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

Beata Zmyślona

ZASTOSOWANIE MODELI HIERARCHICZNYCH W BAYESOWSKIM WNIOSKOWANIU STATYSTYCZNYM W PRZYPADKU NIEPEŁNYCH DANYCH

1. Wstęp

W badaniach społeczno-ekonomicznych występuje często problem brakujących danych. Na przykład w badaniach przeprowadzanych za pomocą eksperymentów niektóre pomiary nie są dokonywane, w przypadku badań ankietowych występuje zjawisko tzw. braku odpowiedzi (respondenci nie udzielają pewnych informacji). W innych sytuacjach nie wszystkie dane są uzyskiwane z powodu ograniczeń czasowych, technicznych bądź finansowych.

Brakujące dane utrudniają realizację kolejnych etapów badania, m.in. wstępną obróbkę danych, właściwą analizę statystyczną. Wnioskowanie statystyczne na bazie niekompletnego zbioru danych nie gwarantuje reprezentatywności uzyskanych wyników. W celu przeciwdziałania niedogodnościom związanym z występowaniem brakujących danych wykorzystuje się różne podejścia. W artykule przedstawiono ujęcie bayesowskie do analizy niepełnych danych uwzględniające mechanizm powstawania brakujących danych. W podejściu uwzględniającym mechanizm powstawania brakujących danych niezbędne jest oszacowanie prawdopodobieństwa udzielenia odpowiedzi. Prawdopodobieństwo to szacowane jest albo metodami klasycznymi, albo metodami bayesowskimi.

Wartości p zmiennych dla n-elementowej próby prostej traktowane są jak realizacje macierzy losowej:

$$\mathbf{Y}_{n\times p} = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1p} \\ Y_{21} & Y_{22} & \dots & Y_{2p} \\ \vdots & \vdots & & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{np} \end{bmatrix}. \tag{1}$$

Każdy wiersz macierzy $\mathbf{Y}$ może być zapisany w następujący sposób:

$$\mathbf{Y}_i = \left[\mathbf{Y}_{i(\text{obs})} \quad \mathbf{Y}_{i(\text{br})}\right] \text{ (dla } i = 1, 2, \ldots, n\text{)}. \tag{2}$$

Pierwsza składowa wektora (2) odnosi się do obserwowanej części danych w i-tym wierszu, natomiast druga składowa – do nieobserwowanej części danych.

Przyjmuje się założenie, że wiersze macierzy (1) są statystycznie niezależne i o identycznych rozkładach. Dystrybuanta rozkładu prawdopodobieństwa macierzy losowej $\mathbf{Y}$ zależy od wektora nieznanych parametrów $\boldsymbol{\theta}$, co symbolicznie zapisujemy w następujący sposób: $F_{\mathbf{y}}(\mathbf{y}|\theta)$. Bayesowskie podejście wymaga określenia rozkładu *a priori* $p(\theta)$ wektora parametrów θ. Dystrybuantę brzegowego rozkładu wektora $\mathbf{Y}_i$ oznacza się przez $F_{\mathbf{Y}_i}(\mathbf{y}_i|\theta)$. Przy założeniu niezależności wektorów losowych $\mathbf{Y}_1$, $\mathbf{Y}_2$, ..., $\mathbf{Y}_n$ dystrybuantę rozkładu prawdopodobieństwa macierzy $\mathbf{Y}$ można przedstawić jako:

$$F_{\mathbf{Y}}(\mathbf{y}|\theta) = \prod_{i=1}^{n} F_{\mathbf{Y}_i}(\mathbf{y}_i|\theta).$$

Celem wnioskowania statystycznego jest znalezienie rozkładu *a posteriori* $p(\theta|\mathbf{y})$ wektora parametrów θ.

2. Mechanizm powstawania brakujących danych

W metodach wnioskowania statystycznego w przypadku występowania w zbiorze brakujących obserwacji w sposób jawny bądź niejawny wprowadza się założenia dotyczące mechanizmu powstawania brakujących danych. Zasadniczo metody analizy niepełnych danych można podzielić na takie, które uwzględniają mechanizm powstawania brakujących danych, oraz takie, które tego mechanizmu nie uwzględniają.

Rozpatrywana jest macierz wskaźników odpowiedzi (por. [6])

$$\mathbf{R}_{n\times p} = \begin{bmatrix} R_{11} & R_{12} & \ldots & R_{1p} \\ R_{21} & R_{22} & \ldots & R_{2p} \\ \vdots & \vdots & & \vdots \\ R_{n1} & R_{n2} & \ldots & R_{np} \end{bmatrix} \tag{3}$$

identyfikująca macierz (1). Przez $\mathbf{R}_i = \left[R_{i1} \quad R_{i2} \quad \ldots \quad R_{ip}\right]$ oznacza się i-ty wiersz macierzy (3). Mechanizm powstawania brakujących danych jest zdeterminowany warunkowym rozkładem prawdopodobieństwa wektora $\mathbf{R}_i$.

