PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1097 — 2005
Statystyka ekonomiczna

Beata Zmyślona

ZASTOSOWANIE BAYESOWSKIEJ METODY ITERACYJNEGO DOPASOWANIA PROPORCJONALNEGO DO UZUPEŁNIANIA BRAKUJĄCYCH DANYCH

1. Wstęp

W badaniach ankietowych wykorzystuje się następujące techniki badawcze: ankiety, wywiady, spisy, rejestracje oraz inne narzędzia socjometryczne. Na etapie zbierania danych z wykorzystaniem wyżej wymienionych technik występuje problem braku odpowiedzi, wynikający z przeoczenia, niewiedzy bądź też odmowy udzielenia odpowiedzi przez respondenta na określone pytanie.

Brakujące dane utrudniają realizację kolejnych etapów badania, a mianowicie wstępnej obróbki danych (czyli edycji, kodowania danych), właściwej analizy statystycznej (m.in. estymacji i testowania hipotez statystycznych) czy też prezentacji i interpretacji wyników. Z teoretycznego punktu widzenia brakujące dane są problemem uniemożliwiającym uzyskanie reprezentatywności wyników pomimo wykorzystywania w analizach dużych, losowo dobranych prób.

W celu przeciwdziałania niedogodnościom związanym z występowaniem brakujących danych wykorzystuje się metody służące do predykcji brakujących danych (nazywane także metodami imputacji). Mimo że metody te mogą istotnie zmniejszać bądź też eliminować obciążenie estymatorów, mają jednak istotną wadę – zaniżają oceny wariancji estymatorów. Nie są to więc wskazane metody, jeżeli celem analizy jest konstrukcja przedziałów ufności bądź testowanie hipotez statystycznych w przypadku niepełnych danych (wyznaczone przedziały ufności oraz wartości statystyk testowych są obarczone błędem wynikającym z niedoszacowania bądź przeszacowania wariancji estymatorów) (por. [3, s. 128-130; 9, s. 11-18]; zob. także [4; 5; 6]).

Metodami umożliwiającymi eliminację zarówno obciążenia estymatora, jak i poprawę oceny wariancji tego estymatora są bayesowskie metody analizy niepełnych danych oparte na metodach Monte Carlo. W metodach bayesowskich brakujące dane są traktowane jak nie obserwowane realizacje pewnych zmiennych losowych o określonym warunkowym rozkładzie prawdopodobieństwa. Z tego rozkładu generowane są próbki pseudolosowe, które wykorzystuje się do analiz statystycznych. Tak więc w przypadku podejścia bayesowskiego uzupełnianie brakujących danych nie służy bezpośrednio do predykcji utraconej informacji, jest tylko etapem pośrednim umożliwiającym otrzymanie wyników „lepszych" ze statystycznego punktu widzenia (por. [2; 7; 8; 9; 10]).

W artykule przedstawiono przykład dotyczący badania zależności między zmiennymi kategorialnymi za pomocą modeli hierarchicznych w razie wystąpienia braku odpowiedzi. W omawianym przykładzie skupiono głównie uwagę na prezentacji metody generowania próbek pseudolosowych z warunkowego rozkładu brakujących danych za pomocą bayesowskiej metody iteracyjnego dopasowania proporcjonalnego.

2. Model log-liniowy

Przykład dotyczy wielowymiarowej analizy trzech zmiennych losowych kategorialnych. Zmienne Y_1, Y_3 przybierają wartości 0 lub 1, zmienna Y_2 przybiera wartości 0, 1 lub 2. Realizacje zmiennych Y_1, Y_2, Y_3 uzyskane dla n-elementowej próby mogą być bez straty informacji zredukowane do wektora (por. [1; 10; 11]):

$$\mathbf{x} = (x_{000}, x_{001}, x_{011}, \ldots, x_{121}). \tag{1}$$

Składowa x_{ijk} oznacza liczbę jednostek, dla których $Y_1 = i$, $Y_2 = j$ oraz $Y_3 = k$. Wektor $\mathbf{x}$ może być traktowany albo jako tablica kontyngencji, albo jako realizacja wektora losowego $\mathbf{X}$, mającego rozkład wielomianowy, co symbolicznie zapisuje się w następujący sposób:

