Justyna Brzezińska
Uniwersytet Ekonomiczny w Katowicach

MODELOWE METODY ANALIZY DANYCH WIELOWYMIAROWYCH TABLIC KONTYNGENCJI W BADANIACH OPINII PUBLICZNEJ

Wprowadzenie

Tablice, które stanowią podstawową formę zapisu zmiennych niemetrycznych, znane były w historii już ponad 2000 lat przed naszą erą. Babilończycy wykorzystywali je do przedstawienia zależności w pewnym systemie liczbowym. Matematycy chińscy używali tablic liczbowych w obliczeniach, które niewiele różniły się od znanej dziś tabliczki mnożenia. Część etymologów uważa za źródłosłów terminu „tablica" słowo „stół" (*table*), który w czasach średniowiecznych wykorzystywany był do układania na nim należności podatkowych od obywateli danego państwa[1].

W XVIII wieku, kiedy rozwinęła się statystyka państwowa, tablice były wykorzystywane do opisu zasobów państwa. Kluczowym okresem z punktu widzenia statystyki jako nauki jest przełom XIX i XX wieku, kiedy zaczęto analizować formalne własności tablic. Pionierem w tym zakresie był Karl Pearson, który wprowadził po raz pierwszy pojęcie korelacji należące do najbardziej fundamentalnych narzędzi opisu i interpretacji zjawisk w wielu dyscyplinach naukowych, a także zdefiniował pojęcie tablicy kontyngencji. Pearson, zainspirowany problemem losowości wyników ruletki Monte Carlo, zdefiniował także jako pierwszy współczynnik chi-kwadrat, dzięki czemu analiza zmiennych niemetrycznych wkroczyła w epokę rozwoju i zainteresowania naukowego, która nadal trwa. W latach 1900-1912, równolegle do Pearsona, prace nad analizą tablic kontyngencji prowadził także Yule, który zdefiniował miarę zależności

[1] Z. Sawiński: Zastosowania tablic w badaniach zjawisk społecznych. IFiS PAN, Warszawa 2010.

zwaną współczynnikiem Yule'a, a także pojęcie ilorazu szans. W 1935 roku Bartlett jako pierwszy zaproponował metodę estymacji największej wiarygodności, a w latach następnych Deming i Stephan wykorzystanie algorytmu dopasowania iteracyjno-proporcjonalnego. Wilks natomiast zaproponował iloraz wiarygodności, który jest alternatywny dla statystyki chi-kwadrat Pearsona, natomiast jego modyfikację zaproponował Neyman.

Forma zapisu zmiennych niemetrycznych w postaci łącznego rozkładu zmiennych sprawiła, że tablice stały się najdogodniejszym sposobem zapisu wielu zmiennych. Wraz ze wzrostem liczby badanych zmiennych komplikuje się sposób ich analizy. Zapotrzebowanie na wyspecjalizowane narzędzia umożliwiające analizę dużych zbiorów danych jest obecnie tak duże, że wywołało konieczność rozwoju wyspecjalizowanych technologii. Wiek XX stał się kluczowym okresem przełomowym w analizie danych jakościowych. W niniejszym artykule przez dane jakościowe rozumiane będą zmienne mierzone na słabych skalach pomiaru (skala nominalna, porządkowa). W latach 60. powstały wyspecjalizowane metody analizy wielowymiarowych tablic kontyngencji pozwalające na przedstawienie zależności zachodzącej pomiędzy dowolną liczną zmiennych. Metody te nazywane są modelowymi metodami analizy danych (*model--based methods*), gdyż w wyniku analizy buduje się formalny model opisujący charakter zależności zachodzącej pomiędzy zmiennymi.

W niniejszym artykule zaprezentowane zostaną modelowe metody przeznaczone do analizy danych wielowymiarowych tablic kontyngencji. Celem artykułu jest prezentacja zastosowania analizy logarytmiczno-liniowej w opisywaniu zjawisk o charakterze ekonomicznym, a także wykorzystanie prezentowanej metody w programie **R**.