W bayesowskich metodach analizy niepełnych danych wyróżnia się trzy typy założeń dotyczących mechanizmu powstawania brakujących danych, które definiuje się za pomocą warunkowego rozkładu prawdopodobieństwa wektora $\mathbf{R}_i$ (por. np. [6; 7]).

Definicja 1

Nieobserwowane realizacje składowych wektora losowego $\mathbf{Y}_i$ są typu *MCAR* (skrót od angielskiej nazwy *missing completely at random*) w przypadku, gdy warunkowy rozkład prawdopodobieństwa wektora losowego $\mathbf{R}_i$ można przedstawić jako:

$$p(\mathbf{r}_i|\mathbf{y}_{i(\text{obs})}, \mathbf{y}_{i(\text{br})}, \xi) = p(\mathbf{r}_i|\xi) \text{ (dla } i = 1, 2, \ldots, n). \tag{4}$$

Oznacza to, że warunkowy rozkład wektora losowego $\mathbf{R}_i$ zależy tylko od wektora nieznanych parametrów ξ i nie zależy od wektora $\mathbf{Y}_i$.

Nieobserwowane realizacje składowych wektora $\mathbf{Y}_i$ w przypadku spełnienia tego założenia są traktowane jako podpróba losowa wszystkich obserwacji z próby. W badaniach społeczno-ekonomicznych zasadne jest przyjęcie założenia *MCAR*, jeżeli nie uzyskano danych od części jednostek z próby ze względu na przeoczenie lub jeżeli pewne informacje zostały nieczytelnie zapisane i w późniejszej analizie traktowane są jako brakujące dane.

Definicja 2

Nieobserwowane realizacje składowych wektora losowego $\mathbf{Y}_i$ są typu *MAR* (skrót od angielskiej nazwy *missing at random*), jeżeli warunkowy rozkład prawdopodobieństwa wektora losowego $\mathbf{R}_i$ zależy od nieznanego wektora parametrów ξ oraz tylko od obserwowanych składowych wektora $\mathbf{Y}_i$. W tym przypadku rozkład ten można przedstawić w następujący sposób:

$$p(\mathbf{r}_i|\mathbf{y}_{i(\text{obs})}, \mathbf{y}_{i(\text{br})}, \xi) = p(\mathbf{r}_i|\mathbf{y}_{i(\text{obs})}, \xi) \text{ (dla } i = 1, 2, \ldots, n). \tag{5}$$

Założenie *MAR* jest stosunkowo trudne do spełnienia w przypadku badań ankietowych, w przeciwieństwie do eksperymentów, w których badacz ma większą kontrolę nad wyborem jednostek, dla których realizacje pewnych zmiennych będą nieobserwowane.

Definicja 3

Nieobserwowane realizacje składowych wektora losowego $\mathbf{Y}_i$ są typu *NMAR* (skrót od angielskiej nazwy: *not missing at random*), jeżeli warunkowy rozkład prawdopodobieństwa wektora losowego $\mathbf{R}_i$ zależy od nieznanego wektora parametrów ξ, a także od obserwowanych i nieobserwowanych składowych wektora $\mathbf{Y}_i$, co symbolicznie można zapisać w następujący sposób:

$$p(\mathbf{r}_i|\mathbf{y}_{i(\text{obs})}, \mathbf{y}_{i(\text{br})}, \xi) \text{ (dla } i = 1, 2, \ldots, n). \tag{6}$$

Pomocna w ustaleniu założeń odnośnie do warunkowego rozkładu prawdopodobieństwa wektorów losowych $\mathbf{R}_1$, $\mathbf{R}_2$, ..., $\mathbf{R}_n$ okazuje się wiedza *a priori* badacza dotycząca przyczyn nieuzyskania pewnych informacji o badanych jednostkach.

