$

Jeśli dana jest dwuwymiarowa tablica kontyngencji $H \times J$, wówczas istnieje jeden wolny parametr λ, $H-1$ wolnych parametrów λ_h^X, $J-1$ wolnych parametrów λ_j^Y oraz $(H-1)(J-1)$ wolnych parametrów λ_{hj}^{XY}.

Liczebności teoretyczne modeli wyznaczane są za pomocą algorytmu iteracyjno-proporcjonalnego (*iterative proportional fiting*) [Deming, Stephan 1940], który pozwala na wyznaczenie parametrów modelu będących estymatorami największej wiarygodności (MLE) [Bishop, Fienberg, Holland 1975; Christensen 1997; Mair 2006].

Hierarchiczne modele logarytmiczno-liniowe wybierane są według jednej z procedur krokowych (*stepwise procedures*), tj. selekcji w przód (*forward selection*) lub eliminacji wstecznej (*backward elimination*). Oznacza to, że analizę rozpoczyna się od modelu wyjściowego (model zerowy, model pełny), dodając lub usuwając kolejne elementy według zasady hierarchiczności.

Podstawową statystyką oceny jakości dopasowania modelu logarytmiczno-liniowego do danych jest statystyka chi-kwadrat postaci:

$$\chi^2 = \sum_{h=1}^{H}\sum_{j=1}^{J}\frac{\left(n_{hj} - m_{hj}\right)^2}{m_{hj}} \tag{4}$$

oraz iloraz wiarygodności:

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J} n_{hj}\ln\left(\frac{n_{hj}}{m_{hj}}\right). \tag{5}$$

Innymi kryteriami pozwalającymi na ocenę jakości modelu są kryteria informacyjne AIC [Akaike 1973] oraz BIC [Raftery 1986] zdefiniowane jako:

$$AIC = G^2 - 2df\,, \tag{6}$$

$$BIC = G^2 - df \cdot \ln n\,. \tag{7}$$

Mierniki te są wyznaczane dla każdego modelu, natomiast ich najmniejsza wartość wskazuje na model najlepiej dopasowany do danych.

Wybór modelu końcowego odbywa się w dwóch etapach. W etapie pierwszym testowana jest hipoteza o równości liczebności empirycznych z oczekiwanymi, tj. $H_0 : n_{hj} = m_{hj}$. Modelem pożądanym jest model, którego różnice między tymi liczebnościami są jak najmniejsze. W etapie drugim natomiast wybrane poprzednio modele

są porównywane względem siebie i badana jest różnica między ilorazami wiarygodności (dewiancjami). Hipoteza zerowa głosi wówczas, iż różnica ta wynosi zero, tj. $H_0 : \Delta G^2 = 0,$ przy czym liczba stopni swobody dla tego testu jest różnicą między liczbą stopni swobody porównywanych modeli, tj. $\Delta df = df_2 - df_2$.

W zaprezentowanej analizie logarytmiczno-liniowej sposób definiowania hipotez H_0 jest odmienny od standardowego i dąży się do sytuacji, by decyzją był brak podstaw do ich odrzucenia. Wzrasta wówczas ryzyko popełnienia błędu II rodzaju i w tym celu poziom istotności ustalany jest na poziomie wyższym, tj. $[0{,}1; 0{,}35]$ [Bishop 1975].

3. Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE

W programie **R** analiza logarytmiczno-liniowa dostępna jest w pakiecie `MASS` (funkcja `loglm`) oraz w pakiecie `stats` (funkcja `glm`). Zbiór danych wykorzystany do zaprezentowania algorytmu analizy logarytmiczno-liniowej pochodzi z Europejskiego Urzędu Statystycznego i dotyczy przyczyn umieralności w Unii Europejskiej w 2009 r.

Analizie poddano dwie zmienne o charakterze dyskretnym: „`Kraj" (C)` (32 kraje europejskie) oraz „`Przyczyna zgonów" (D) ("Cancer", "Heart", "Nervous", "Pneumonia", "Liver", "Diabetes", "Accidents", "Suicide", "Alcohol", "Homicide", "AIDS", "Drug")`. Zbudowano dwa modele: model pełny (*saturated model*) i model niezależności (*independence model*), a następnie oceniono je za pomocą ilorazu wiarygodności G^2 z odpowiadającą mu liczbą stopni swobody df, poziomem `p-value` oraz kryterium informacyjnym *AIC*.

Tabela 1. Kryteria oceny jakości modelu

s	Model	G^2	df_s	*p-value*	*AIC*
2	[CD]	0	0	1	3 935
1	[C][D]	152 400	341	0	155 600

Źródło: opracowanie własne w **R**.

