Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 37
Ekonometria 23 2009

Dariusz Biskup
Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE BAYESOWSKICH METOD WYBORU MODELU DO IDENTYFIKACJI CZYNNIKÓW WPŁYWAJĄCYCH NA JAKOŚĆ ŻYCIA

Streszczenie: W artykule opisana została bayesowska procedura wyboru zmiennych oraz funkcji łączącej w uogólnionym modelu liniowym. Procedura ta wykorzystuje metodę Monte Carlo, a zwłaszcza algorytm *reversible jump*. Opisany algorytm zastosowany został do identyfikacji czynników wpływających na jakość życia. Analizie poddane zostały dane ankietowe, w których analizowaną zmienną jest odpowiedź na pytanie, czy dana osoba określa swoje życie jako szczęśliwe. Stwierdzono, że czynnikami wpływającymi na szczęście człowieka są m.in. płeć, dochód, wiek oraz wykształcenie.

Słowa kluczowe: bayesowski wybór modelu, algorytm *reversible jump*, uogólnione modele liniowe.

1. Wstęp

Uogólnione modele liniowe (por. np. [Agresti 2002]) stosowane są do modelowania zależności pomiędzy zmienną objaśnianą Y a zbiorem zmiennych objaśniających X_1, X_2, …, X_k w ten sposób, że jeśli $\mu = E(Y)$, to $g(\mu)$ jest kombinacją liniową zmiennych objaśniających, natomiast g jest tzw. funkcją łączącą. Do najpopularniejszych funkcji łączących, gdy zmienna Y ma rozkład Bernoullego, należą: funkcja logitowa – $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, funkcja probitowa – $g(\mu) = \Phi^{-1}(\mu)$ (gdzie $\Phi(\cdot)$ jest dystrybuantą rozkładu normalnego), funkcja log-log – $g(\mu) = -\log\left(-\log(\mu)\right)$, funkcja komplementarna log-log – $g(\mu) = -\log\left(-\log(1-\mu)\right)$. Ogólna postać uogólnionego modelu liniowego jest zatem następująca:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k .$$

Wybór modelu w przypadku uogólnionego modelu liniowego może obejmować dwa elementy. Po pierwsze istotne jest wybranie odpowiedniego zbioru

zmiennych objaśniających, a po drugie należy wybrać właściwą funkcję łączącą. Zagadnieniom tym poświęcona zostanie dalsza część artykułu. Jego część teoretyczna opierać się będzie na pracy [Ntzoufras i in. 2003]. Implementacja numeryczna obliczeń prawdopodobieństwa modelu wykorzystywać będzie algorytm *reversible jump* Greena (por. [Green 1995]).

Przestrzeń modeli w rozpatrywanym zagadnieniu będzie składać się zatem z elementów zbioru $\{0,1\}^k \times \mathcal{L}$. Pierwsza część wzoru określa, które z *k* potencjalnych zmiennych wchodzą w skład modelu, natomiast $\mathcal{L}$ oznacza zbiór rozpatrywanych funkcji łączących. W dalszej części rozpatrywane będą funkcje logitowa, probitowa, log-log oraz funkcja komplementarna log-log.

2. Rozkład *a priori*

Jednym z najistotniejszych elementów bayesowskiego zagadnienia wyboru modelu jest ustalenie właściwego rozkładu *a priori* na przestrzeni parametrów modelu oraz na przestrzeni samych modeli. Przyjęte zostanie naturalne założenie, że *a priori* każda z rozpatrywanych funkcji łączących jest jednakowo prawdopodobna, tzn. $p(L) = 1/|\mathcal{L}| = 0{,}25,\ L \in \mathcal{L}$. Ponadto przyjęte zostanie, że jednakowo prawdopodobne są wszystkie kombinacje zmiennych objaśniających. Pozostaje więc określenie rozkładu *a priori* dla parametrów regresyjnych β.