Założenia dotyczące mechanizmu powstawania brakujących danych mają wpływ na postać rozkładu *a posteriori* wektora parametrów θ, na którym bazuje bayesowskie wnioskowanie statystyczne. Łączny rozkład *a posteriori* parametrów θ oraz ξ wyraża się następującym wzorem (por. np. [6]):

$$p(\theta, \xi|\mathbf{y}_{obs}, \mathbf{y}_{br}, \mathbf{r}) = \frac{p(\mathbf{r}, \mathbf{y}_{obs}, \mathbf{y}_{br}|\theta, \xi)p(\theta, \xi)}{\iint p(\mathbf{r}, \mathbf{y}_{obs}, \mathbf{y}_{br}|\theta, \xi)p(\theta, \xi)d\theta d\xi} = \frac{p(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{br}, \xi)p(\mathbf{y}_{obs}, \mathbf{y}_{br}|\theta)p(\theta, \xi)}{\iint p(\mathbf{r}|\mathbf{y}_{obs}, \mathbf{y}_{br}, \xi)p(\mathbf{y}_{obs}, \mathbf{y}_{br}|\theta)p(\theta, \xi)d\theta d\xi}. \quad (7)$$

Jeżeli w analizie niepełnych danych jest uwzględniany mechanizm powstawania brakujących danych, to wnioskowanie statystyczne bazuje na rozkładzie *a posteriori* (7).

3. Szacowanie prawdopodobieństwa uzyskania odpowiedzi

Jak już wspomniano na wstępie, w metodach wnioskowania statystycznego uwzględniającego mechanizm powstawania brakujących danych niezbędne jest oszacowanie tzw. prawdopodobieństwa uzyskania odpowiedzi.

W celu uproszczenia rozważań przyjmuje się, że liczba zmiennych p jest równa jeden. W takim przypadku zmienne dla n-elementowej próby można zapisać jako następujący wektor $\mathbf{Y} = [Y_1\ Y_2\ \ldots\ Y_n]^T$. Realizacje zmiennej Y (dla $i = 1, 2, \ldots, n$) są nieznane dla pewnej części elementów z próby. Wektor $\mathbf{R} = [R_1\ R_2\ \ldots\ R_n]^T$ jest wektorem wskaźników odpowiedzi. Składowe tego wektora są zmiennymi losowymi o rozkładzie zero-jedynkowym z prawdopodobieństwem sukcesu (por. [1; 2]):

$$\pi_i = P(R_i = 1). \quad (8)$$

Prawdopodobieństwo (8) w analizie niepełnych danych jest nazywane prawdopodobieństwem uzyskania odpowiedzi lub, w skrócie, prawdopodobieństwem odpowiedzi. W klasycznym podejściu prawdopodobieństwo odpowiedzi jest szacowane za pomocą modelu regresji, np. modelu probitowego, logitowego lub log-log-liniowego.

W bayesowskim wnioskowaniu przyjmuje się, że parametr π_i jest zmienną losową. Jednostki z próby są dzielone na K odrębnych klas o liczebnościach n_k takich, że $n = \sum_{k=1}^{K} n_k$. W obrębie każdej z klas są szacowane prawdopodobieństwa

uzyskania odpowiedzi π_k (dla $k = 1, 2, \ldots, K$). Przez A_k oznacza się zbiór jednostek należących do k-tej klasy. Liczbę uzyskanych odpowiedzi w k-tej klasie oznacza się przez $\sum_{i \in A_k} r_{ik} = r_k$. Klasy wyznaczane są na podstawie wartości pewnych zmiennych pomocniczych. Wektor odpowiedzi w poszczególnych klasach $[r_1 \; r_2 \; \ldots \; r_K]$ ma rozkład wielomianowy z wektorem parametrów $\pi = [\pi_1 \; \pi_2 \; \ldots \; \pi_K]$, czyli z wektorem prawdopodobieństw odpowiedzi.

Bayesowska estymacja polega na wyznaczaniu charakterystyk rozkładu *a posteriori* wektora π. W zależności od dostępności dodatkowych informacji pochodzących spoza próby przyjmuje się informacyjne bądź nieinformacyjne rozkłady *a priori* (por. np. [3]).

Przykładem rozkładu *a priori* parametru $\pi = [\pi_1 \; \pi_2 \; \ldots \; \pi_K]$ jest wielowymiarowy rozkład beta (czyli rozkład Dirichleta) z wektorem nieznanych parametrów $\xi = [\xi_1 \; \xi_2 \; \ldots \; \xi_K]$ z następującą funkcją gęstości:

$$p(\pi|\xi) = \frac{\Gamma(\xi_1 + \xi_2 + \ldots + \xi_K)}{\Gamma(\xi_1)\Gamma(\xi_2)\ldots\Gamma(\xi_K)} \pi_1^{\xi_1 - 1} \pi_2^{\xi_2 - 1} \ldots \pi_K^{\xi_K - 1}.$$