$$\mathbf{X} \sim M(n, \boldsymbol{\theta}). \tag{2}$$

Wektor $\boldsymbol{\theta} = (\theta_{000}, \theta_{001}, \theta_{011}, \ldots, \theta_{121})$ jest wektorem nieznanych parametrów. Składowa θ_{ijk} oznacza prawdopodobieństwo, że $Y_1 = i$, $Y_2 = j$ oraz $Y_3 = k$.

Do wielowymiarowej analizy zależności stochastycznych między zmiennymi Y_1, Y_2, Y_3 wykorzystuje się modele log-liniowe. Modele log-liniowe nie zmieniają

postaci rozkładu wektora losowego **X**, wprowadzają natomiast pewne dodatkowe ograniczenia na składowe wektora $\boldsymbol{\theta}$. Niech

$$\eta_{ijk} = \log\theta_{ijk} \quad \text{dla} \quad i,k = 0,\ 1 \ \text{oraz} \ j = 0,\ 1,\ 2. \tag{3}$$

Model log-liniowy można przedstawić w następującej postaci:

$$\boldsymbol{\eta} = \mathbf{M}\boldsymbol{\lambda} \tag{4}$$

gdzie (por. [10, s. 290]):

- $\boldsymbol{\eta} = (\eta_{000}, \eta_{001}, \ldots, \eta_{121})$ oznacza D-wymiarowy wektor parametrów, gdzie D jest liczbą kombinacji poziomów zmiennych Y_1, Y_2, Y_3 (w omawianym przypadku D = 12),
- $\boldsymbol{\lambda}$ – r-wymiarowy wektor parametrów (w omawianym przypadku r = 8),
- **M** – stała macierz o wymiarze $D \times r$, składająca się z 1 lub –1.

Oprócz warunku sumowania się elementów θ_{ijk} wektora $\boldsymbol{\theta}$ do jedynki dodatkowo narzuca się warunek, żeby wektor $\boldsymbol{\eta} = \log\boldsymbol{\theta}$ leżał w liniowej podprzestrzeni rozpiętej na kolumnach macierzy **M**.

Dla uproszczenia zapisu oznacza się przez *I, J, K* odpowiednio pierwszą, drugą oraz trzecią zmienną, tzn. $Y_1 \equiv I$, $Y_2 \equiv J$ oraz $Y_3 \equiv K$.

Każda składowa η_{ijk} wektora $\boldsymbol{\eta}$ może być przestawiona jako następująca suma [10]:

$$\eta_{ijk} = \lambda_0 + \lambda_i^I + \lambda_j^J + \lambda_k^K + \lambda_{ij}^{IJ} + \lambda_{ik}^{IK} + \lambda_{jk}^{JK} + \lambda_{ijk}^{IJK}, \tag{5}$$

przy dodatkowym założeniu, że $\sum_i \lambda_i^I = \sum_j \lambda_j^J = \sum_{ik} \lambda_{ik}^{IK} = \ldots = \sum_{ijk} \lambda_{ijk}^{IJK} = 0$.

Parametry $\lambda_i^I, \lambda_j^J, \lambda_k^K$ są nazywane efektami głównymi, $\lambda_{ij}^{IJ}, \lambda_{ik}^{IK}, \lambda_{jk}^{JK}$ interakcjami drugiego rzędu, a λ_{ijk}^{IJK} interakcją trzeciego rzędu.

Składnik

$$\lambda_0 = -\log\left\{\sum_{i=0}^{1}\sum_{j=0}^{2}\sum_{k=0}^{1}\exp\left(\lambda_i^I + \lambda_j^J + \ldots + \lambda_{ijk}^{IJK}\right)\right\}$$

jest stałą wybraną w taki sposób, aby prawdopodobieństwa θ_{ijk} (dla wszystkich i, j, k) sumowały się do jedynki (por. np. [10; 11]).