1. Modelowe metody analizy tablic kontyngencji

Analiza tablic kontyngencji pozwala na badanie zależności pomiędzy kilkoma zmiennymi niemetrycznymi (nominalnymi lub porządkowymi). Tradycyjnym sposobem analizy związku pomiędzy zmiennymi niemetrycznymi w tablicach dwuwymiarowych jest wyznaczenie współczynnika chi-kwadrat lub innych statystyk na nim opartych (Yule'a, Czuprowa, Cramera, Pearsona), które mówią jedynie o sile i kierunku zależności. Taki sposób pomiaru zależności należy do metod niemodelowych. Gdy analizie poddana jest wielowymiarowa tablica kontyngencji, współczynniki te stają się niewystarczalne i powinny wówczas zostać zastosowane metody modelowe, których wynikiem jest formalny model opisu zależności.

W programie **R** tablice kontyngencji zapisane mogą zostać w postaci: *case form*, *frequency form* lub *table form*. Mogą one także zostać przekształcane z jednej postaci w inną dzięki funkcjom: `expand.dft()`, `as.data.frame()`, `xtabs(~A+B)`, `table(~A,B)`.

Jedną z modelowych metod pozwalających na opisanie struktury zależności pomiędzy zmiennymi nominalnymi oraz porządkowymi jest analiza logarytmiczno-liniowa. Metoda ta pozwala na zbudowanie wielu modeli określających strukturę zależności pomiędzy zmiennymi mierzonymi na skali nominalnej, jak i porządkowej. Ponadto metodę tę wyróżniają liczne własności, których nie posiadają niemodelowe metody analizy danych takie jak: możliwość wizualizacji wyników w postaci zaawansowanych graficznych wykresów, wykorzystanie znanej metody estymacji parametrów, szczegółowa analiza charakteru zależności oraz możliwość analizy nieograniczonej liczby zmiennych i kategorii.

2. Analiza logarytmiczno-liniowa

Analiza logarytmiczno-liniowa pozwala na zbadanie zależności pomiędzy zmiennymi niemetrycznymi, mierzonymi zarówno na skali nominalnej, jak i porządkowej, bez podziału na zmienną zależną i niezależną. W analizie logarytmiczno-liniowej rolę zmiennej zależnej odgrywają liczebności teoretyczne o rozkładzie Poissona, natomiast zmiennymi objaśniającymi są zmienne niemetryczne oraz ich kategorie.

Dla trójwymiarowej tablicy o liczebnościach empirycznych n_{hjk} $(h = 1,...H, j = 1,...,J, k = 1,...,K)$ model logarytmiczno-liniowy określony jest równaniem[2]:

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

gdzie:
m_{hjk} oznaczają liczebności empiryczne, a λ są parametrami modelu wyznaczonymi metodą największej wiarygodności.

[2] Y.M.M. Bishop et al.: Discrete multivariate analysis. MIT Press, Cambridge, Massachusetts 1975.

W celu wyznaczenia parametrów modelu spełniony jest warunek:

$$\sum_{h=1}^{H}\lambda_h^X=\sum_{j=1}^{J}\lambda_j^Y=\sum_{k=1}^{K}\lambda_k^Z=0,$$
$$\sum_{h=1}^{H}\lambda_{hj}^{XY}=\sum_{j=1}^{J}\lambda_{hj}^{XY}=\sum_{h=1}^{H}\lambda_{hk}^{XZ}=\sum_{k=1}^{K}\lambda_{hk}^{XZ}=\sum_{j=1}^{J}\lambda_{jk}^{YZ}=\sum_{k=1}^{K}\lambda_{jk}^{YZ}=0, \quad (2)$$
$$\sum_{h=1}^{H}\lambda_{hjk}^{XYZ}=\sum_{j=1}^{J}\lambda_{hjk}^{XYZ}=\sum_{k=1}^{K}\lambda_{hjk}^{XYZ}=0.$$

Gdy zmienne mają charakter porządkowy, możliwymi do zbudowania modelami są: jednorodny model asocjacji (*uniform association*), model efektów wierszowych i kolumnowych (*row-effects and column-effects model*) oraz model RC Goodmana (*Goodman's RC model, row and column effects model*).