Rozkładem *a posteriori* jest również wielowymiarowy rozkład beta z wektorem parametrów $[\xi_1 + r_1 \; \xi_2 + r_2 \; \ldots \; \xi_K + r_K]$ o następującej funkcji gęstości

$$P(\pi|\xi) = \begin{cases} \dfrac{\Gamma(\xi_1 + \ldots + \xi_K + r_1 + \ldots + r_K)}{\Gamma(\xi_1 + r_1)\ldots\Gamma(\xi_K + r_K)} \pi_1^{\varepsilon_1 + r_1 - 1} \ldots \pi_K^{\varepsilon_K + r_K - 1}, & \text{jeśli } \sum_{i=1}^{K} \pi_i = 1, \\ 0, & \text{pozostałe.} \end{cases}$$

Jeżeli rozkład *a priori* wektora parametrów π jest rozkładem niesprzężonym, to do estymacji bayesowskiej wykorzystuje się metody numeryczne, np. metody Monte Carlo. Bayesowskimi estymatorami $\hat{\pi} = [\hat{\pi}_1 \; \hat{\pi}_2 \; \ldots \; \hat{\pi}_K]$ wektora prawdopodobieństw odpowiedzi są charakterystyki rozkładów *a posteriori* $p(\pi|\mathbf{r})$. Pomocna w ustaleniu rozkładu *a priori* wektora parametrów π jest wiedza spoza próby dotycząca przyczyn nieudzielania przez respondentów odpowiedzi.

4. Zastosowanie modeli hierarchicznych uwzględniających mechanizm powstawania brakujących danych

Ankiety wykorzystywane w badaniach społeczno-ekonomicznych zawierają m.in. pytania, na które respondenci z różnych przyczyn nie chcą udzielać odpowiedzi, np. pytania związane z nałogami, poglądami politycznymi, światopoglądem, popełnionymi przestępstwami. W takich sytuacjach do analizy niepełnych danych

przyjmuje się założenie *NMAR* dotyczące mechanizmu generowania brakujących danych. Do wnioskowania statystycznego o parametrach opisowych populacji w przypadku przyjęcia założenia *NMAR* w podejściu bayesowskim wykorzystuje się modele hierarchiczne (por. [2, s. 241-251]). Poniżej zostanie opisany hipotetyczny przykład wykorzystania takiego modelu do szacowania odsetka studentów ściągających na egzaminach.

Przyjęcie założenia *NMAR* oznacza, że fakt nieudzielenia odpowiedzi przez *i*-tego studenta jest powiązany z badaną zmienną Y_i, która przyjmuje wartość jeden, jeśli student ściąga, oraz zero – w przeciwnym wypadku. Wiadomo *a priori*, że skłonność do nieudzielania tego typu informacji jest zróżnicowana w zależności od płci oraz rodzaju studiów. Zatem badani studenci (dwa tysiące osób) dzieleni są na cztery grupy ze względu na płeć oraz rodzaj studiów (studia dzienne, studia zaoczne). Numer grupy indeksowany będzie przez k (dla $k = 1, 2, 3, 4$). Wszystkich studentów podzielono na cztery grupy: w pierwszej grupie znajdują się studentki studiów zaocznych, w drugiej – studentki studiów dziennych, w trzeciej grupie studenci studiów zaocznych, a w czwartej – studenci studiów dziennych. Otrzymane liczebności oraz liczby brakujących odpowiedzi z uwzględnieniem podziału na grupy zostały przedstawione w tab. 1. Liczebności studentów w k-tej grupie oznaczane są przez n_k (w omawianym przykładzie $n_1 = 400$, a pozostałe grupy $n_2 = n_3 = n_4 = 500$). Dla ustalonego k zmienne Y_{ik} $(i = 1, 2, \ldots, n_k)$ mają rozkład zero-jedynkowy z prawdopodobieństwem sukcesu p_k.

Tabela 1. Liczebności studentów z uwzględnieniem podziału na płeć oraz rodzaj studiów

Wyszczególnienie	Osoby ściągające	Osoby nieściągające	Brak odpowiedzi	Suma
Grupa 1	180	160	60	400
Grupa 2	105	310	85	500
Grupa 3	230	180	90	500
Grupa 4	120	190	190	500

Źródło: obliczenia własne.