Parametryzacja (5) nie wnosi żadnej dodatkowej informacji, jeśli na parametry λ nie zostaną narzucone dodatkowe warunki. Wyrażenie logarytmu prawdopodobieństwa θ_{ijk} za pomocą modelu log-liniowego pozwala na eliminację związków między zmiennymi przez przyjęcie, że określone składniki sumy (5) są równe zeru.

Na przykład jeżeli założy się, że wszystkie elementy zbiorów $\lambda^{IJ} = \{\lambda_{ij}^{IJ}\}$, $\lambda^{IK} = \{\lambda_{ik}^{IK}\}$, $\lambda^{JK} = \{\lambda_{jk}^{JK}\}$ oraz $\lambda^{IJK} = \{\lambda_{ijk}^{IJK}\}$ są równe zeru, to model log-liniowy nazywany modelem wzajemnej niezależności przybiera następującą postać:

$$\eta_{ijk} = \lambda_0 + \lambda_i^I + \lambda_j^J + \lambda_k^K.$$

Przyjęcie modelu wzajemnej niezależności jest równoznaczne z tym, że zmienne *I, J* oraz *K* są wzajemnie niezależne, czyli łączne prawdopodobieństwo $\theta_{ijk} = P(Y_1 = i, Y_2 = j, Y_3 = k)$ jest następującym iloczynem:

$$\theta_{ijk} = P(Y_1 = i)P(Y_2 = j)P(Y_3 = k) \text{ (dla } i, k = 0, 1; j = 0, 1, 2).$$

Innym często rozpatrywanym modelem log-liniowym jest model jednorodnej zależności. Model jednorodnej zależności ma taką własność, że na każdym poziomie jednej zmiennej istnieje ten sam rodzaj zależności między parą pozostałych zmiennych w sensie równości ilorazów szans, jednakże inną postać ma brzegowa tablica kontyngencji dla poszczególnych zmiennych [11]. Jest to równoznaczne z przyjęciem założenia, że $\lambda^{IJK} = 0$, co jest jednoznaczne z założeniem braku interakcji trójczynnikowej. Model jednorodnej zależności można przedstawić jako

$$\eta_{ijk} = \lambda_0 + \lambda_i^I + \lambda_j^J + \lambda_k^K + \lambda_{ij}^{IJ} + \lambda_{ik}^{IK} + \lambda_{jk}^{JK}. \quad (6)$$

W modelu jednorodnej zależności nie można podać dokładnego wzoru na prawdopodobieństwo θ_{ijk}, w związku z czym do jego oszacowania wykorzystuje się metody numeryczne, np. metodę iteracyjnego dopasowania proporcjonalnego.

3. Modele log-liniowe w przypadku niepełnych danych

Dodatkowe utrudnienie pojawia się, gdy w zbiorze danych występują brakujące obserwacje. Dla uproszczenia prezentacji w artykule rozpatruje się przypadek, w którym zmienne Y_1 oraz Y_2 są zawsze obserwowane, a Y_3 jest dla pewnych bada-

nych jednostek nie obserwowana. W takim przypadku elementy próby można podzielić na dwie części, oznaczone odpowiednio symbolami *A* oraz *B*. Do części *A* zalicza się jednostki, dla których wszystkie trzy zmienne są obserwowane, a do części *B* – takie, dla których tylko Y_1 oraz Y_2 są obserwowane. Składową x_{ijk} wektora (1) można przedstawić jako następującą sumę (por. [10]):

$$x_{ijk} = x_{ijk}^{A} + x_{ijk}^{B} \quad (\text{dla } i,\ k = 0,\ 1 \text{ oraz } j = 0,\ 1,\ 2). \tag{7}$$

Liczebności x_{ijk}^{B} nie są obserwowane, znane są jedynie brzegowe liczebności $x_{ij.}^{B} = x_{ij0}^{B} + x_{ij1}^{B}$ (dla $i = 0,\ 1$ oraz $j = 0,\ 1,\ 2$). Przez $\mathbf{y}_{obs} = \left\{x_{ij.}^{B}, x_{ijk}^{A}\right\}$ oznacza się obserwowane dane, a przez $\mathbf{y}_{br} = \left\{x_{ij0}^{B}, x_{ij1}^{B}\right\}$ – brakujące.

