PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1195 — 2008
Ekonometria 20

Beata Zmyślona

ZASTOSOWANIE MODELU REGRESJI DO GENEROWANIA DANYCH

1. Wstęp

W praktyce badań społeczno-ekonomicznych występuje problem istnienia niekompletnych zbiorów danych. Tego typu zbiory utrudniają analizę statystyczną, niekorzystnie wpływają na wartości estymatorów (powodują obciążenie estymatorów) oraz oceny ich wariancji (przeszacowanie lub niedoszacowanie). Popularnym podejściem do analizy niepełnych danych jest uzupełnianie brakujących danych (tzw. metody imputacji). Wyróżnia się dwie podstawowe grupy metod imputacji. Pierwsza grupa metod polega na wyznaczaniu warunkowych rozkładów brakujących danych, z których potem generowane są próby losowe. Druga grupa zaś polega na wykorzystaniu różnego rodzaju zależności pomiędzy zmiennymi, na podstawie których buduje się model (w literaturze nazywany modelem imputacji) wykorzystywany do generowania danych (por. [3; 6; 7]).

W pracy przedstawiono zastosowanie modelu regresji do generowania danych. Metoda ilustrowana jest dwoma przykładami opartymi na rzeczywistych danych.

2. Generowanie danych za pomocą modelu regresji

Poniżej zostanie przedstawiony przykład metody generowania danych na podstawie regresji. Przykład ten dotyczy analizy danych uzyskanych poprzez dokonywanie pomiarów pewnej cechy y w n momentach. Zbiór danych przedstawia się w formie poniższego wektora:

$$\mathbf{y} = (y_1, y_2, \ldots, y_n). \tag{1}$$

W praktyce zdarza się, że pewne pomiary z różnych względów nie są dokonywane, w związku z czym w zbiorze danych występują braki. W takiej sytuacji wek-

tor (1) zostaje podzielony na dwa podwektory: $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{br})$. Pierwszy podwektor dotyczy tej części danych, których wartości są znane, drugi – tej części danych, których wartości nie są znane.

Zastosowanie modelu regresji wymaga wykorzystania pewnych zmiennych pomocniczych X_1, X_2, ..., X_q. Wartości zmiennych pomocniczych muszą być zawsze znane, aby mogły być uwzględniane w analizie niepełnych danych. W celu uproszczenia prezentacji przyjmuje się, że w modelu regresji jest wykorzystywana tylko jedna zmienna pomocnicza X.

Model regresji liniowej wyraża się następującym wzorem (por. [3]):

$$y_i = \theta_1 + \theta_2 x_i + e_i, \text{ dla } i = 1, 2, \ldots, n, \tag{2}$$

gdzie: $\theta = (\theta_1, \theta_2)$ jest wektorem nieznanych parametrów, e_i jest składnikiem losowym. Przyjmuje się, że e_i dla $i = 1, 2, \ldots, n$ są realizacjami zmiennych losowych ε_i. Do estymacji nieznanych parametrów θ_1 oraz θ_2 ze względu na brakujące dane wykorzystuje się algorytm Gibbsa (por. [1; 3; 7]).

3. Bayesowska estymacja parametrów modelu regresji

Aby wyznaczyć estymatory bayesowskie modelu (2), niezbędne jest najpierw wyspecyfikowanie rozkładów *a priori* wektora parametrów $\theta = (\theta_1, \theta_2)$ oraz rozkładu prawdopodobieństwa zmiennej losowej ε_i.

Za rozkład *a priori* wektora $\theta = (\theta_1, \theta_2)$ przyjmuje się nieinformacyjny rozkład normalny o wektorze wartości oczekiwanych $\mu_0 = (\mu_{01}, \mu_{02})$ i macierzy kowariancji $\Sigma_0 = \sigma^2 \mathbf{I}$, gdzie $\mathbf{I}$ jest macierzą jednostkową. Za rozkład błędu ε_i przyjmuje się rozkład normalny. Ponadto zakłada się, że wartość oczekiwana zmiennej ε_i jest równa zeru, natomiast wariancja jest równa σ^2. Dodatkowo przyjmuje się, że rozkład *a priori* precyzji (odwrotności wariancji) $\tau = \sigma^{-2}$ jest rozkładem gamma z parametrami $\alpha > 0$ oraz $\beta > 0$ (por. [1; 2]).

Konsekwencją przyjęcia rozkładu normalnego zmiennej ε_i jest to, że rozkład zmiennej Y_i jest rozkładem normalnym o wartości oczekiwanej $\theta_1 + \theta_2 x_i$ oraz wariancji równej σ^2.