W modelach hierarchicznych dotyczących analizy niepełnych danych niezbędne jest określenie prawdopodobieństwa uzyskania oraz nieuzyskania odpowiedzi. Wprowadza się następujące oznaczenia [2, s. 242]:

$$\begin{aligned} \pi_{k0} &= P\left(R_{ik} = 1 \middle| Y_{ik} = 0\right) \text{ dla } i = 1, 2, \ldots, n_k \text{ oraz } k = 1, 2, 3, 4, \\ \pi_{k1} &= P\left(R_{ik} = 1 \middle| Y_{ik} = 1\right) \text{ dla } i = 1, 2, \ldots, n_k \text{ oraz } k = 1, 2, 3, 4. \end{aligned} \tag{9}$$

Prawdopodobieństwa $1 - \pi_{k0}$ oraz $1 - \pi_{k1}$ są nazywane prawdopodobieństwami nieudzielenia odpowiedzi pod warunkiem, że odpowiednio student nie ściąga oraz ściąga na egzaminach. Prawdopodobieństwo nieudzielenia odpowiedzi przez *i*-tego studenta należącego do k-tej grupy można wyrazić za pomocą następującego wzoru:

$$P(R_{ik}=0)=P(R_{ik}=0|Y_{ik}=0)P(Y_{ik}=0)+P(R_{ik}=0|Y_{ik}=1)P(Y_{ik}=1)= \\ =(1-\pi_{k0})(1-p_k)+(1-\pi_{k1})p_k. \tag{10}$$

Przez t_k, s_k oraz u_k oznacza się odpowiednio liczbę studentów w k-tej grupie, którzy odpowiedzieli na postawione pytanie i ściągają na egzaminie, liczbę studentów, którzy odpowiedzieli i nie ściągają na egzaminie oraz liczbę studentów, którzy nie udzielili odpowiedzi na postawione pytanie. Symbolem v_k oznacza się nieznaną liczbę studentów, którzy nie udzielili odpowiedzi i ściągają na egzaminie.

Nieznane wielkości p_k, π_{k0}, π_{k1} traktowane są jak zmienne losowe, dla których przyjmuje się pewne rozkłady prawdopodobieństwa. Wartość v_k traktuje się jak realizację zmiennej losowej V_k. Dla parametrów p_k, π_{k0}, π_{k1} (dla $k=1, 2, 3, 4$) przyjmuje się następujące rozkłady *a priori* (por. np. [3]):

$$\begin{aligned} p_k &\sim Beta(1,1), \\ \pi_{k0} &\sim Beta(\alpha_{k0},\beta_{k0}), \\ \pi_{k1} &\sim Beta(\alpha_{k1},\beta_{k1}), \end{aligned} \tag{11}$$

czyli nieinformacyjny rozkład parametru p_k oraz informacyjne rozkłady pozostałych parametrów. Rozkłady *a priori* wyznaczane są na bazie informacji pochodzących spoza próby dotyczących przyczyn nieudzielania przez respondentów odpowiedzi na określone pytania. Takie dodatkowe informacje pochodzą albo z innych badań dotyczących podobnych zbiorowości, poprzednich badań dotyczących badanej zbiorowości lub wykorzystywana jest wiedza ekspertów zajmujących się dziedziną związaną z zagadnieniem, którego dotyczy badanie. W omawianym przykładzie przyjęto następujące osiem rozkładów *a priori*:

$$\begin{gathered} \pi_{10} \sim Beta(17{,}1;\,0{,}9),\ \pi_{20} \sim Beta(6{,}7;\,0{,}13), \\ \pi_{30} \sim Beta(31{,}5;\,3{,}5),\ \pi_{40} \sim Beta(42{,}5;\,7{,}5), \\ \pi_{11} \sim Beta(44{,}1;\,53{,}9),\ \pi_{21} \sim Beta(57;\,38), \\ \pi_{31} \sim Beta(50{,}4;\,12{,}6),\ \pi_{41} \sim Beta(55{,}5;\,18{,}5). \end{gathered} \tag{12}$$

Funkcje gęstości dla rozkładów (12) zostały przedstawione na rys. 1.

Wartość v_k jest traktowana jak nieznana realizacja zmiennej losowej V_k o rozkładzie dwumianowym:

$$V_k \sim B(u_k, w_k), \tag{13}$$

gdzie $w_k=(1-\pi_{k1})p_k/\left((1-\pi_{k0})(1-p_k)+(1-\pi_{k1})p_k\right)$ jest prawdopodobieństwem sukcesu dla rozkładu (13). Rozkład (13) jest nazywany warunkowym rozkładem brakujących danych.

Dla tak wyspecyfikowanych rozkładów (11) oraz (13) otrzymuje się następujące warunkowe rozkłady *a posteriori* (por. np. [2; 3]):

$$\begin{aligned} p_k|v_k &\sim Beta(v_k + s_k + 1; t_k + 1 + u_k - v_k), \\ \pi_{k0}|v_k &\sim Beta(t_k + \alpha_{k0}; u_k - v_k + \beta_{k0}), \\ \pi_{k1}|v_k &\sim Beta(s_k + \alpha_{k1}; v_k + \beta_{k1}). \end{aligned} \tag{14}$$

Do bayesowskiej estymacji parametru wykorzystuje się algorytm Gibbsa. Każda iteracja algorytmu polega na losowaniu z warunkowych rozkładów (13) oraz (14). Po 50 000 iteracji przyjmuje się, że osiągnięta została zbieżność do brzegowych rozkładów *a posteriori* parametrów p_k, π_{k0}, π_{k1} oraz do rozkładu *a posteriori* brakujących danych V_k.