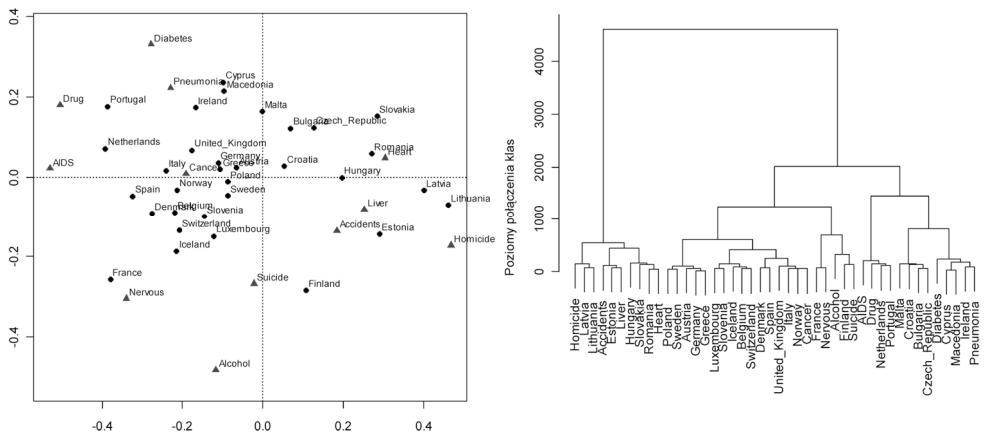
Z informacji zawartych w tab. 1 widać, iż spośród zbudowanych modeli tylko model pełny [CD] jest modelem dopasowanym do danych, ponieważ jedynie dla tego modelu wartość prawdopodobieństwa testowego przekracza 0,15. Jednocześnie wartość kryterium informacyjnego *AIC* dla tego modelu jest najmniejsza, można zatem uznać, że model jest dopasowany do danych. Z własności modelu pełnego wiadomo również, że liczebności empiryczne w idealny sposób odzwierciedlają teoretyczne i dopasowanie takiego modelu jest doskonałe, jednak z praktycznego punk-

tu widzenia model ten zawiera wszystkie parametry wpływu i interakcje, zatem jest modelem bardzo złożonym. Ze względu na brak innego modelu akceptowalnego model ten zostanie wybrany jako najlepszy. W równaniu modelu pełnego występują parametry wpływu zmiennej wierszowej, kolumnowej oraz interakcja pomiędzy nimi. Oznacza to, że na zależność pomiędzy krajem a przyczyną śmierci wpływają zarówno same zmienne, jak i interakcja pomiędzy nimi. Model pełny w tym przypadku przedstawić można jako równanie postaci:

$$\ln(m_{hj}) = \lambda + \lambda_h^C + \lambda_j^D + \lambda_{hj}^{CD}. \tag{8}$$

Parametry modelu estymowane są metodą największej wiarygodności z wykorzystaniem algorytmu dopasowania iteracyjno-proporcjonalnego (IPF) i mówią o kierunku i charakterze zależności pomiędzy zmiennymi. Większe wartości parametrów z interakcją zaobserwowano dla krajów i przyczyn umieralności, które są położone dalej od siebie, a mniejsze dla tych, które leżą bliżej.

W przypadku analizy zmiennych dyskretnych dzięki analizie korespondencji możliwa jest graficzna prezentacja wyników w postaci mapy percepcji. Współwystępowanie pomiędzy kategoriami badanych zmiennych pozwoli zaobserwować charakter i rodzaj zależności. Do weryfikacji otrzymanych wyników, tj. modelu pełnego w analizie logarytmiczno-liniowej w postaci graficznej, wykorzystać można zarówno mapę percepcji, która jest rezultatem analizy korespondencji, jak i metody taksonomiczne, które pozwalają na graficzną prezentację wyników w postaci dendrogramu.



Rys. 1. Mapa percepcji oraz dendrogram ilustrujący wynik analizy taksonomicznej metodą Warda

Źródło: opracowanie własne w **R**.

Aby poprawnie określić liczbę klas, zbadano wartość indeksu Silhouette dla trzech, czterech, pięciu i sześciu klas, a następnie dokonano oceny jakości klasyfikacji. Maksymalna wartość indeksu Silhouette uzyskana jest dla trzech klas

($S(u) = 0{,}45$), co oznacza, że podział obiektów na trzy klasy jest najlepszy. Po zastosowaniu funkcji `cutree` otrzymano przynależność każdego obiektu do danej klasy i pierwszą grupę krajów stanowią takie kraje, jak: Belgia, Dania, Niemcy, Grecja, Hiszpania, Francja, Włochy, Luksemburg, Austria, Polska, Słowenia, Finlandia, Szwajcaria, Wielka Brytania, Islandia, Szwecja i Norwegia, w których przyczyną umieralności są choroby nerwowe, nowotworowe, samobójstwa oraz alkoholowe. Grupę drugą tworzą: Bułgaria, Czechy, Irlandia, Cypr, Malta, Holandia, Portugalia, Chorwacja, Macedonia, w których dominującą przyczyną umieralności jest cukrzyca, AIDS oraz narkotyki. Pozostałe kraje należą do klasy trzeciej, w której głównymi przyczynami umieralności są choroby serca, wypadki i bezdomność. Podział ten potwierdza także metoda taksonomiczna, której graficznym wynikiem jest dendrogram.