Modelem, który wykorzystuje porządek kategorii zmiennych wierszowych oraz kolumnowych, jest jednorodny model o równaniu:

$$\ln\left(m_{hj}\right)=\mu+\lambda_h^X+\lambda_j^Y+\beta\left(u_h-\bar{u}\right)\left(v_j-\bar{v}\right). \quad (3)$$

Model efektu wierszowego zdefiniowany jest jako:

$$\ln\left(m_{hj}\right)=\mu+\lambda_h^X+\lambda_j^Y+\tau_h\left(v_j-\bar{v}\right), \quad (4)$$

natomiast model efektu kolumnowego można zapisać jako:

$$\ln\left(m_{hj}\right)=\mu+\lambda_h^X+\lambda_j^Y+\tau_j\left(u_h-\bar{u}\right). \quad (5)$$

Szczegółowy opis modeli logarytmiczno-liniowych dla zmiennych porządkowych znaleźć można w pracy Masako Ishii-Kuntz[3].

W analizie logarytmiczno-liniowej wykorzystywane są formalne kryteria pozwalające na ocenę stopnia dopasowania modelu do danych. Należą do nich współczynnik chi-kwadrat i iloraz wiarygodności, kryteria informacyjne oraz współczynnik determinacji.

W celu wyboru modelu najlepiej dopasowanego do danych wykorzystuje się współczynnik iloraz wiarygodności G^2 zdefiniowany jako[4]:

---

[3] M. Ishii-Kuntz: Ordinal log-linear models. Sage University Paper Series on Quantitative Applications in the Social Science, Series No. 07-097. Sage, Beverly Hills, London 1994.

[4] A. Agresti: Analysis of ordinal categorical data. John Wiley & Sons, Hoboken, New Jersey 2010; A. Agresti: Categorical data analysis. John Wiley & Sons, Hoboken, New Jersey 2002; R. Christensen: Log-linear models and logistic regression. Springer-Verlag, New York 1997; Y.M.M Bishop et al., op. cit.

$$G^2 = 2\sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{hjk} \ln\left(\frac{n_{hjk}}{m_{hjk}}\right). \quad (6)$$

Współczynnik ten wykorzystuje się do porównywania ze sobą modeli sąsiednich, budowanych wedle zasady hierarchiczności[5].

Kolejną statystyką służącą do porównania większej ilości modeli jest kryterium informacyjne Akaike *AIC* (*Akaike Information Criteria*)[6]:

$$AIC = G^2 - 2df\ , \quad (7)$$

gdzie df oznacza liczbę stopni swobody.

Kryterium Bayesowskie BIC (*Bayesian Information Criteria*) jest drugim kryterium postaci:

$$BIC = G^2 - df \cdot \ln n\ , \quad (8)$$

gdzie n oznacza liczebność tablicy kontyngencji[7].

Minimalna wartość kryteriów informacyjnych pozwala na wybór najlepszego modelu logarytmiczno-liniowego. Ich istotą nie jest wskazanie 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.

3. Wykorzystanie analizy logarytmiczno-liniowej w programie R

W niniejszym badaniu wykorzystano dane sondażowe opublikowane przez Radę Monitoringu Społecznego w raporcie „Diagnoza Społeczna 2013. Warunki i jakość życia Polaków". Raport dotyczy czasu spędzanego przed telewizorem przez osoby powyżej 18. roku życia względem wieku. Próba liczyła 26 307 respondentów. W badaniu uwzględniono dwie zmienne: *czas oglądania telewizji* (0-1, 1-3, 3 i więcej godzin) oraz *wiek* (18-24, 25-34, 35-44, 45-59, 60-64, 65 i więcej lat).

---

[5] S.E. Fienberg: The analysis of multidimensional contingency tables. „Ecology" 1970, No. 51, s. 419-433; D. Knoke, P.J. Burke: Log-linear models. Sage University Paper Series on Quantitative Applications in the Social Science. Series No. 07-020. Sage, Beverly Hills and London 1980.

[6] H. Akaike: Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd International Symposium on Information. Akademiai Kiado, Budapest 1973.

[7] A.E. Raftery: Choosing models for cross-classification. „American Sociological Review" 1986, No. 51, s. 145, 146; G. Schwartz: Estimating the dimensions of a model. „Annals of Statistics" 1978, No. 6, s. 461-464.