Następnie na bazie wartości wylosowanych z rozkładów brzegowych w kolejnych iteracjach wyznacza się charakterystyki brzegowych rozkładów *a posteriori*, które są estymatorami bayesowskimi. Charakterystyki rozkładów *a posteriori* wyznaczane są za pomocą programu WINBUGS.

Charakterystyki brzegowych rozkładów a posteriori parametrów p_k, π_{k0}, π_{k1} oraz rozkładu *a posteriori* brakujących danych przedstawiono w tab. 2-4.

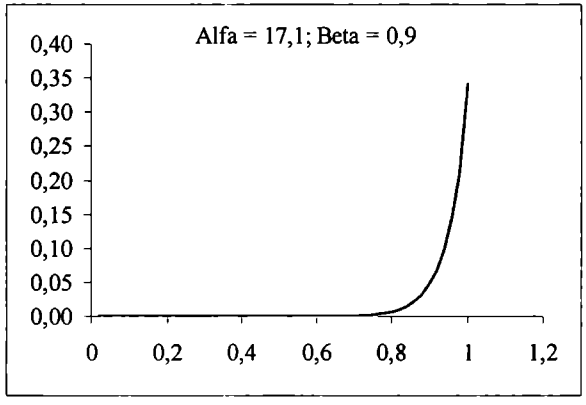
Wartość estymatora odsetka osób ściągających na egzaminach w badanej populacji studentów oszacowanego za pomocą modelu hierarchicznego wynosi 0,4818. Bayesowski obszar wiarygodności na poziomie ufności 0,95 wynosi (0,4463; 0,5201). Wartość estymatora odsetka osób ściągających wyznaczona tylko na podstawie obserwowanych danych (por. tab. 1) wynosi 0,4032.

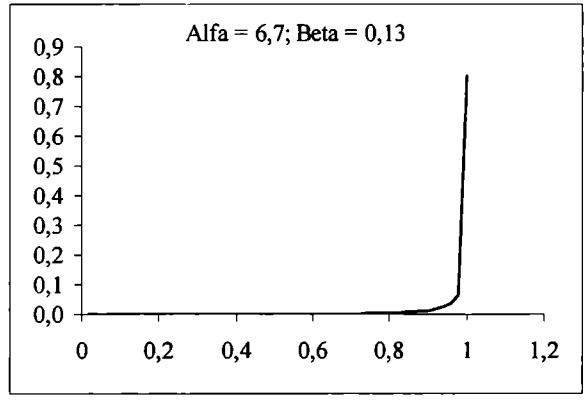
Zastosowanie podejścia bayesowskiego uwzględniającego mechanizm powstawania brakujących danych umożliwia zatem zmniejszenie obciążenia estymatora parametru, wynikające z faktu ukrywania przez niektórych studentów informacji dotyczących ściągania na egzaminie.