Rozkładem *a posteriori* precyzji τ jest rozkład gamma z następującymi parametrami (por. [2]):

$$\alpha_1 = \alpha + (n/2),$$

$$\beta_1 = \beta + \frac{1}{2}\left((\mathbf{y} - \mathbf{x}\mu_1)\mathbf{y}^{\mathrm{T}} + (\mu_0 - \mu_1)\Sigma_0^{-1}\mu_0^{\mathrm{T}}\right), \tag{3}$$

gdzie $\mu_1^{\mathrm{T}} = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x} + \Sigma_0^{-1}\right)^{-1}\left(\mathbf{x}^{\mathrm{T}}\mathbf{y}^{\mathrm{T}} + \Sigma_0^{-1}\mu_0^{\mathrm{T}}\right)$, $\mathbf{x} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$.

Warunkowy rozkład *a posteriori* wektora parametrów θ jest dwuwymiarowym rozkładem Studenta o $2\alpha + n$ stopniach swobody, wektorze niecentralności μ_1 i macierzy precyzji [2]

$$(\alpha_1/\beta_1)\left(\Sigma_0^{-1} + \mathbf{x}^{\mathrm{T}}\mathbf{x}\right).$$

Na podstawie znanych wielkości y_i oraz x_i (dla $i \in \mathscr{C}$, gdzie $\mathscr{C}$ jest zbiorem indeksów odpowiadającym obserwowanym wartościom $\mathbf{y}_{\mathrm{obs}}$ zmiennej Y) wyznacza się wartości bayesowskich estymatorów $\hat{\theta}$, $\hat{\tau}$ parametrów θ oraz τ, przyjmując kwadratową funkcję straty. Wartości tych estymatorów przyjmuje się w procedurze Gibssa jako wartości startowe $\theta^{(0)}$ oraz $\tau^{(0)}$.

Iteracja $t+1$ algorytmu Gibbsa składa się z trzech kroków. W pierwszym kroku losowana jest wartość zmiennej losowej $\varepsilon_i^{(t+1)}$ (dla $i \in \mathscr{B}$, gdzie $\mathscr{B}$ jest zbiorem indeksów odpowiadającym nieznanym wartościom $\mathbf{y}_{\mathrm{br}}$ zmiennej Y) z rozkładu $N\left(0, \tau^{(t)}\right)$.

W kroku drugim obliczane są wartości zmiennej y_i w następujący sposób (por. [3]):

$$y_i^{(t+1)} = \theta_1^{(t)} + \theta_2^{(t)} x_i + e_i^{(t+1)}, \text{ dla } i \in \mathscr{B}. \tag{4}$$

W kroku trzecim na podstawie obserwowanych danych x_i (dla $i = 1, 2, \ldots, n$) oraz y_i dla $i \in \mathscr{C}$ i uzupełnionych brakujących danych $y_i^{(t+1)}$ dla $i \in \mathscr{B}$ wyznacza się wartości oczekiwane rozkładu *a posteriori* wektora parametrów θ oraz wartość oczekiwaną rozkładu a posteriori parametru τ.

Wartość oczekiwaną rozkładu *a posteriori* wektora θ oblicza się zgodnie z poniższym wzorem (por. [2]):

$$E\left(\theta^{(t+1)}\right) = \left(\mu_1^{(t+1)}\right)^{\mathrm{T}} = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x} + \left(\Sigma_0^{(t+1)}\right)^{-1}\right)^{-1}\left(\mathbf{x}^{\mathrm{T}}\left(\mathbf{y}^{(t+1)}\right)^{\mathrm{T}} + \left(\Sigma_0^{(t+1)}\right)^{-1}\mu_0^{\mathrm{T}}\right), \tag{5}$$

gdzie $\mathbf{y}^{(t+1)}$ jest wektorem, którego składowymi są obserwowane wartości zmiennej y oraz uzupełnione w kroku drugim w $(t+1)$-ej iteracji wartości obliczone zgodnie ze wzorem (4).

Z kolei wartość oczekiwaną rozkładu *a posteriori* parametru τ wyznacza się na podstawie wzoru (por. [2]):

$$E\left(\tau^{(t+1)}\right) = \frac{\alpha_1}{\beta_1^{(t+1)}}, \tag{6}$$

gdzie $\beta_1^{(t+1)}$ są wartościami parametru rozkładu gamma policzonymi na podstawie znanych wielkości oraz uzupełnionych brakujących danych w $(t+1)$-ej iteracji.

Wartości $E\left(\theta^{(t+1)}\right)$ oraz $E\left(\tau^{(t+1)}\right)$ traktuje się jako wartości parametrów $\theta^{(t+1)} = \left(\theta_1^{(t+1)}, \theta_2^{(t+1)}\right)$ oraz $\tau^{(t+1)}$, które wykorzystuje się w pierwszym i drugim kroku kolejnej iteracji.

Iteracje algorytmu Gibbsa przeprowadza się do momentu stabilizacji do rozkładów stacjonarnych. Następnie w kolejnych dziesięciu iteracjach wyznacza się wartości zgodnie ze wzorem:

$$y_i^{(l)} = \theta_1^{(l)} + \theta_2^{(l)} x_i + e_i^{(l)} \text{ dla } i \in \mathscr{B} \text{ oraz } l = 1, 2, \ldots, 10. \tag{7}$$

Wygenerowanymi za pomocą modelu regresji wartościami $y_i^{(l)}$ dla $i \in \mathscr{B}$ oraz $l = 1, 2, \ldots, 10$ uzupełnia się brakujące dane. Uzyskane w taki sposób zbiory danych stanowią bazę, na podstawie której możliwa jest dalsza analiza statystyczna, np. z zastosowaniem bayesowskich metod, np. metody imputacji wielokrotnej (por. [5; 6]).