Wyznaczanie estymatorów parametrów modelu log-liniowego w przypadku braków odpowiedzi wymaga wyznaczania rozkładu *a posteriori* brakujących danych. Pobieranie prób z rozkładów brakujących danych bez konieczności wyznaczania go w sposób analityczny jest możliwe dzięki wykorzystaniu algorytmu Gibbsa (por. [2; 12]).

Każda iteracja algorytmu Gibbsa składa się z dwóch kroków. W pierwszym kroku wyznacza się następujące warunkowe rozkłady brakujących danych:

$$\begin{aligned} X_{ij0}^{B(t+1)} &\sim M\left(x_{ij.}, \frac{\theta_{ij0}^{(t)}}{\theta_{ij.}^{(t)}}\right) \quad \text{dla } i = 0,\ 1;\ j = 0,\ 1,\ 2, \\ X_{ij1}^{B(t+1)} &\sim M\left(x_{ij.}, \frac{\theta_{ij1}^{(t)}}{\theta_{ij.}^{(t)}}\right) \quad \text{dla } i = 0,\ 1;\ j = 0,\ 1,\ 2. \end{aligned} \tag{8}$$

Z rozkładów (8) losowane są liczby $x_{ijk}^{B(t+1)}$, którymi następnie uzupełniany jest zbiór danych, czyli obliczane są liczebności tabeli kontyngencji zgodnie ze wzorem (7).

W drugim kroku iteracji losuje się wartości parametrów $\boldsymbol{\theta}$ z jego rozkładu *a posteriori*. Wyznaczenie rozkładu *a posteriori* wektora parametrów $\boldsymbol{\theta}$ wymaga wyspecyfikowania rozkładu *a priori*. Przyjmuje się, że rozkładem *a priori* jest rozkład Dirichleta z wektorem parametrów $\boldsymbol{\alpha} = (\alpha_{000}, \alpha_{001}, \ldots, \alpha_{121})$. Przy tak wyspecyfikowanym rozkładzie *a priori* oraz przy założeniu, że wektor $\mathbf{X}$ ma rozkład wielomianowy, rozkładem *a posteriori* wektora $\boldsymbol{\theta}$ jest rozkład Dirichleta z wekto-

rem parametrów $\boldsymbol{\alpha}' = (\alpha_{000} + x_{000}, \alpha_{001} + x_{001}, \ldots, \alpha_{121} + x_{121})$, co w skrócie zapisuje się jako $\boldsymbol{\theta} \sim D(\boldsymbol{\alpha}')$.

4. Bayesowska metoda iteracyjnego dopasowania proporcjonalnego

Do wyznaczenia estymatorów parametrów modelu jednorodnej zależności wykorzystuje się niżej podaną bayesowską metodę iteracyjnego dopasowania proporcjonalnego ([10, s. 308, 309]; por. także [2; 11]). Wspomniana metoda służy także do osiągnięcia celu pośredniego w analizie niepełnych danych, a mianowicie do uzupełniania braków odpowiedzi.

Wartości $\boldsymbol{\theta}^{(t+1)}$ wektora $\boldsymbol{\theta}$ w $(t+1)$-ej iteracji otrzymuje się w podany niżej sposób w trzech etapach. W pierwszym etapie otrzymuje się

$$\theta_{ijk}^{(t+1/3)} = \theta_{ijk}^{(t+0/3)} \left(\frac{g_{ij.} / g_{...}}{\theta_{ij.}^{(t+0/3)}} \right) \text{ dla każdego } i, j, k . \tag{9}$$

Numer 3 w górnym indeksie oznacza, że wartości parametru θ_{ijk} w każdej iteracji są otrzymywane w trzech krokach. Wartości $g_{ij.}$ (dla wszystkich i oraz j) są niezależnymi realizacjami zmiennej losowej o standardowym rozkładzie gamma z parametrem kształtu

$$\alpha'_{ij.} = \sum_{k=0}^{1} (\alpha_{ijk} + x_{ijk})$$

oraz parametrem skali równym 1, a $g_{...} = \sum_{i=0}^{1} \sum_{j=0}^{2} g_{ij.}$.