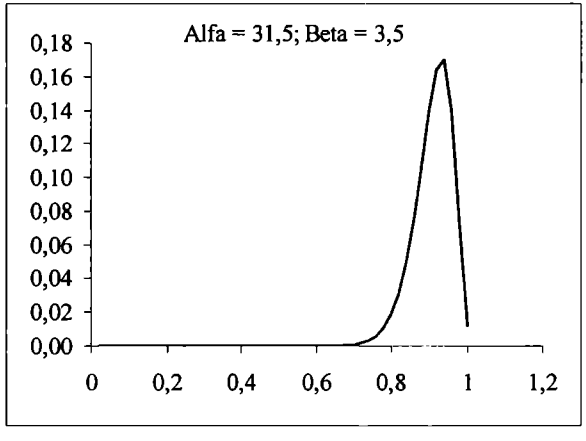
5. Zakończenie

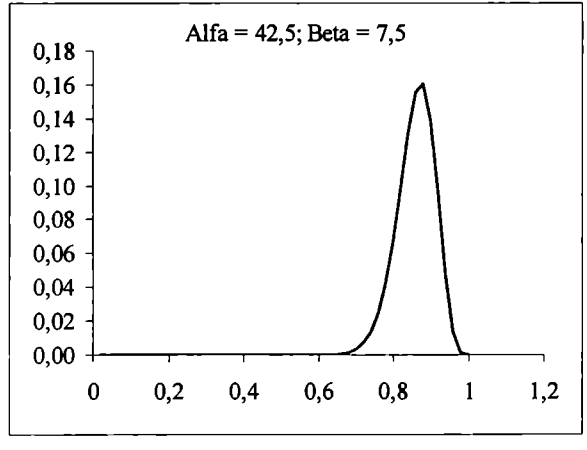
Wykorzystanie podejścia bayesowskiego we wnioskowaniu statystycznym w przypadku niepełnych danych umożliwia wykorzystanie dodatkowej wiedzy dotyczącej badanych jednostek, dla których występują brakujące dane, co jest niewątpliwie przewagą metod bayesowskich nad metodami klasycznymi. Modele hierarchiczne umożliwiają dodatkowo oszacowanie prawdopodobieństwa nieuzyskania pewnych informacji.

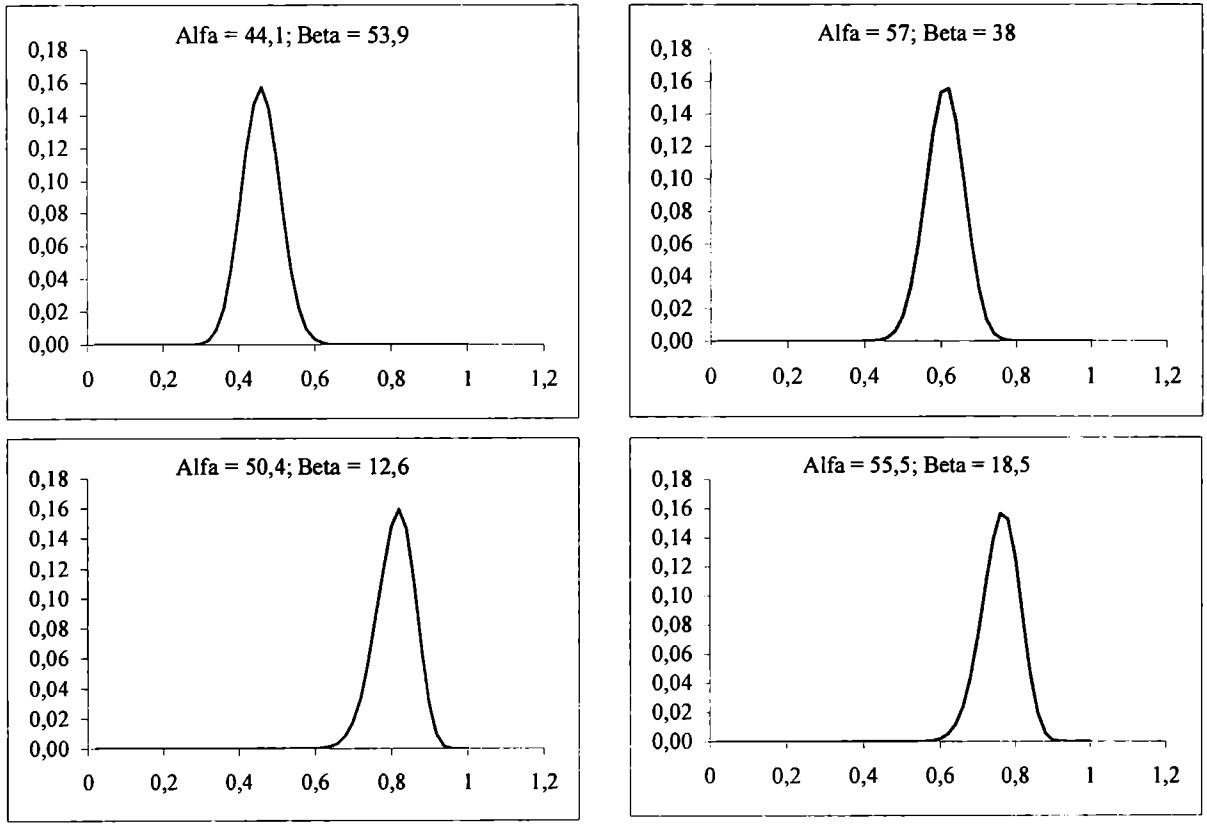
W praktyce często stosuje się pewne metody łagodzące niekorzystny wpływ brakujących danych, np. metody uzupełniania danych, nazywane także metodami imputacji. Metody imputacji mogą w istotny sposób zmniejszać bądź też eliminować obciążenie estymatorów parametrów, mają jednak poważną wadę – zaniżają oceny wariancji estymatorów. Powoduje to, że konstrukcja przedziałów ufności dla szacowanych parametrów oraz testowanie hipotez są obarczone błędem wynikającym z tego niedoszacowania (por. [4, s. 128-130; 6, s. 11-18]).