W celu oceny współwystępowania kategorii zmiennych oraz oceny siły zależności przeprowadzono klasyczną analizę korespondencji, traktując wszystkie zmienne jako nominalne. Wartość inercji całkowitej $\lambda = 0{,}0614$ jest niewielka i wskazuje na brak zależności pomiędzy zmiennymi. Liczba wymiarów rzutowania wynosi 2, z czego pierwszy wymiar wyjaśnia 91,3% inercji całkowitej, natomiast dwa wymiary wyjaśniają łącznie 100% inercji całkowitej. W sytuacji tej metoda współwystępowania okazała się nieskuteczna, a zależność pomiędzy zmiennymi nie została wykryta.

W celu przeprowadzenia pogłębionej analizy zależności pomiędzy opisanymi zmiennymi można zastosować analizę logarytmiczno-liniową. Metoda ta jest modelową analizą zależności i pozwala na analizę zmiennych o niewielkiej liczbie kategorii. Uwzględnia także porządek kategorii zmiennych, co w analizie korespondencji jest niemożliwe. Ponadto w wyniku przeprowadzonej metody wybrany zostanie model opisujący liczebności teoretyczne. Zbudowane zostaną modele prezentujące różne rodzaje zależności i uwzględniające porządek kategorii zmiennych w różnych konfiguracjach, tzn. raz traktowane są jako nominalne, a kolejny raz jako porządkowe. Zbudowane zostaną następujące modele: jednorodny model asocjacji, model efektów wierszowych oraz model efektów kolumnowych.

Wartości mierników oceny modeli przedstawiono w tabeli 1.

Tabela 1

Wartości mierników oceny dopasowania modeli do danych

Model	G^2	df	AIC
Model pełny	0	0	196,93
Jednorodny model asocjacji	190,10	9	369,03
Model efektów wierszowych	182,64	8	363,58
Model efektów kolumnowych	85,52	5	272,46
Model niezależności	1606,60	10	1783,50
Model RC Goodmana	79,15	0	79,15

Spośród zbudowanych modeli porządkowych najlepsze dopasowanie do danych zapewnia model efektów kolumnowych. Dla tego modelu zarówno współczynnik $G^2 = 85{,}52$ przy liczbie stopni swobody równej $df = 5$, jak i kryteria informacyjne osiągają najmniejsze wartości. Widoczne jest to, że wartości parametrów dla interakcji pomiędzy wiekiem a rangami przypisanymi zmiennej kolumnowej stale rosną (0,0595; 0,1763; 0,4348; 0,8631; 0,9656). Oznacza to, że dla danej kolumny dodatnie znaki parametrów wskazują, iż więcej obserwacji pojawia się w kolumnach reprezentujących wysokie wartości zmiennej porządkowej, a mniej w kolumnach o niższych wartościach w porównaniu z występowaniem niezależności zmiennych.

Z przeprowadzonej analizy wynika, że dla porządkowych modeli logarytmiczno-liniowych uzyskane wyniki są znacznie lepsze niż w przypadku modeli dla zmiennych nominalnych. Współczynniki G^2 we wszystkich trzech przypadkach (jednorodny model asocjacji, model efektów wierszowych oraz model efektów kolumnowych) osiągają znacznie mniejszą wartość niż w przypadku modelu niezależności. Przeprowadzone badanie pokazuje, że modele porządkowe wypełniają obszerną lukę istniejącą pomiędzy modelem pełnym a modelem niezależności, zapewniając tym samym znaczną część informacji, której analiza nie jest możliwa w przypadku zmiennych nominalnych.

W badaniach ekonomicznych nie zawsze dysponuje się pełną informacją na temat zjawiska, a badana tablica kontyngencji może zawierać zerowe liczebności. W niniejszym badaniu przeprowadzono analizę logarytmiczno-liniową dla tablicy kontyngencji zawierającej zerowe komórki. Dane wykorzystane do analizy logarytmiczno-liniowej pochodzą z Wyższego Urzędu Górniczego w Polsce (www.wug.gov.pl) i dotyczą łącznej liczby wypadków w pracy w górnictwie w 2013 roku. Zbudowano trójwymiarową tablicę przedstawiającą łączny rozkład liczby ofiar wypadków dla następujących zmiennych:

- górnictwo (G) (górnictwo węgla kamiennego, górnictwo rud miedzi, górnictwo odkrywkowe, górnictwo otworkowe, pozostałe),
- załoga (Z) (załoga własna, firmy usługowe),
- wypadki (W) (śmiertelne, ciężko ranni, inne).

