PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1123 — 2006
Ekonometria 17

Iwona Kasprzyk

MODELE LOG-LINIOWE DLA ZMIENNYCH PORZĄDKOWYCH

1. Wstęp

W badaniach marketingowych spotyka się często zmienne mierzone na skalach nominalnej oraz porządkowej. Za pomocą modeli log-liniowych można pokazać, jak powiązane są zmienne uwzględnione w tablicy kontyngencji, oraz czy zmienne te są od siebie zależne.

Celem tego artykułu jest zaprezentowanie modeli log-liniowych uwzględniających zmienne mierzone na skali porządkowej, tzw. *ordinal log-linear model*. Tego typu modele pozwalają nie tylko na określenie zależności zmiennych w tablicy kontyngencji, ale także pokazują, jak silny jest związek pomiędzy zmiennymi X oraz Y. Modele te wprowadził L.A. Goodman [4]. Następnie były rozwijane przez C.C. Clogga [3] oraz A. Agrestiego [1], a później zajmował się nimi J. Vermunt [14]. W prezentowanym artykule zostaną przedstawione różne modele dla zmiennych porządkowych, przede wszystkim z uwzględnieniem tablicy dwudzielczej. Program *LEM* pozwala m.in. na estymację parametrów modeli log-liniowych.

2. Iloraz szans

Do opisania parametrów modelu log-liniowego można posłużyć się tzw. ilorazem szans (*odds ratio*). Iloraz szans jest podstawową miarą opisu stopnia związku zmiennych zawartych w tablicy kontyngencji. Dla dwudzielczej tablicy kontyngencji o wymiarach $r \times c$ można rozważyć podzbiór $(r-1)(c-1)$ „lokalnych ilorazów szans". Miarę tę można zapisać za pomocą poniższej formuły:

$$\theta_{ij} = \frac{p_{i,j}p_{i+1,j+1}}{p_{i,j+1}p_{i+1,j}} \text{ dla } i = 1, 2, \ldots, r-1,\ j = 1, 2, \ldots, c-1.$$

Każdy iloraz szans wskazuje kombinację dwóch kategorii zmiennej X i dwóch kategorii zmiennej Y. W modelach log-liniowych posługujemy się logarytmem ilorazu szans ($\log\theta_{ij}$).

3. Modele log-liniowe dla zmiennych nominalnych

Niech F_{ij} oznacza częstość oczekiwaną ($i=1,2,\ldots,r$ oraz $j=1,2,\ldots,c$. Podstawowy model uwzględniający efekt interakcji pomiędzy dwiema zmiennymi (λ_{ij}^{XY}) zawartymi w tablicy kontyngencji można przedstawić za pomocą poniższej formuły:

$$\log F_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \tag{1}$$

gdzie: μ – średnia logarytmów częstości oczekiwanych,

λ_i^X – główny efekt zmiennej X,

λ_j^Y – główny efekt zmiennej Y.

Parametry modelu (1) są obliczane za pomocą formuł:

$$\mu = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\log F_{ij}}{rc},\quad \lambda_i^X = \frac{\sum_{j=1}^{c}\log F_{ij}}{c} - \mu,\quad \lambda_j^Y = \frac{\sum_{i=1}^{r}\log F_{ij}}{r} - \mu,$$

$$\lambda_{ij}^{XY} = \log F_{ij} - \mu - \lambda_i^X - \lambda_j^Y. \tag{2}$$

Model (1) spełnia następujące warunki:

$$\sum_{i=1}^{r} u_i^X = \sum_{j=1}^{c} u_j^Y = 0,\ \sum_{i=1}^{r} u_{ij}^{XY} = \sum_{j=1}^{c} u_{ij}^{XY} = 0.$$

Gdy parametr λ_{ij}^{XY} jest równy zeru, model (1) przybiera następującą postać:

$$\log F_{ij} = \mu + \lambda_i^X + \lambda_j^Y. \tag{3}$$

Model (3) ma $(r-1)(c-1)$ stopni swobody oraz $1+(r-1)+(c-1)=r+c-1$ liniowo niezależnych parametrów.

Model (1) nazywamy modelem pełnym (nasyconym), natomiast model (3) będziemy nazywać modelem niezależnym.

4. Jednorodny model asocjacji (*uniform association model*)

W przypadku tego modelu zmienne X oraz Y traktowane są jako zmienne mierzone na skali porządkowej. Stosujemy tutaj rangi całkowite dla obu zmiennych uwzględnionych w tablicy kontyngencji, które to pozwolą na uporządkowanie kategorii zmiennych. Niech u_i oznacza rangi wierszowe, gdzie $u_1 < u_2 < \ldots < u_r$ oraz v_j rangi kolumnowe, gdzie $v_1 < v_2 < \ldots < v_c$.