Przykład 1

Badania społeczno-ekonomiczne obejmują swoim zasięgiem także zagadnienia związane z ochroną środowiska. Omawiany dalej przykład dotyczy analizy niepełnych danych uzyskanych w wyniku pomiarów stężeń określonych związków chemicznych w powietrzu.

W ramach realizacji programu monitoringu stanu środowiska na Dolnym Śląsku przeprowadzono badania zanieczyszczenia powietrza od stycznia do października 2003 roku (www.wroclaw.pios.gov.pl). Dokonano $n = 244$ pomiarów stężenia pyłu zawieszonego PM10 $\left[\mu g/m^3\right]$ na stanowisku pomiarowym Wojewódzkiego Inspektoratu Ochrony Środowiska we Wrocławiu przy ulicy Wierzbowej. Uzyskane dane traktuje się jako realizacje zmiennej losowej Y.

Dwadzieścia dziewięć pomiarów z pewnych względów nie zostało dokonanych. Do generowania danych wykorzystano model regresji. Do modelu wybrano jedną zmienną objaśniającą X, a mianowicie stężenie dwutlenku azotu NO_2 $\left[\mu g/m^3\right]$. Wartość współczynnika korelacji dla stężenia dwutlenku azotu oraz stężenia pyłu zawieszonego PM10 obliczona na podstawie 215 obserwowanych danych wynosi ok. 0,6.

Tabela 1. Dane otrzymane w wyniku 215 pomiarów przeprowadzonych przez Wojewódzki Inspektorat Ochrony Środowiska

NO_2	PM10	NO_2	PM10	NO_2	PM10	NO_2	PM10	NO_2	PM10
4,8	33,5	17,9	28,8	21,6	31	28,1	13,3	34,8	134,6
7,3	35,9	17,9	12,2	21,7	26,3	28,3	40,1	35,3	39,6
9	8,9	18,1	21,9	21,9	18,1	28,4	36,2	35,3	62,1
9,7	20,6	18,2	17,2	21,9	25,2	28,4	63,2	35,5	42,6
10,1	21,9	18,2	32,7	22,3	28,9	28,5	80,6	35,7	19,3
10,5	26	18,3	13,8	22,3	25,2	28,9	89,4	35,8	16,1
10,8	21	18,5	25,2	22,4	34,2	29	49,8	36	38,9
10,9	22,8	18,7	17,9	22,5	28,2	29,1	40,9	36	46,9
11,7	23,6	18,8	48,1	22,6	13,6	29,4	64,9	36,1	64,6
11,9	10	18,8	31,5	22,6	19,1	29,4	35,9	36,7	65,6
12,1	26,1	18,8	52,4	22,6	41,7	29,6	15,6	37,7	8,6
12,1	14,1	18,9	32,1	22,7	43,8	29,7	34,7	37,8	59,9
12,3	20,4	19,1	14,6	23,1	17,7	29,7	111,4	37,9	64,8
12,4	21,6	19,2	16,1	23,1	23,3	29,7	42	38,4	54,5
13	14,7	19,3	31,9	23,1	15,4	29,9	37,9	38,6	29,6
13,3	29,8	19,3	25,2	23,1	60,3	30	88,8	38,8	64,5
13,6	22,2	19,4	45	23,6	14,8	30	25,4	38,8	40,9
13,6	32,8	19,4	8,6	23,9	21,4	30,4	48,3	39,2	67
13,8	39,3	19,4	30,1	24	29,1	30,5	66,1	39,3	92,9
13,9	18,7	19,6	13,6	24,2	30	31,2	66,7	39,3	45,9
14,2	27,9	19,8	14,6	24,4	38,9	31,5	26,2	39,5	102,6
14,4	23	19,9	13,5	24,6	25,5	31,6	42,2	40,2	25,5
14,6	21,8	20,1	60,4	24,7	36,4	31,7	35,8	40,2	48,8
14,9	22,4	20,1	12	24,8	33,4	31,9	43,3	40,2	88,2
15	20,5	20,1	32,9	25	39,2	32,1	56,1	41,9	94,5
15,1	23,7	20,1	23	25,1	80,9	32,2	42,9	42,2	78,1
15,5	20,1	20,2	35	25,3	40	32,7	50,9	42,4	42,7
15,7	18,5	20,3	29	25,4	12,1	32,9	67	42,4	64,3
15,8	42,5	20,3	9,9	25,4	71	33,1	36,2	43	35,2
16,1	27,3	20,4	33,6	25,9	41,6	33,3	96,3	44,3	113,7
16,2	28,2	20,6	26,1	25,9	23,8	33,4	32,3	44,5	111,5
16,2	17,5	20,8	21,4	26	45	33,4	58,6	44,5	103,7
16,6	20,9	20,9	23,9	26,1	35,3	33,4	78,2	46,6	96
16,8	39,4	21	29,8	26,3	58,1	33,5	45,6	46,9	113,8
16,9	45,2	21	25,4	26,4	24,2	33,7	43,3	46,9	53,2
17	21,4	21,2	15,3	27	57	33,7	33,7	49	21,3
17,1	25,3	21,3	23,9	27,1	12,8	34	64,7	50	26,5
17,2	4,25	21,3	31,4	27,3	10,2	34,2	16,4	50,4	45,7
17,2	27,5	21,4	26,7	27,5	49,2	34,4	88	51,3	65,7
17,4	28,5	21,5	27,4	27,6	66,7	34,5	22,8	54,9	84,7
17,5	14,6	21,5	35,2	27,8	56,9	34,6	134,1	58,8	60
17,7	13,1	21,5	13	27,8	36,1	34,6	19,8	58,9	154,4
17,9	40,5	21,6	38,3	28	41,8	34,8	51,5	72,6	52,1