W drugim etapie oblicza się następującą wartość:

$$\theta_{ijk}^{(t+2/3)} = \theta_{ijk}^{(t+1/3)} \left(\frac{g_{i.k} / g_{...}}{\theta_{i.k}^{(t+1/3)}} \right) \text{ dla każdego } i, j, k . \tag{10}$$

Wartości $g_{i.k}$ (dla wszystkich i oraz k) są niezależnymi realizacjami zmiennej losowej o standardowym rozkładzie gamma z parametrem kształtu

$$\alpha'_{i.k} = \sum_{j=0}^{2} \left(\alpha_{ijk} + x_{ijk} \right)$$

oraz parametrem skali równym 1, a $g_{...} = \sum_{i=0}^{1} \sum_{k=0}^{1} g_{i.k}$.

W trzecim etapie otrzymuje się

$$\theta_{ijk}^{(t+3/3)} = \theta_{ijk}^{(t+2/3)} \left(\frac{g_{.jk} / g_{...}}{\theta_{.jk}^{(t+2/3)}} \right) \text{ dla każdego } i, j, k . \tag{11}$$

Wartości $g_{.jk}$ (dla wszystkich j oraz k) są niezależnymi realizacjami zmiennej losowej o standardowym rozkładzie gamma z parametrem kształtu

$$\alpha'_{.jk} = \sum_{i=0}^{1} \left(\alpha_{ijk} + x_{ijk} \right)$$

oraz parametrem skali równym 1, a $g_{...} = \sum_{j=0}^{2} \sum_{k=0}^{1} g_{.jk}$.

W drugim kroku każdej iteracji oblicza się wartości parametrów $\boldsymbol{\theta}^{(t+1)}$, wykorzystując algorytm iteracyjnego dopasowania proporcjonalnego.

W wyniku kolejnych iteracji otrzymuje się stochastyczne ciągi $\left\{ \boldsymbol{\theta}^{(t)} : t = 0,\ 1,\ 2 \right\}$, które zbiegają do rozkładu Dirichleta z wektorem parametrów

$$\boldsymbol{\alpha}' = \left(\alpha_{000} + x_{000}, \alpha_{001} + x_{001}, \ldots, \alpha_{121} + x_{121} \right) \text{ dla każdego } i, j, k.$$

Po osiągnięciu zbieżności do rozkładu Dirichleta losuje się w m kolejnych iteracjach liczby $x_{ijk}^{B(l)}$ (dla l = 1, 2, ..., m), na podstawie których jest obliczanych m liczebności x_{ijk} dla każdego i, j, k. Liczebności te tworzą zbiory danych, na postawie których przeprowadzana jest dalsza analiza statystyczna.

Jeżeli do wnioskowania statystycznego wykorzystuje się metodę imputacji wielokrotnej, to przyjmuje się, że liczba uzupełnień brakujących danych m powinna wynosić od 3 do 10 w zależności od procentu informacji utraconych z powodu brakujących danych. Gdy zaś do wyznaczania estymatorów parametrów oraz wariancji tych estymatorów wykorzystuje się metody Monte Carlo, liczba m powinna wynosić co najmniej 1000 (przy czym zwiększanie liczby uzupełnień brakujących danych poprawia dokładność otrzymanych wyników). Od wybranej metody anali-

zy statystycznej zależy sposób wyznaczania estymatorów parametrów oraz sposób szacowania wariancji tych estymatorów (por. [2; 8; 9; 10]).