Jednorodny model asocjacji jest wyrażony za pomocą poniższej formuły (4):

$$\log F_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta(u_i - \bar{u})(v_j - \bar{v}), \tag{4}$$

gdzie parametry modelu (4) spełniają następujące ograniczenia:

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0.$$

Parametr β opisuje związek miedzy zmiennymi X i Y. Model (4) ma tylko o jeden więcej parametrów niż model niezależny oraz $rc - [1 + (r-1) + (c-1)] - 1 = rc - r - c$ stopni swobody.

Jednorodny model asocjacji może być postrzegany jako szczególny przypadek modelu efektu wierszowego albo modelu efektu kolumnowego, które zostaną omówione dalszej części artykułu.

5. Modele efektu wierszowego oraz efektu kolumnowego

Jeśli rangi całkowite są zastosowane tylko dla zmiennej Y, to taki model będziemy nazywać modelem efektu wierszowego (*row effect model*). Odpowiednio, gdy rangi całkowite zostaną wprowadzone dla zmiennej X, wówczas w rezultacie otrzymamy model, który będziemy nazywać modelem efektu kolumnowego (*column effect model*).

Model efektu wierszowego jest skonstruowany przez zastąpienie efektu interakcji między zmiennymi X oraz Y (λ_{ij}^{XY}) w modelu pełnym, składnikiem asocjacji $\tau_i(v_j - \bar{v})$:

$$\log F_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \tau_i(v_j - \bar{v}), \tag{5}$$

gdzie τ_i są parametrami efektów wierszowych, natomiast v_j są rangami kolumnowymi, gdzie rangi te muszą być uporządkowane w sposób rosnący, tj. $v_1 < v_2 < \ldots < v_c$. Model (5) ma o r-1 więcej parametrów niż model niezależny oraz $rc - [1 + (r-1) + (c-1) + (r-1)] = (r-1)(c-2)$ stopni swobody.

Parametry modelu (5) spełniają następujące ograniczenia:

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \tau_i .$$

Wyznaczone stałe $\left(v_j - \bar{v}\right)$, są równe rangą (–1, 0, +1) w przypadku, gdy w tablicy kontyngencji znajdują się trzy kategorie zmiennej Y. Składnik asocjacji, czyli $\left(v_j - \bar{v}\right)$, jest traktowany jako pewne odchylenie od modelu niezależnego.

Jak pokazano w pracy [7], parametry modelu efektu wierszowego są powiązane z ilorazem szans. Dla przypadkowych dwóch kategorii zmiennej X (h, i) oraz sąsiadujących dwóch kategorii zmiennej Y (j, $j+1$) logarytm ilorazu szans ma następującą postać:

$$\begin{aligned}\log\frac{F_{hj}F_{i,j+1}}{F_{h,j+1}F_{ij}} &= \left(\mu+\lambda_h^X+\lambda_j^Y+\tau_h\left(v_j-\bar{v}\right)+\mu+\lambda_i^X+\lambda_{j+1}^Y+\tau_i\left(v_{j-1}-\bar{v}\right)\right)- \\ &-\left(\mu+\lambda_h^X+\lambda_{j+1}^Y+\tau_h\left(v_{j+1}-\bar{v}\right)+\mu+\lambda_i^X+\lambda_j^Y+\tau_i\left(v_j-\bar{v}\right)\right)= \\ &=\left[\tau_h\left(v_j-\bar{v}\right)+\tau_i\left(v_{j+1}-\bar{v}\right)\right]-\left[\tau_h\left(v_{j+1}-\bar{v}\right)+\tau_i\left(v_j-\bar{v}\right)\right]= \\ &=\tau_h\left(v_j-v_{j+1}\right)+\tau_i\left(v_{j+1}-v_j\right)=\tau_i-\tau_h .\end{aligned} \tag{6}$$

Tabela 1. Modele log-liniowe, liczba stopni swobody oraz logarytm ilorazu szans

Modele	Liczba stopni swobody	Logarytm ilorazu szans
Model niezależny	$(r-1)(c-1)$	
Jednorodny model asocjacji	$rc-r-c$	$\beta(\mu_b-\mu_a)(\nu_d-\nu_c)$
Model efektu wierszowego	$(r-1)(c-2)$	$\tau_i-\tau_h$
Model efektu kolumnowego	$(r-1)(c-2)$	$\nu_{j+1}-\nu_j$
Model RC	$(r-2)(c-2)$	$\beta(\mu_{i+1}-\mu_i)(\nu_{j+1}-\nu_j)$

Źródło: opracowanie własne.