Źródło: strona internetowa Wojewódzkiego Inspektoratu Ochrony Środowiska we Wrocławiu: http://www.wroclaw.pios.gov.pl.

Uzyskane średniodobowe wyniki dla 215 pomiarów, dla których wartości zmiennych Y oraz X są znane, zostały przedstawione w tab. 1.

Dwadzieścia dziewięć pomiarów dla zmiennej X odpowiadających tym momentom czasowym, w których nie dokonano pomiarów dla zmiennej Y, przedstawiono w pierwszej kolumnie tab. 2.

Tabela 2. Wartości stężenia dwutlenku azotu oraz wygenerowane poziomy stężenia pyłu zawieszonego PM10

i	NO_2	$y_i^{(1)}$	$y_i^{(2)}$	$y_i^{(3)}$	$y_i^{(4)}$	$y_i^{(5)}$	$y_i^{(6)}$	$y_i^{(7)}$	$y_i^{(8)}$	$y_i^{(9)}$	$y_i^{(10)}$
216	14,3	6,3	12,6	16,1	3,5	27,7	23,3	35,2	31,4	20,6	18,5
217	15,8	20,2	5,8	6,7	29,8	20,2	8,9	32,6	17,7	21,4	13,8
218	16,5	31,3	31	35,3	1,5	13,9	21,3	20,2	22,5	13,8	23,6
219	21,7	34,5	4,7	23,8	23,7	21,6	26,3	27,6	27,3	20,4	6,8
220	23,7	33	21	24	32,5	18,4	16,7	27	23,6	29,8	27,4
221	26,1	16,4	14,1	17,6	19,7	14,3	17,1	32,2	21,1	19,3	11,8
222	26,1	15,9	33,7	23,6	26,4	33,3	20,5	26,6	17,6	17,5	24,7
223	26,2	19,6	16,6	41,4	21,9	28,4	19,4	14,7	22,5	29,6	13,6
224	27,1	19,6	24,3	16,6	35,1	28	30,5	35,9	22,7	26,8	27,3
225	27,6	15,7	27	17,9	20,1	37,5	30,1	17,8	40,2	25,5	20,5
226	29	9,8	7,4	13,1	16,9	17,6	35	23,4	17,6	25,1	24,5
227	30,6	25,9	-0,4	25,3	26,8	22,8	22,4	35,5	40,1	28,3	29,9
228	32,5	21,5	33,4	29,2	26,3	20,7	26,5	38,5	37,9	31,4	36,9
229	35,5	25,3	15,9	14,6	41,6	28,7	12,5	45,2	39,9	33,9	21,9
230	38	28,6	36,8	13,8	23,1	27,9	24,5	24,6	36,6	24,7	18,5
231	39,3	31,1	35,7	28,7	32,2	36,5	23	18,9	19,8	18,4	31,5
232	42,1	18,3	26,4	19,7	16,9	19,5	34,7	10,8	11,9	34,4	26,9
233	43,8	26,5	21,9	15,3	28,6	19,9	49,5	19,4	29,8	17,8	39,2
234	44,1	33,7	25,4	25,6	30,8	4,9	34,7	28,1	4,1	21,8	37
235	44,8	36,9	20,5	25,9	32,5	27,9	20,5	32,5	26,3	22,9	18,3
236	48,5	19	31,4	34,7	31,1	41,1	29,2	29,8	10,1	22,8	26,5
237	49,8	46,4	22,9	28,7	38,4	36	23,3	23,3	40	30,6	33,5
238	55,6	37,1	24,1	33,4	34,9	38	22,2	20,7	34,7	38,5	20,6
239	63	8,8	42,7	28,4	27,8	20,3	26	45,6	36,9	27,2	37,9
240	71	30,7	22,9	20,4	30,2	37,7	24,3	32,7	38,3	46,7	23,9
241	74,2	36,7	30,6	37,6	29,4	28,9	48,8	35,2	37,1	43,3	32,3
242	102,8	32,1	40	27,1	52,6	39,1	33,6	44,4	37,7	39,5	37,3
243	137,7	44,8	52,9	57,4	46,1	54,6	61,9	47,1	52,5	36,2	56,1
244	177,3	51,4	63	52,1	68,3	61,5	73	61,5	51,3	57,3	51,2

Źródło: pomiary NO_2: strona internetowa Wojewódzkiego Inspektoratu Ochrony Środowiska we Wrocławiu, http://www.wroclaw.gov.pl; pozostałe kolumny: opracowanie własne.