5. Przykład

W 1997 r. Bank Światowy przeprowadził badanie dotyczące sytuacji ekonomicznej gospodarstw domowych w krajach Europy Środkowo-Wschodniej. Dane uzyskane w wyniku tego badania przedstawiono w raporcie RAD Project „Poverty and Targeting of Social Assistance in Eastern Europe and Former Soviet Union". Badaniem objęto 16 051 gospodarstw domowych w Polsce. Zebrane informacje dotyczyły m.in. dochodów, wydatków na różne cele oraz wyposażenia gospodarstw domowych w dobra trwałego użytku.

Jednym z celów badania było ustalenie, czy istnieje związek między posiadaniem telefonu oraz samochodu a miejscem zamieszkania. Zebrane informacje dotyczyły tego, czy gospodarstwo domowe ma telefon, samochód oraz jakie jest jego miejsce zamieszkania (możliwe odpowiedzi: „stolica", „miasto", „wieś").

Odpowiedzi na pytanie, czy gospodarstwo domowe ma telefon oraz samochód, traktuje się jak realizacje binarnych zmiennych losowych Y_1 oraz Y_3 (odpowiedzi są kodowane w następujący sposób: odpowiedzi „nie" były zapisywane jako 0, odpowiedzi „tak" – jako 1). Odpowiedzi udzielane na trzecie pytanie, dotyczące

Tabela 1. Informacje o 14 007 gospodarstwach domowych dotyczące miejsca zamieszkania, posiadania samochodu oraz telefonu

	Gospodarstwo posiada samochód			Gospodarstwo nie posiada samochodu		
	Miejsce zamieszkania					
	stolica	miasto	wieś	stolica	miasto	wieś
Gospodarstwo posiada telefon	250	1780	354	279	1545	136
Gospodarstwo nie posiada telefonu	174	1494	1438	332	4028	2197

Źródło: RAD Project „Powerty and Targeting of Social Assistance in Eastern Europe and Former Soviet Union".

miejsca zamieszkania traktuje się jako realizacje zmiennej losowej Y_2, która przybiera trzy wartości: 0, 1, 2 (odpowiedzi są kodowane w następujący sposób: „stolica" – 0, „miasto" – 1, „wieś" – 2). Uzyskane informacje o 14 007 gospodarstwach domowych przedstawiono w tab. 1.

Od 2044 gospodarstw domowych nie uzyskano informacji o tym, czy gospodarstwo ma samochód. Informacje dotyczące miejsca zamieszkania oraz posiadania telefonu przez te gospodarstwa domowe przedstawiono w tab. 2.

W celu prezentacji zastosowania bayesowskiego algorytmu iteracyjnego dopasowania proporcjonalnego do uzupełniania brakujących danych przyjmuje się, że do badania zależności między zmiennymi kategorialnymi Y_1, Y_2, Y_3 wykorzystuje się model jednorodnej zależności.

Tabela 2. Informacje dla 2044 gospodarstw domowych dotyczących miejsca zamieszkania oraz posiadania telefonu

	Miejsce zamieszkania		
	stolica	miasto	wieś
Gospodarstwo posiada telefon	36	267	114
Gospodarstwo nie posiada telefonu	78	481	1068

Źródło: RAD Project „Powerty and Targeting of Social Assistance in Eastern Europe and Former Sowiet Union".

Wszystkie 16 051 gospodarstw domowych podzielono na dwie grupy oznaczone literami A oraz B. Do pierwszej grupy A zaliczono te gospodarstwa, o których uzyskano informacje dotyczące trzech badanych zmiennych (czyli 14 007 gospodarstw), a do grupy B – te, o których nie uzyskano informacji, czy gospodarstwo domowe ma telefon (czyli 2044 gospodarstw). Przedstawione w tab. 1 i 2 dane traktuje się jako realizacje wektora losowego $\mathbf{x} = (x_{000}, x_{001}, \ldots, x_{121})$ o rozkładzie wielomianowym. Ze względu na brakujące dane każdą składową wektora $\mathbf{x}$ można przedstawić jako sumę

$$x_{ijk} = x_{ijk}^{A} + x_{ijk}^{B} \text{ dla } i, k = 0, 1 \text{ oraz } j = 0, 1, 2,$$

gdzie x_{ijk}^{A} oraz x_{ijk}^{B} są liczbami gospodarstw domowych, dla których $Y_1 = i$, $Y_2 = j$, $Y_3 = k$, zakwalifikowanych odpowiednio do grupy A oraz grupy B.