Tabela 1 pokazuje dodatkowo iloraz szans dla pozostałych modeli rozważanych w tym artykule.

W modelu efektu kolumnowego zmienne wierszowe są opisane na skali porządkowej, zmienne kolumnowe opisane są zaś za pomocą skali nominalnej. Model tego typu można zapisać przy pomocy poniższej formuły:

$$\log F_{ij} = \mu+\lambda_i^X+\lambda_j^Y+\tau_i\left(u_j-\bar{u}\right), \tag{7}$$

gdzie:

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \tau_i .$$

Podobnie jak model efektu wierszowego, model ten ma $(r-1)(c-2)$ stopni swobody.

6. Model RC

Model ten został zaproponowany przez Goodmana w roku 1979. Potocznie jest on nazywany modelem RC (*row and column effects model*), będącym specjalnym przypadkiem modelu log-multiplikatywnego, w którym to rangi wierszowe oraz kolumnowe są parametrami.

Kiedy mamy do czynienia z dwiema zmiennymi, model RC przyjmuje następującą postać:

$$\log F_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j, \tag{8}$$

gdzie β jest parametrem asocjacji, a u_i oraz v_j są nieznanymi rangami, które będą estymowane. Parametry tego modelu muszą spełnić następujące ograniczenia, a mianowicie:

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i u_i = \sum_j v_j = 0,$$
$$\sum_i u_i^2 = \sum_j v_j^2 = 1.$$

Jeśli $\beta = 0$, to model (6) sprowadza się do modelu niezależnego. Model ten ma $rc - [1 + (r-1) + (c-1) + 1 + (r-2) + (c-2)] = (r-1)(c-2)$ stopni swobody. Liczba stopni swobody dla modeli rozważanych do tej pory zaprezentowana jest w tabeli 1.

Model (8) nie wymaga uporządkowania kategorii wierszy czy też kolumn. Estymacja ocen ujawnia ukryty porządek kategorii w modelu [11].

Model log-multiplikatywny (tzw. model RC) może być uogólniony w celu zastosowania go dla trójdzielczej i wielodzielczej tablicy. Taki model nazywamy modelem RC(M) i przyjmuje on następującą postać:

$$\log F_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \sum_{m=1}^{M} \beta_m \mu_{im} \nu_{jm} \tag{9}$$

z warunkami:

$$\sum_i u_i = \sum_j v_j = 0,$$
$$\sum_i u_i^2 = \sum_j v_j^2 = 1, \tag{10}$$

gdzie β_m mierzy siłę związku dla m-tego wymiaru.

7. Przykład

CBOS często prowadzi badania dotyczące aktualnej sytuacji na rynku pracy. W jednej z ankiet poproszeno respondentów o odpowiedź na m.in. pytanie: „Jak Pan/i ocenia obecną sytuację na rynku pracy w Polsce?" W budowie tego typu modeli nie zostały uwzględnione odpowiedzi „trudno powiedzieć". Przykład ma na celu pokazanie, czy istnieje zależność oceny sytuacji na rynku pracy wśród osób biernych zawodowo. Na podstawie danych zawartych w tab. 2 zostanie pokazany model niezależny oraz model efektu wierszowego.

Tabela 2. Obserwowane i estymowane częstości dla modelu niezależnego i modelu efektu wierszowego

Bierni zawodowo	Jak Pan(i) ogólnie ocenia sytuację na rynku pracy w Polsce? Czy jest ona:				
	bardzo dobra	dobra	ani dobra, ani zła	zła	bardzo zła
Renciści	0 (0,386)[1] (0,119)[2]	1 (0,386) (0,189)	1 (3,478) (2,461)	39 (41,355) (38,032)	62 (57,394) (62,199)
Emeryci	2 (0,732) (0,594)	0 (0,732) (0,720)	6 (6,585) (7,099)	82 (78,293) (83,257)	105 (108,659) (103,329)
Uczniowie i studenci	0 (0,281) (1,149)	1 (0,281) (0,862)	6 (2,533) (5,268)	41 (30,113) (38,282)	27 (41,792) (29,440)
Bezrobotni	0 (0,443) (0,057)	0 (0,443) (0,116)	2 (3,985) (1,912)	38 (47,377) (37,615)	78 (65,752) (78,300)
Gospodynie domowe i inni	0 (0,158) (0,082)	0 (0,158) (0,112)	3 (1,418) (1,261)	14 (16,863) (16,814)	25 (23,403) (23,732)

[1] Częstości oczekiwane dla modelu niezależnego.

[2] Częstości oczekiwane dla modelu efektu wierszowego.