Do generowania danych, którymi są uzupełniane braki, wykorzystano model regresji liniowej wyrażający się następującym wzorem:

$$y_i = \theta_1 + \theta_2 x_i + e_i,$$

gdzie: y_i, x_i są wartościami uzyskanymi w i-tym pomiarze odpowiednio stężenia pyłu zawieszonego PM10 oraz stężenia dwutlenku azotu NO_2.

Przyjmuje się następujące nieinformacyjne rozkłady *a priori*: dla wektora parametrów θ dwuwymiarowy rozkład normalny, dla precyzji τ – rozkład gamma, czyli

$$\begin{gathered}\left(\theta\middle|\sigma^{-2}=\tau\right)\sim N(\mathbf{0},\ 0{,}001\cdot\mathbf{I}),\\ \tau\sim\Gamma(0{,}001;\ 0{,}001).\end{gathered}\tag{8}$$

Symbolem Γ oznacza się rozkład gamma.

Ponadto zakłada się, że błąd ε_i ma rozkład normalny $\left(\varepsilon_i\sim N\left(0,\ \sigma^{-2}\right)\right)$, zatem zmienna Y_i ma także rozkład normalny $\left(Y_i\sim N\left(\theta_1+\theta_2 x_i,\ \sigma^{-2}\right)\right)$. Na bazie 215 pomiarów, dla których realizacja zmiennej Y była obserwowana, wyznacza się wartości estymatorów bayesowskich, które wykorzystuje się jako wartości parametrów w zerowej iteracji. W analizowanym przykładzie otrzymano $\theta^{(0)}=(16{,}9;\ 0{,}329)$ oraz $\tau^{(0)}=0{,}015167$.

Dane zostały wygenerowane zgodnie z opisaną wcześniej procedurą Gibbsa. Każda iteracja Gibbsa składała się z trzech kroków. W pierwszym kroku losowana jest wartość zmiennej losowej $\varepsilon_i^{(t+1)}$ (dla $i\in\mathcal{B}$) z rozkładu normalnego o zerowej wartości oczekiwanej i precyzji $\tau^{(t)}$. W drugim kroku wyznacza się wartości y_i (dla $i\in\mathcal{B}$) zmiennej Y zgodnie ze wzorem (2). Wartościami tymi uzupełniany jest zbiór danych. W trzecim kroku na bazie uzupełnionego już zbioru danych wyznacza się wartości oczekiwane rozkładów *a posteriori* wektora parametrów θ oraz parametru τ zgodnie ze wzorami (5) oraz (6).

Po wykonaniu kilkuset iteracji po osiągnięciu zbieżności do rozkładów stacjonarnych w kolejnych dziesięciu iteracjach otrzymuje się wygenerowane dane, którymi zastępowane są brakujące obserwacje (por. [7]).

Wygenerowane dane zostały przedstawione w tab. 2. Pierwsza kolumna tab. 2 zawiera dane dotyczące stężenia dwutlenku azotu dla pomiarów dokonanych w dwudziestu dziewięciu momentach czasowych, w których nie dokonano pomiaru stężenia pyłu zawieszonego PM10. Pozostałe kolumny zwierają po dziesięć wygenerowanych danych dla każdej brakującej obserwacji dotyczącej PM10.

Przykład 2

Przykład dotyczy sytuacji opisanej w przykładzie 1, w której podejrzewa się, że w zbiorze danych występują także obserwacje odstające. Odporność na obserwacje nietypowe zależy od założonego rozkładu błędów obserwacji. W przypadku wystąpienia w zbiorze danych obserwacji odstających zakłada się, że zmienna losowa ε_i ma rozkład należący do rozkładów o tzw. „grubych ogonach", np. Laplace'a lub Studenta. Gdy błąd losowy ma rozkład o grubych ogonach, wówczas estymator

funkcji regresji wykazuje znacznie większą odporność na obserwacje nietypowe niż w przypadku rozkładu normalnego.

Przykład zostanie przedstawiony na bazie danych wykorzystanych w przykładzie 1, czyli na bazie danych dotyczących 244 pomiarów stężeń dwutlenku azotu oraz pyłu zawieszonego PM10. W przypadku pomiaru stężeń pyłu zawieszonego nie dokonano 29 pomiarów.