Wartości x_{ijk}^{B} są nieznane, znane są jedynie następujące brzegowe sumy:

$$x_{ij.}^{B} = \sum_{k=0}^{1} x_{ijk}^{B} \text{ dla wszystkich } i, k.$$

Na podstawie danych przedstawionych w tab. 1 i 2 uzyskano następujące liczebności:

$$x^{A}_{000}=332,\ x^{A}_{010}=4028,\ x^{A}_{020}=2197,\ x^{A}_{001}=174,\ x^{A}_{011}=1494,\ x^{A}_{021}=1438,$$
$$x^{A}_{100}=279,\ x^{A}_{110}=1545,\ x^{A}_{120}=136,\ x^{A}_{101}=250,\ x^{A}_{111}=1780 \text{ oraz } x^{A}_{121}=354,$$

oraz następujące sumy brzegowe:

$$x^{B}_{00.}=78,\ x^{B}_{01.}=481,\ x^{B}_{02.}=1068,\ x^{B}_{10.}=36,\ x^{B}_{11.}=267 \text{ oraz } x^{B}_{12.}=114.$$

Liczebności te stanowią obserwowane dane. Każda z brzegowych sum $x^{B}_{ij.}$ składa się z dwóch składników: x^{B}_{ij0} oraz x^{B}_{ij1}. Składniki te są nieznane i traktowane jak nieobserwowane realizacje zmiennych losowych o rozkładach wielomianowych wyrażonych wzorem (8). Rozkłady (8) są określane jako rozkłady brakujących danych. Nieznane wielkości składników sumy generowane są z tych rozkładów za pomocą opisanej wyżej bayesowskiej metody iteracyjnego dopasowania proporcjonalnego.

W analizie przyjęto jako rozkład *a priori* nieinformacyjny rozkład Dirichleta z wektorem parametrów $\boldsymbol{\alpha}=(1,1,...,1)$. Za startowe wartości parametrów przyjęto $\theta^{(t+0/3)}_{ijk}=\frac{1}{12}$ dla $i, k = 0, 1$ oraz $j = 0, 1, 2$.

Po przeprowadzeniu 1000 wstępnych iteracji uznano, że została osiągnięta zbieżność do rozkładu *a posteriori* wektora parametrów $\boldsymbol{\theta}$. W kolejnych iteracjach wygenerowano liczebności, które wykorzystano do wyznaczenia liczebności tabeli kontyngencji. Liczebności te wyznaczono zgodnie ze wzorem (7). Przykładowych dziesięć liczebności uzyskanych po uzupełnieniu brakujących danych przedstawiono w tab. 3.

Na podstawie uzupełnionych zbiorów danych (w analizowanym przykładzie są to tabele kontyngencji) możliwa jest dalsza analiza, czyli np. estymacja parametrów modelu log-liniowego czy testowanie hipotez dotyczących zależności między badanymi zmiennymi kategorialnymi.

Tabela 3. Suma liczebności z części *A* oraz *B* po uzupełnieniu brakujących obserwacji

Nr zbioru danych	$l = 1$	$l = 2$	$l = 3$	$l = 4$	$l = 5$	$l = 6$	$l = 7$	$l = 8$	$l = 9$	$l = 10$
$x_{000}^{(l)}$	367	372	378	370	375	370	366	377	371	364
$x_{010}^{(l)}$	4270	4289	4284	4286	4281	4269	4251	4263	4293	4295
$x_{020}^{(l)}$	2756	2782	2767	2815	2787	2748	2731	2782	2750	2778
$x_{001}^{(l)}$	217	212	206	214	209	214	218	207	213	220
$x_{011}^{(l)}$	1733	1714	1719	1717	1722	1734	1752	1740	1710	1708
$x_{021}^{(l)}$	1947	1921	1936	1888	1916	1955	1972	1921	1953	1925
$x_{100}^{(l)}$	293	294	299	295	297	294	288	298	297	294
$x_{110}^{(l)}$	1659	1679	1661	1668	1652	1658	1647	1662	1667	1677
$x_{120}^{(l)}$	192	187	174	194	184	191	183	192	189	182
$x_{101}^{(l)}$	272	271	266	270	268	271	277	267	268	271
$x_{111}^{(l)}$	1933	1913	1931	1924	1940	1934	1945	1930	1925	1915
$x_{121}^{(l)}$	412	417	430	410	420	413	421	412	415	422