Źródło: opracowanie własne na podstawie [15].

Do testowania niezależności zmiennych zawartych w tablicy kontyngencji można wykorzystać statystykę chi-kwadrat. Bardziej jednak powszechnym testem stosowanym w przypadku modelowania log-liniowego jest współczynnik największej wiarygodności L^2:

$$L^2 = 2\left(\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{ij}}{E_{ij}}\right)\right). \tag{11}$$

Oprócz statystyki χ^2 oraz L^2 można znaleźć w literaturze również inny test niezależności, a mianowicie test Cressiego-Roada, który można przedstawić za pomocą poniższej formuły:

$$CR = \frac{2}{\lambda(\lambda+1)} \sum_i \sum_j n_{ij} \left[\left(\frac{n_{ij}}{E_{ij}} \right)^{\lambda} - 1 \right]. \tag{12}$$

Formuła ta wprowadzona przez Reada i Cressiego w pracy [12] jest równoważna statystyce χ^2 dla $\lambda = 1$ oraz statystyce L^2, gdy $\lambda \to 0$. Read i Cressie sugerują, żeby za λ przyjmować wartość 2/3.

W przypadku wyboru modelu najlepszego można posłużyć się kryterium Akaikego, które można zapisać za pomocą formuły:

$$A = L^2(M) - 2 \cdot (df), \tag{13}$$

gdzie $L^2(M)$ oznacza wartość statystyki chi-kwadrat największej wiarygodności dla analizowanego modelu *M*. Im mniejsza wartość statystyki, tym wybrany model jest lepiej dopasowany.

W tabeli 2 zostały ujęte obserwowane oraz oczekiwane częstości dla modelu niezależnego i modelu efektu wierszowego. Miary jakości dopasowania są podsumowane w tab. 3. Jakość modelu niezależnego jest znacznie uboższa niż modelu efektu wierszowego z $L^2 = 12{,}64$ opartego na 12 stopniach swobody. Dodatkowo można stwierdzić, że istnieje silny związek pomiędzy zmiennymi zawartymi w tab. 1. Różnica testu chi-kwadrat największej wiarygodności pomiędzy modelem niezależnym a modelem efektu wierszowego ($\Delta L^2 = 19{,}38$ z $df = 4$) pokazuje, że parametry efektu wierszowego znacznie poprawiają jakość modelu. Innymi słowy, występuje tutaj znacząca różnica w ocenie aktualnej sytuacji rynku pracy Polski w maju 2004 r. wśród osób biernych zawodowo.

Tabela 3. Miary jakości dopasowania dla oceny sytuacji i biernych zawodowo

Model	L^2	*df*	*p*	*CR*	*p*	*AIC*
Model niezależny	32,0214	16	0,0099	31,1245	0,0130	0,0214
Model efektu wierszowego	12,6416	12	0,3956	12,7042	0,3909	–11,3584

Źródło: opracowanie własne z wykorzystaniem programu *LEM*.

Tabela 3 pokazuje wartości współczynnika największej wiarygodności oraz testu Cressiego-Reada wraz z poziomem prawdopodobieństwa *p*. Jeśli testy te są istotne, to wówczas dany model zostaje odrzucony. W rozpatrywanym przykładzie model efektu wierszowego zadowalająco wyjaśnia liczebności w tab. 2 (testy nie są istotne).

Za pomocą programu *LEM* dokonano estymacji parametrów modelu efektu wierszowego:

$$\mu = 1{,}1546,$$
$$\lambda_1^S = -0{,}1799,\ \lambda_2^S = 0{,}8796,\ \lambda_3^S = 0{,}5812,\ \lambda_4^S = -0{,}4321,\ \lambda_5^S = -0{,}8488,$$
$$\lambda_1^P = -2{,}7318,\ \lambda_2^P = -2{,}4512,\ \lambda_3^P = -0{,}0743,\ \lambda_4^P = 2{,}4763,\ \lambda_5^P = 2{,}7810.$$
$$\tau_1 = 0{,}1873,\ \tau_2 = -0{,}0886,\ \tau_3 = -0{,}5672,\ \tau_4 = 0{,}4285,\ \tau_5 = 0{,}0400.$$

Dodatnie wartości τ wskazują, że badani respondenci wykazują większą tendencję do oceny sytuacji rynku pracy w końcu skali porządkowej, czyli określają ją jako bardzo złą.