Alfa = 17,1; Beta = 0,9
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
0
0,2
0,4
0,6
0,8
1
1,2


Alfa = 6,7; Beta = 0,13
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0
0,2
0,4
0,6
0,8
1
1,2


Alfa = 31,5; Beta = 3,5
0,18
0,16
0,14
0,12
0,10
0,08
0,06
0,04
0,02
0,00
0
0,2
0,4
0,6
0,8
1
1,2


Alfa = 42,5; Beta = 7,5
0,18
0,16
0,14
0,12
0,10
0,08
0,06
0,04
0,02
0,00
0
0,2
0,4
0,6
0,8
1
1,2



Rys. 1. Wykresy rozkładów *a priori* dla parametrów π_{i0} i π_{i1}

Źródło: opracowanie własne.

Tabela 2. Charakterystyki rozkładu *a posteriori* parametrów p_1, p_2, p_3, p_4

Parametr	Wartość oczekiwana	Odchylenie standardowe	Kwantyl 2,5%	Mediana	Kwantyl 97,5%
p_1	0,5953	0,02488	0,5457	0,5955	0,6436
p_2	0,3725	0,02662	0,3133	0,3742	0,4201
p_3	0,5926	0,03162	0,527	0,594	0,6509
p_4	0,3897	0,05469	0,2971	0,3842	0,5096

Źródło: obliczenia własne.

Tabela 3. Charakterystyki rozkładu *a posteriori* parametrów π_{10}, π_{20}, π_{30}, π_{40}, π_{11}, π_{21}, π_{31}, π_{41}

Parametr	Wartość oczekiwana	Odchylenie standardowe	Kwantyl 2,5%	Mediana	Kwantyl 97,5%
π_{10}	0,9855	0,01439	0,9472	0,99	0,9997
π_{20}	0,9876	0,02358	0,9147	0,9994	1,0
π_{30}	0,8888	0,04650	0,7898	0,8924	0,9672
π_{40}	0,6578	0,05527	0,5679	0,6508	0,7854
π_{11}	0,6665	0,02597	0,6147	0,6667	0,7164
π_{21}	0,5772	0,0342	0,5149	0,5755	0,6506
π_{31}	0,7807	0,03359	0,7189	0,7791	0,8507
π_{41}	0,6588	0,06654	0,5274	0,6606	0,7833

Źródło: obliczenia własne.

Tabela 4. Charakterystyki rozkładu *a posteriori* brakujących danych v_1, v_2, v_3, v_4

Parametr	Wartość oczekiwana	Odchylenie standardowe	Kwantyl 2,5%	Mediana	Kwantyl 97,5%
v_1	58,27	2,144	52	59	60
v_2	80,98	7,85	56	85	85
v_3	66,5	11,5	41	68	85
v_4	74,67	25,26	34	71	131

Źródło: obliczenia własne.

Podejście bayesowskie, w przeciwieństwie do metod imputacji, nie jest wykorzystywane do predykcji brakujących danych i traktowania ich w późniejszej analizie tak, jak gdyby były obserwowane. Podstawowym celem tego podejścia jest eliminacja obciążenia estymatora parametru (w omawianym przykładzie – odsetka studentów ściągających na egzaminach) oraz poprawa oceny wariancji tego estymatora (por. [4, s. 131, 132; 5; 6]).

Literatura

[1] Berger J.O., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York 1985.
[2] Congdon P., *Bayesian Statistical Modelling*, J. Wiley & Sons, New York 2001.
[3] DeGroot M., *Optymalne decyzje statystyczne*, PWN, Warszawa 1981.
[4] Fairclough D.L., *Design and Analysis of Quality of Life Studies in Clinical Trials*, Chapman Hall/CRC, Washington 2002.
[5] Rubin D., *Multiple Imputation after 18+Years*, JASA 91 (1996), s. 473-520.
[6] Rubin D., *Multiple Imputation for Nonresponse in Surveys*, John Willey & Sons, New York 1987.
[7] Schafer J., *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York 2000.

THE APPLICATION OF BAYESIAN HIERARCHICAL MODELS IN STATISTICAL INFERENCE IN THE CASE OF INCOMPLETE DATA

Summary

In this paper we present the application of the Bayesian hierarchical models to analyse categorical data in the case of incomplete data sets. The Bayesian approach enables to use the knowledge about sample and missing data mechanisms. The main aim of using this approach is to improve statistical inference through the elimination of estimator bias and correct estimation of standard errors.

Dr Beata Zmyślona jest pracownikiem Katedry Statystyki w Akademii Ekonomicznej we Wrocławiu.