Źródło: obliczenia własne.

6. Podsumowanie

W bayesowskim podejściu do analizy niepełnych danych niezbędne jest generowanie próbek pseudolosowych z warunkowych rozkładów brakujących danych. Głównym celem stosowania podejścia bayesowskiego jest eliminacja zarówno obciążenia wyznaczanych estymatorów, jak i błędu wynikającego z niedoszacowania bądź przeszacowania wariancji tych estymatorów, powstałego na skutek braku odpowiedzi.

Bayesowska metoda iteracyjnego dopasowania proporcjonalnego jest skutecznym narzędziem używanym w przypadku stosowania modeli jednorodnej zależności do badania związków między zmiennymi kategorialnymi.

W algorytmie przyjęto nieinformacyjny rozkład *a priori*. Niemniej warto wspomnieć o możliwości zastosowania informacyjnych rozkładów *a priori*, umożliwiających wykorzystanie dodatkowej wiedzy spoza próby dotyczącej badanej populacji. Możliwość wykorzystania dodatkowych informacji, gdy informacje z próby są niepełne, jest niewątpliwie bardzo istotną zaletą metod bayesowskich.

Literatura

[1] Agrestii A., *Categorical Data Analysis*, John Wiley & Sons, New York 1990.

[2] Berger J.O., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York 1985.

[3] Fairclough D.L., *Design and Analysis of Quality of Life Studies in Clinical Trials*, Chapman Hall/CRC, Washington 2002.

[4] Rao J.N.K., *On Variance Estimation with Imputed Survey Data*, JASA 91 (1996), s. 499-506.

[5] Rao J.N.K., Shao J., *Jackknife Variance Estimation with Survey Data under Hot Deck Imputation*, Biometrika 79 (1992), s. 811-822.

[6] Rubin D., Schenker N., *Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse*, JASA 81 (1986), s. 366-380.

[7] Rubin D., *Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician*, The Annals of Statistics 12 (1984), s. 1151-1172.

[8] Rubin D., *Multiple Imputation after 18+Years*, JASA 91 (1996), s. 473-520.

[9] Rubin D., *Multiple Imputation for Nonresponse in Surveys*, John Willey & Sons, New York 1987.

[10] Schafer J., *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York 2000.

[11] Sobolewski M., *Analiza logarytmiczno-liniowa w eksploracji wielowymiarowych tabel kontyngencji*, [w:] *Metodologia pomiaru jakości życia*. Materiały ogólnopolskiej konferencji, red. W. Ostasiewicz, Karpacz: 22, 23 listopada 2001.

[12] Tanner M., Wong W., *The Calculation of Posterior Distributions by Data Augmentation*, JASA 82 (1987), s. 528-540.

THE APPLICATION OF BAYESIAN ITERATIVE PROPORTIONAL FITTING ALGORITHM TO IMPUTATION MISSING DATA

Summary

The study of associations among categorical features requires using some techniques, for example a loglinear model. The Bayesian iterative proportional fitting algorithm (Bayesian IPF) is the simulation Monte Carlo technique of estimation of loglinear model parameters in case of incomplete data sets. In this technique we create pseudorandom draws from the posterior distribution of parameters and from the conditional distribution of missing values. The main aim of using this approach is to improve statistical inference through elimination of estimator bias and to correct estimation of standard errors.

In this paper we present the theoretical background of Bayesian IPF and its application to impute missing data through generating values from conditional distribution.