Do generowania danych, którymi będą uzupełniane brakujące dane, wykorzystuje się model regresji liniowej wyrażający się poniższym wzorem:

$$Y_i = \theta_1 + \theta_2 x_i + \varepsilon_i .$$

Przyjmuje się, że zmienna ε_i ma rozkład Laplace'a z parametrem położenia równym zeru oraz nieznanym parametrem skali λ, co w skrócie można zapisać jako

$$\varepsilon_i \sim L(0, \lambda) .$$

Zmienna Y_i ma także rozkład Laplace'a z parametrem położenia $\theta_1 + \theta_2 x_i$ oraz parametrem skali λ, czyli

$$Y_i \sim L(\theta_1 + \theta_2 x_i, \lambda) .$$

Przyjmuje się następujące nieinformacyjne rozkłady *a priori*: dla wektora parametrów θ dwuwymiarowy rozkład normalny, dla parametru λ – rozkład gamma, co w skrócie można zapisać w następujący sposób (por. [1; 2]):

$$(\theta|\lambda) \sim N(\mathbf{0}, 1000 \cdot \mathbf{I}),$$

$$\lambda \sim \Gamma(0{,}001, 0{,}001) .$$

Ponieważ ani rozkład normalny, ani rozkład gamma nie jest rozkładem sprzężonym dla rozkładu Laplace'a, nie można w sposób analityczny wyznaczyć rozkładów *a posteriori* wektora parametrów θ oraz parametru λ.

Do generowania danych wykorzystuje się procedurę Gibbsa ze startowymi wartościami $\theta^{(0)} = (0, 0)$ oraz $\lambda^{(0)} = 1$. Każda iteracja algorytmu Gibbsa będzie się składać z dwóch kroków.

W pierwszym kroku są losowane wartości $y_i^{(t+1)}$ (dla $i \in \mathcal{B}$, gdzie $\mathcal{B}$ jest zbiorem indeksów odpowiadającym nieobserwowanym wartościom zmiennej Y, w omawianym przypadku $i \in \{216, 217, \ldots, 244\}$) z dwudziestu dziewięciu warunkowych rozkładów brakujących danych, czyli

$$Y_i^{(t+1)} \sim p\left(y_i | \mathbf{x}, \theta_1^{(t)} + \theta_2^{(t)}, \lambda^{(t)}\right).$$

W drugim kroku losuje się wartości parametrów $\theta_1^{(t+1)}$, $\theta_2^{(t+1)}$ oraz $\lambda^{(t+1)}$ z następujących warunkowych rozkładów (por. [1; 2]):

$$\lambda^{(t+1)} \sim p\left(\lambda|\mathbf{x}, \mathbf{y}^{(t+1)}\right),$$

$$\theta_1^{(t+1)} \sim p\left(\theta_1|\mathbf{x}, \mathbf{y}^{(t+1)}, \lambda^{(t+1)}\right),$$

$$\theta_2^{(t+1)} \sim p\left(\theta_2|\mathbf{x}, \mathbf{y}^{(t+1)}, \lambda^{(t+1)}\right).$$

Wylosowanymi w pierwszym kroku wartościami uzupełniany jest zbiór danych. Zbiór uzupełnionych danych w $(t+1)$-ej iteracji oznacza się symbolem $\mathbf{y}^{(t+1)}$.

Iteracje algorytmu Gibbsa przeprowadza się do momentu stabilizacji do rozkładów stacjonarnych, którymi są rozkłady *a posteriori* brakujących danych

Tabela 3. Wygenerowane dane dla nieobserwowanych pomiarów stężenia pyłu zawieszonego PM10