Dla badanej grupy liczącej 2588 wypadków tablica ma wymiary $2 \times 3 \times 5$ i spośród 30 komórek 8 zawiera zerowe liczebności. Analiza liczebności trójwymiarowej tablicy wypadków w górnictwie pokazuje, iż kopalniami, w których najczęściej dochodziło do obrażeń, były kopalnie węgla kamiennego (w sumie 1482 wypadki wśród górników pracujących w załodze własnej i 455 wypadków wśród górników zatrudnionych w firmach usługowych). Najliczniejszą grupę wśród rannych stanowili górnicy należący do komórki opisującej poszkodowanych w załodze własnej, którzy odnieśli inne obrażenia (1471 osób). Brak śmiertelnych wypadków odnotowano natomiast w górnictwie otworkowym i innym, zarówno wśród załogi własnej, jak i wśród górników zatrudnionych w firmach usługowych. Brak poważnych wypadków odnotowano w górnictwie odkrywkowym, otworkowym oraz innym u górników zatrudnionych w firmach usługowych. Ze względu na to, że badana tablica zawiera zerowe liczebności, nie jest możliwe przeprowadzenie analizy korespondencji, gdyż metoda ta nie powinna być stosowana w przypadku tablic zawierających zera. W badanym przykładzie liczba wypadków jest rezultatem wpływu procesów opisywanych zmiennymi oraz interakcjami pomiędzy nimi. Ze zbioru wszystkich możliwych modeli

z trzema zmiennymi wybrano jeden model optymalny. Z przeprowadzonego badania wynika, że najlepszym modelem opisującym liczbę osób poszkodowanych w wypadkach górniczych jest model zależności homogenicznej [GZ][GW][ZW]. Dla tego modelu iloraz wiarygodności 12, 804 przy df = 8, z prawdopodobieństwem testowym p = 0, 118. Model ten jest modelem złożonym, gdyż zawiera wszystkie możliwe interakcje pomiędzy zmiennymi. Równanie modelu zapisać można w postaci:

$$lm(m_{hjkl}) = \lambda + \lambda_h^G + \lambda_j^Z + \lambda_k^W + \lambda_{hj}^{GZ} + \lambda_{hk}^{GW} + \lambda_{jk}^{ZW} . \tag{9}$$

Model ten pozwala opisać strukturę zależności zachodzącej pomiędzy zmiennymi opisującymi tablicę kontyngencji z zerowymi komórkami.

Dla modelu zależności homogenicznej można wyznaczyć reszty Pearsona zdefiniowane jako:

$$\widetilde{r}_{hj} = \frac{n_{hj} - \hat{m}_{hj}}{\sqrt{\hat{m}_{hj}}} . \tag{10}$$

Ich wartości wskazują na odchylenia każdej liczebności tablicy od wyznaczonych na podstawie modelu liczebności teoretycznych. Im większe odchylenia liczebności, tym model wykazuje słabsze dopasowanie do danych. Model z zerowymi odchyleniami jest modelem doskonale dopasowanym do danych i takie reszty zaobserwować można jedynie dla modelu pełnego, w którym liczebności empiryczne są równe liczebnościom teoretycznym. Dla badanej tablicy kontyngencji wyznaczono reszty Pearsona (10).

```
, , Wypadki = Śmiertelne

                          Załoga
Kopalnia                  Własna            Firmy
 Węgla kamiennego      -0.04881384       0.06643288
 Rud miedzi             0.88052887      -1.59913788
 Odkrywkowa            -0.68504600       1.03616781
 Otworkowa              0.00000000       0.00000000
 Inne                   0.00000000       0.00000000

, , Wypadki = Inne

                          Załoga
Kopalnia                  Własna            Firmy
 Węgla kamiennego       0.05118689      -0.09290097
 Rud miedzi            -0.04997587       0.07650800
 Odkrywkowa             0.06189362      -0.15664216
 Otworkowa             -0.14914589       0.24983695
 Inne                  -0.13960221       0.29520364
```

```
, , Wypadki = Ciężko ranni
                          Załoga
Kopalnia                  Własna                 Firmy
 Węgla kamiennego       -1.0247628             0.9399970
 Rud miedzi             -0.3443789             0.2915457
 Odkrywkowa              0.4826143            -1.1015727
 Otworkowa               0.5884987            -0.9790257
 Inne                    0.7487361            -1.4945057
```