4. Podsumowanie

Analiza logarytmiczno-liniowa pozwala na ocenę zależności pomiędzy dowolną liczbą zmiennych dyskretnych, uwzględniając interakcje pomiędzy zmiennymi w dowolnych konfiguracjach. Metoda ta pozwala także na wyodrębnienie zmiennych wpływających na charakter zależności. Liczne statystyki oceny jakości modelu oraz kryteria informacyjne pozwalają na wybór modelu o jak najprostszej postaci, który jest zarazem w wystarczający sposób dopasowany do danych. Taki model cechuje się tym, iż jego liczebności empiryczne są zbliżone do teoretycznych.

Modele logarytmiczno-liniowe są techniką analityczną, która z powodzeniem wykorzystywana jest w badaniach demograficznych, marketingowych, społecznych, psychologicznych, socjologicznych oraz medycznych. W niniejszym artykule metoda ta została wykorzystana do badania o charakterze demograficznym, którego celem było zbadanie zależności pomiędzy przyczynami umieralności w 27 krajach Unii Europejskiej i określenie charakteru tej zależności. Dla dwuwymiarowej tablicy kontyngencji zbudowane zostały modele: pełny oraz model niezależności. Kryteria oceny modelu pozwoliły na wybór modelu dopasowanego do danych, którym okazał się model pełny zawierający parametry wpływu pojedynczych zmiennych, a także interakcję pomiędzy nimi. Do graficznej prezentacji wyników analizy logarytmiczno-liniowej wykorzystano analizę korespondencji, której rezultatem jest mapa oraz dendrogram powstały w wyniku zastosowania analizy taksonomicznej metodą Warda. Z przeprowadzonej analizy można zaobserwować, iż widoczne są trzy grupy ze względu na kilka przyczyn umieralności, które są identyczne zarówno na mapie percepcji, jak i na dendrogramie. Pierwszą grupę stanowią kraje, w których dominują choroby nerwowe, nowotworowe, samobójstwa oraz alkoholowe. W grupie drugiej dominuje cukrzyca, AIDS oraz narkotyki. Klasę trzecią stanowią choroby serca, wypadki i bezdomność.

Analiza logarytmiczno-liniowa, którą zaprezentowano w niniejszym artykule, pozwoliła na analizę zależności pomiędzy zmiennymi: kraj i przyczyna umieralności.

Model pełny, który za pomocą metody eliminacji wstecznej został wybrany jako najlepiej dopasowany do danych, wskazuje, iż istotną rolę w modelu odgrywa interakcja pomiędzy zmiennymi. Metoda ta jest szczególnie użyteczna w sytuacjach, gdy badaniu poddanych jest wiele zmiennych, dzięki czemu można zaobserwować wpływ poszczególnych zmiennych w różnych konfiguracjach (interakcje wyższych rzędów) na zależność pomiędzy badanymi zmiennymi.

Literatura

Agresti A., *Categorical Data Analysis*, 2nd edition, Wiley, 2002.

Aitkin M., *The analysis of unbalced cross-classification*s, „Journal of the Royal Statistical Society", Series B, 1978, no 141.

Aitkin M., *A simultaneous test procedure for contingency tables*, „Applied Statistics" 1979, no 28.

Akaike H., *Information theory and an extension of the maximum likelihood principle*, [w:] Proceedings of the 2nd International Symposium on Information, B.N. Petrow, F. Czaki, Akademiai Kiado, Budapest 1973.

Bishop Y.M.M., Fienberg E.F., Holland P.W., *Discrete multivariate analysis*, MIT Press, Cambridge, Massachusetts 1975.

Christensen R., *Log-linear Models and Logistic Regression*, Springer-Verlag, New York 1997.

Deming W., Stephan F., *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*, „Annals of Mathematical Statistics" 1940, no 11.

Everitt B., *The analysis of contingency tables*, Chapman and Hall, London 1977.

Mair P., *Interpreting Standard and Nonstandard Log-linear Models*, Waxmann Verlag, 2006.

Raftery A.E., *Choosing models for cross-classification*, „Amer. Sociol. Rev." 1986, no 51.

LOG-LINEAR ANALYSIS IN THE STUDY OF MORTALITY IN EU

Summary: Log-linear models are a widely used tool for modeling qualitative data in contingency table. Hierarchical log-linear models include lower order terms implied by any higher order ones. The fit of log-linear model can be assessed with the Pearson or likelihood-ratio chi-square, as well as information criteria: *AIC* [Akaike 1973]*, BIC* [Raftery 1986] and Aitkin`s method. Identifying the simplest model with the fewest parameters is the main goal of the analysis. Log-linear analysis is available in **R** software with the use of `loglm` function in library `MASS`. The empirical use of log-linear analysis is based on European Union countries mortality dataset.

Keywords: log-linear analysis, log-linear models, independence of categorical variable.