i	$y_i^{(1)}$	$y_i^{(2)}$	$y_i^{(3)}$	$y_i^{(4)}$	$y_i^{(5)}$	$y_i^{(6)}$	$y_i^{(7)}$	$y_i^{(8)}$	$y_i^{(9)}$	$y_i^{(10)}$
216	6,3	12,6	16,1	3,5	27,7	23,3	35,2	31,4	20,6	18,5
217	20,2	5,8	6,7	29,8	20,2	8,9	32,6	17,7	21,4	13,8
218	31,3	31	35,3	1,5	13,9	21,3	20,2	22,5	13,8	23,6
219	34,5	4,7	23,8	23,7	21,6	26,3	27,6	27,3	20,4	6,8
220	33	21	24	32,5	18,4	16,7	27	23,6	29,8	27,4
221	16,4	14,1	17,6	19,7	14,3	17,1	32,2	21,1	19,3	11,8
222	15,9	33,7	23,6	26,4	33,3	20,5	26,6	17,6	17,5	24,7
223	19,6	16,6	41,4	21,9	28,4	19,4	14,7	22,5	29,6	13,6
224	19,6	24,3	16,6	35,1	28	30,5	35,9	22,7	26,8	27,3
225	15,7	27	17,9	20,1	37,5	30,1	17,8	40,2	25,5	20,5
226	9,8	7,4	13,1	16,9	17,6	35	23,4	17,6	25,1	24,5
227	25,9	-0,4	25,3	26,8	22,8	22,4	35,5	40,1	28,3	29,9
228	21,5	33,4	29,2	26,3	20,7	26,5	38,5	37,9	31,4	36,9
229	25,3	15,9	14,6	41,6	28,7	12,5	45,2	39,9	33,9	21,9
230	28,6	36,8	13,8	23,1	27,9	24,5	24,6	36,6	24,7	18,5
231	31,1	35,7	28,7	32,2	36,5	23	18,9	19,8	18,4	31,5
232	18,3	26,4	19,7	16,9	19,5	34,7	10,8	11,9	34,4	26,9
233	26,5	21,9	15,3	28,6	19,9	49,5	19,4	29,8	17,8	39,2
234	33,7	25,4	25,6	30,8	4,9	34,7	28,1	4,1	21,8	37
235	36,9	20,5	25,9	32,5	27,9	20,5	32,5	26,3	22,9	18,3
236	19	31,4	34,7	31,1	41,1	29,2	29,8	10,1	22,8	26,5
237	46,4	22,9	28,7	38,4	36	23,3	23,3	40	30,6	33,5
238	37,1	24,1	33,4	34,9	38	22,2	20,7	34,7	38,5	20,6
239	8,8	42,7	28,4	27,8	20,3	26	45,6	36,9	27,2	37,9
240	30,7	22,9	20,4	30,2	37,7	24,3	32,7	38,3	46,7	23,9
241	36,7	30,6	37,6	29,4	28,9	48,8	35,2	37,1	43,3	32,3
242	32,1	40	27,1	52,6	39,1	33,6	44,4	37,7	39,5	37,3
243	44,8	52,9	57,4	46,1	54,6	61,9	47,1	52,5	36,2	56,1
244	51,4	63	52,1	68,3	61,5	73	61,5	51,3	57,3	51,2

Źródło: opracowanie własne.

$p\left(y_{i(\mathrm{br})}\middle|\mathbf{y}_{\mathrm{obs}},\mathbf{x}\right)$ (dla $i=216,217,\ldots,244$ oraz rozkłady *a posteriori* parametrów $p\left(\theta\middle|\mathbf{y}_{\mathrm{obs}},\mathbf{x}\right)$ oraz $p\left(\lambda\middle|\mathbf{y}_{\mathrm{obs}},\mathbf{x}\right)$.

Następnie w kolejnych dziesięciu iteracjach losuje się wartości $y_i^{(l)}$ dla $l=1,2,\ldots,10$, $i=216,217,\ldots,244$. Wygenerowane dane w kolejnych dziesięciu iteracjach po osiągnięciu stabilizacji zostały przedstawione w tab. 3.

Uzupełnione wygenerowanymi wartościami zbiory danych stanowią podstawę wnioskowania statystycznego.

Tabela 4. Charakterystyki dwudziestu dziewięciu rozkładów *a posteriori* brakujących danych

Zmienne	Wartość oczekiwana	Odchylenie standardowe	Kwantyl 2,5%	Kwantyl 97,5%
Y_{216}	18,81	9,213	–0,166	37,84
Y_{217}	19,26	9,203	0,221	38,22
Y_{218}	19,35	9,142	–0,235	38,02
Y_{219}	20,75	8,889	2,341	40,16
Y_{220}	21,04	8,944	1,759	40,22
Y_{221}	21,90	8,964	3,118	41,77
Y_{222}	21,87	9,342	1,899	41,06
Y_{223}	21,87	8,867	2,805	41,38
Y_{224}	22,18	9,208	2,440	42,05
Y_{225}	22,18	8,989	2,888	41,12
Y_{226}	22,63	9,041	3,605	41,96
Y_{227}	23,04	9,255	3,595	43,24
Y_{228}	23,67	8,825	5,561	42,29
Y_{229}	24,30	9,562	3,790	44,72
Y_{230}	24,84	9,060	6,052	44,08
Y_{231}	25,41	9,150	5,408	44,64
Y_{232}	25,90	8,887	7,698	46,18
Y_{233}	26,59	9,221	6,816	45,74
Y_{234}	26,64	9,047	7,859	45,44
Y_{235}	26,75	8,837	8,317	45,18
Y_{236}	27,60	8,754	9,371	46,37
Y_{237}	27,93	9,061	8,565	47,07
Y_{238}	29,44	9,160	9,582	48,25
Y_{239}	31,12	9,105	10,640	49,65
Y_{240}	33,48	8,920	14,210	52,46
Y_{241}	34,42	9,024	14,840	53,70
Y_{242}	41,82	9,164	23,170	61,50
Y_{243}	51,08	9,131	32,470	70,59
Y_{244}	60,90	9,626	41,180	81,71

Źródło: opracowanie własne.

Rozkłady *a posteriori* brakujących danych $p\left(y_{i(\mathrm{br})}\middle|\mathbf{y}_{\mathrm{obs}},\mathbf{x}\right)$ (dla $i = 216, 217, ..., 244$) nie należą do żadnej klasy znanych rozkładów prawdopodobieństwa. W tabeli 4 przedstawiono podstawowe charakterystyki tych rozkładów.