Największe odchylenia widoczne są dla komórek odpowiadających górnikom, którzy byli zatrudnieni w firmach usługowych i ulegli śmiertelnemu wypadkowi podczas pracy w kopalni rud miedzi (-1,5991) oraz byli ciężko ranni podczas pracy w innych kopalniach (-1,4945) i w kopalniach odkrywkowych (-1,1016). Zerowe odchylenia, które odpowiadają zerowym liczebnościom empirycznym, widoczne są w komórkach dotyczących wypadków śmiertelnych, które wydarzyły się w załogach własnych, jak i w firmach usługowych.

Podsumowanie

Istnieje wiele metod analizy danych niemetrycznych w postaci tablic kontyngencji. Większość klasycznych metod ograniczona jest jedynie do analizy zależności dwóch zmiennych nominalnych za pomocą klasycznych współczynników zależności. W niniejszym artykule zaprezentowano analizę logarytmiczno-liniową, która wykorzystuje formalny model opisujący zależność zachodzącą pomiędzy zmiennymi. Metoda ta z powodzeniem może być stosowana zarówno dla zmiennych nominalnych, jak i porządkowych. Ponadto zapewnia ona formalny model liniowy opisujący strukturę zależności i uwzględnia interakcje zachodzące między badanymi zmiennymi. Dodatkowo wykorzystuje ona znaną metodę estymacji parametrów – metodę największej wiarygodności. Metoda ta nie wymaga też spełnienia żadnych założeń i może być stosowana dla dowolnie dużej liczby zmiennych. Ponadto zaprezentowano model logarytmiczno-liniowy dla tablicy kontyngencji opisującej liczebność wypadków w górnictwie względem trzech zmiennych nominalnych w roku 2013.

W niniejszym artykule zaprezentowano wykorzystanie analizy logarytmiczno-liniowej do analizy tablic kontyngencji zawierających zarówno zmienne nominalne, jak i porządkowe. Wybrano model najlepiej dopasowany do danych, dla którego odchylenia liczebności empirycznych od teoretycznych są najmniejsze. Wszelkie obliczenia wykonane zostały w programie **R** z wykorzystaniem pakietu `loglm` oraz `glm`. Przy użyciu analizy logarytmiczno-liniowej możliwy jest również opis zależności zachodzących pomiędzy zmiennymi porządkowymi.

Literatura

Agresti A.: Analysis of ordinal categorical data. John Wiley & Sons, Hoboken, New Jersey 2010.

Agresti A.: Categorical data analysis. John Wiley & Sons, Hoboken, New Jersey 2002.

Akaike H.: Information theory and an extension of the maximum likelihood principle. „Proceedings of the 2nd International Symposium on Information". 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.

Fienberg S.E.: The analysis of multidimensional contingency tables. „Ecology" 1970, No. 51.

Ishii-Kuntz M.: Ordinal log-linear models. Sage University Paper Series on Quantitative Applications in the Social Science. Series No. 07-097. Sage, Beverly Hills, London 1994.

Knoke D., Burke P.J.: Log-linear models. Sage University Paper Series on Quantitative Applications in the Social Science. Series No. 07-020. Sage, Beverly Hills and London 1980.

Raftery A.E.: Choosing models for cross-classification. „American Sociological Review" 1986, No. 51.

Sawiński Z.: Zastosowania tablic w badaniach zjawisk społecznych. IFiS PAN, Warszawa 2010.

Schwartz G.: Estimating the dimensions of a model. „Annals of Statistics" 1978, No. 6.

MODEL-BASED METHODS FOR MULTI-WAY FREQUENCY TABLES IN A PUBLIC OPINION SURVEY

Summary

The methods for analyzing cross-classified tables are usually to test relations between two variables taken one pair at a time. Further development of those methods allowed to move from two dimensional tables to high dimensional tables, where dimensionality of a cross-table refers to the number of variables. It allowed to transform non-model-based to model-based methods providing the equation of a mathematical model, the use of estimation method and variety of visualizing tools.

This paper describes how complex qualitative data may be described by a mathematical model. One of the method presented is log-linear analysis.