Za pomocą formuły (6) można obliczyć iloraz szans. Na przykład dla bezrobotnych i emerytów wynosi on:

$$\theta_{24} = e^{\tau_4 - \tau_2} = e^{-0{,}5171} = 0{,}5962.$$

Tabela 4. Ilorazy szans dla osób biernych zawodowo

Bierni zawodowo	Renciści	Emeryci	Uczniowie i studenci	Bezrobotni	Gospodynie domowe i inni
Renciści	–	1,3177	2,1265	0,7857	1,1587
Emeryci	0,7589	–	1,6138	0,5962	0,8793
Uczniowie i studenci	0,4702	0,6197	–	0,3695	0,5449
Bezrobotni	1,2728	1,6772	2,7066	–	1,4748
Gospodynie domowe i inni	0,8630	1,1372	1,8353	0,6781	–

Źródło: opracowanie własne.

W związku z tym, że otrzymana wartość ilorazu szans jest mniejsza niż 1, można zastosować w tym przypadku odwrotność liczby 0,5962 (tj. 1/0,5962 = 1,6772), co w efekcie ułatwi interpretację otrzymanego wyniku. Wartość 1,67 oznacza, że szansa bezrobotnych oceniających sytuację na rynku pracy w Polsce jako bardzo złą jest 1,67 raza większa niż szansa emerytów. Podobnie można zinterpretować inne grupy badanych osób: np. szansa rencistów oceniających sytuację jako bardzo złą jest 2,13 raza większa niż uczniów i studentów.

8. Zakończenie

Wyraźną zaletą modeli log-liniowych uwzględniających zmienne na skali porządkowej jest to, że w porównaniu z modelami log-liniowymi parametry są łatwiejsze w interpretacji. Stosując iloraz szans, można łatwiej opisać związek pomiędzy zmiennymi mierzonymi na tego typu skali. Jak pokazano w przykładzie, zastosowanie modelu efektu wierszowego poprawiło jakość dopasowania modelu oraz okazało się, że istnieje dość duża zależność pomiędzy badanymi zmiennymi. Podany przykład dowodzi, że badani respondenci ocenili aktualną sytuację na rynku pracy w końcu skali porządkowej.

Literatura

[1] Agresti A., *Categorial Data Analysis*, Wiley, New York 1990.

[2] Christensen R., *Log-Linear Models and Logistic Regression*, Springer-Verlag Inc., New York 1997.

[3] Clogg C.C., *Some Models for the Analysis of Association Multiway Cross-Classification Having Ordered Categories*, „Journal of the American Statistical Association" 1982, nr 380, s. 803-815.

[4] Goodman L.A., *Simple Models for the Analysis of Association in Cross-Classificaions Having Ordered Categories*, „Journal of the American Statistical Association" 1979, nr 367, s. 537-552.

[5] Goodman L.A., *The Multivariate Analysis of Qualitative Data: Interactions among Multiple Classifications*, „Journal of the American Statistical Association" 1970, nr 329, s. 226-255.

[6] Green J.A., *Loglinear Analysis of Cross-Classified Ordinal Data: Applications in Developmental Research*, „Society for Research in Child Development" 1988, nr 59, s.1-25.

[7] Ishii-Kuntz M., *Ordinal Log-Linear Models*, Sage Publications, Beverly Hills, London 1994.

[8] Jeansone A., *Loglinear Models*, http://userwww.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.pdf.

[9] Jobson J.D., *Categorical and Multivariate Method*, Vol. 2, Springer-Verlag, New York 1992.

[10] Knoke D., Burke P.J., *Log-Linear Model*, Sage Publications, Beverly Hills, London 1980.

[11] Powers D.A., Xie Y., *Statistical Methods for Categorical Data Analysis*, Academic Press, San Diego 2000.

[12] Read, T.R.C., Cressie, N.A.C., *Goodness-Of-Fit Statistics for Discrete Multivariate Data*, Springer, New York 1988.

[13] Stanisz A., *Przystępny kurs statystyki*, tom II, Statsoft Polska, Kraków 2000.

[14] Vermunt J., *LEM: A General Program for the Analysis of Categorical Data*, Tilburg University, http://www.uvt.nl/faculteiten/fsw/organisatie/ departementen/mto/software2.html.

[15] *Wzrost optymizmu w przewidywaniach dotyczących rynku pracy*, CBOS, BS/89/2004, Warszawa, maj 2004.

LOG-LINEAR MODELS FOR ORDINAL VARIABLE

Summary

The main purpose of this article is to show log-linear models for ordinal variables. The log-linear models examine the relationship between two or more categorical variables in a contingency table. This article shows the various types of log-linear models i.e. uniform association model, row effect model, column effect model and RC model proposed by Goodman in 1979.

Mgr Iwona Kasprzyk jest doktorantem w Katedrze Statystyki w Akademii Ekonomicznej w Katowicach.