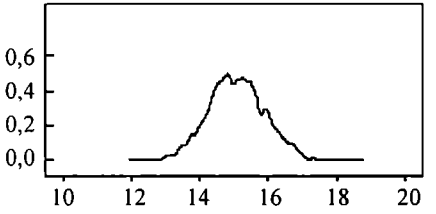
Podobnie rozkłady *a posteriori* parametrów θ_1, θ_2 oraz λ nie należą do żadnej klasy znanych rozkładów.

W tabeli 5 umieszczono charakterystyki rozkładu *a posteriori* parametru θ_1, natomiast na rys. 1 przedstawiono aproksymowaną funkcję gęstości rozkładu *a posteriori* parametru θ_1.

Tabela 5. Charakterystyki rozkładu *a posteriori* parametru θ_1

Wartość oczekiwana	Odchylenie standardowe	Kwantyl 2,5%	Kwantyl 97,5%
15,09	0,8076	13,53	16,71

Źródło: opracowanie własne.



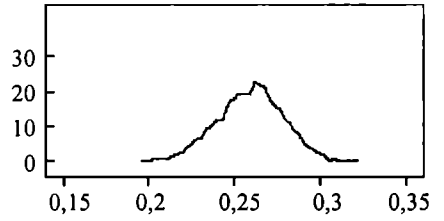
Rys. 1. Aproksymacja wykresu gęstości rozkładu *a posteriori* parametru θ_1

Źródło: opracowanie własne.

Tabela 6. Charakterystyki rozkładu *a posteriori* parametru θ_2

Wartość oczekiwana	Odchylenie standardowe	Kwantyl 2,5%	Kwantyl 97,5%
0,2591	0,01905	0,2208	0,2952

Źródło: opracowanie własne.



Rys. 2. Aproksymacja wykresu gęstości rozkładu *a posteriori* parametru θ_2

Źródło: opracowanie własne.

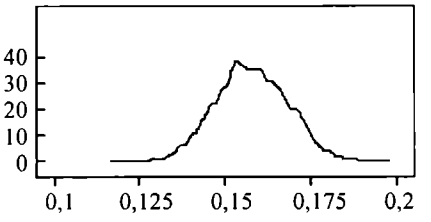
W tabeli 6 z kolei umieszczono charakterystyki rozkładu *a posteriori* parametru θ_2, na rys. 2 zaś przedstawiono jego aproksymowaną funkcję gęstości.

Tabela 7 zawiera charakterystyki rozkładu *a posteriori* parametru θ_2, rys. 3 zaś przedstawia aproksymowaną funkcję gęstości tego rozkładu.

Tabela 7. Charakterystyki rozkładu *a posteriori* parametru λ

Wartość oczekiwana	Odchylenie standardowe	Kwantyl 2,5%	Kwantyl 97,5%
0,1578	0,01059	0,1374	0,1783

Źródło: opracowanie własne.



Rys. 3. Aproksymacja wykresu gęstości rozkładu *a posteriori* parametru λ

Źródło: opracowanie własne.

Wygenerowanymi obserwacjami uzupełniane są zbiory danych, na bazie których przeprowadzana jest dalsza analiza statystyczna.

4. Uwagi końcowe

W artykule przedstawiono dwa przykłady generowania danych na podstawie modelu regresji z założeniem albo normalnego rozkładu błędu, albo rozkładu Laplace'a. Na bazie uzupełnionych zbiorów danych wygenerowanymi obserwacjami możliwa jest dalsza analiza statystyczna, np. z zastosowaniem bayesowskich metod Monte Carlo lub metody imputacji wielokrotnej (por. [3; 5; 6; 7]). Tego typu metody umożliwiają dokładniejsze szacownie nieznanych parametrów poprzez zmniejszanie obciążenia estymatora oraz oszacowanie oceny wariancji z mniejszym błędem (por. np. [3; 7]).

Literatura

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

[2] DeGroot M., *Optymalne decyzje statystyczne*, PWN, Warszawa 1981.

[3] Fairclough D.L., *Design and Analysis of Quality of Life Studies in Clinical Trials*, Chapman Hall/CRC, Washington 2002.
[4] Neal R., *Probabilistic Inference Using Markov Monte Carlo Methods*, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Toronto 1993.
[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 REGRESSION MODEL TO DATA GENERATION

Summary

The most popular methods of statistical inference in the case of incomplete data sets are based on the data augmentation. The multiple imputation and parameter simulation belong to the methods of data augmentation. The main idea of these techniques consists in creating the m complete data sets by replacing each missing data with a set of plausible values, which are generated either from some conditional distribution of missing data or using some imputation model.

In this paper, we present the methods of generating data using some regression model and its applying to the analysis of incomplete data of air pollution measurement.

Beata Zmyślona – dr, adiunkt w Katedrze Statystyki Akademii Ekonomicznej we Wrocławiu.