Niech $\boldsymbol{\gamma}$ oznacza wektor indeksów określający, które spośród *k* zmiennych znajdzie się w określonym modelu. Ponadto niech $\beta_{\gamma L}$ oznacza wektor parametrów regresyjnych skojarzonych ze zbiorem zmiennych γ i z funkcją łączącą. Podobnie jak w [Ntzoufras i in. 2003] przyjęte zostanie, że

$$\boldsymbol{\beta}_{\gamma L} \big| \gamma, L \sim N\left(\theta_{\gamma L}, \Sigma_{\gamma L}\right).$$

Wektor $\boldsymbol{\beta}_{\gamma L}$ zostanie podzielony na dwie części: $\left(\beta_{\gamma L 0}, \beta^*_{\gamma L}\right)$, gdzie $\beta_{\gamma L 0}$ oznacza wyraz wolny. Wartość oczekiwana rozkładu *a priori* dla składników wektora $\beta^*_{\gamma L}$ będzie równa zeru.

Ponieważ rozpatrywane będą modele dla różnych funkcji łączących, wydaje się pożądane, aby istniała zależność pomiędzy parametrami rozkładów *a priori* związanych z poszczególnymi funkcjami łączącymi. Powinien zatem istnieć związek między wektorami $\boldsymbol{\theta}_{\gamma L}$ i macierzami $\boldsymbol{\Sigma}_{\gamma L}$ dla różnych funkcji łączących *L*. Powiązanie takie opierające się na rozwinięciu w szereg Taylora (por. [Ntzoufras i in. 2003]) prowadzi do następujących zależności dotyczących parametrów β:

$$\beta_{\gamma L_1 0} = \frac{g_{L_1}(\mu_0)}{g_{L_2}(\mu_0)} \beta_{\gamma L_2 0} + g_{L_1}(\mu_0) - \frac{g'_{L_1}(\mu_0)}{g'_{L_2}(\mu_0)} g_{L_2}(\mu_0), \tag{1}$$

$$\boldsymbol{\beta}^{*}_{\gamma L_1} = \frac{g'_{L_1}(\mu_0)}{g'_{L_2}(\mu_0)} \boldsymbol{\beta}^{*}_{\gamma L_2} . \tag{2}$$

Analogicznie wyznaczane są zależności dla wartości oczekiwanej *a priori* wyrazu wolnego oraz macierzy kowariancji rozkładu *a priori*:

$$\theta_{\gamma L_1} = \frac{g'_{L_1}(\mu_0)}{g'_{L_2}(\mu_0)} \theta_{\gamma L_2} + g_{L_1}(\mu_0) - \frac{g'_{L_1}(\mu_0)}{g'_{L_2}(\mu_0)} g_{L_2}(\mu_0), \tag{3}$$

$$\Sigma_{\gamma L_1} = \Sigma_{\gamma L_2} \left(\frac{g'_{L_1}(\mu_0)}{g'_{L_2}(\mu_0)} \right)^2 . \tag{4}$$

Wzory (3) i (4) podają związek między parametrami rozkładów *a priori* mających funkcje łączące L_1 i L_2. Wzory te mają charakter przybliżony, wynikający z rozwinięcia Taylora względem punktu μ_0. Sposób ustalenia wartości μ_0 podany zostanie w dalszej części.

Dzięki powyższym wzorom niezbędne staje się określenie tylko wartości oczekiwanej $\theta_{\gamma L}$ rozkładu *a priori* parametru $\beta_{\gamma L 0}$ oraz macierzy kowariancji $\Sigma_{\gamma L}$ rozkładu *a priori* $\boldsymbol{\beta}_{\gamma L}$.

Załóżmy, że dysponujemy zbiorem *n* zmiennych losowych $Y_1, Y_2, \ldots, Y_n$ o rozkładzie dwumianowym, dla których liczba prób wynosi odpowiednio $m_1, m_2, \ldots m_n$.

Określenie macierzy kowariancji $\Sigma_{\gamma L}$ dokonane zostanie przy użyciu metody jednostkowej informacji *a priori* zaproponowanej w [Kass, Wasserman 1995]. W przypadku gdy obserwacje pochodzą z rozkładu Bernoullego i logitowej funkcji łączącej, prowadzi to do następującego wyniku:

$$\Sigma_{\gamma L} = 4\phi \sum_{i=1}^{n} m_i \left(\mathbf{X}_{\gamma}^{\mathrm{T}} \mathbf{X}_{\gamma} \right)^{-1},$$

gdzie ϕ oznacza parametr skali, który proponuje się ustalać zgodnie z zależnością $\phi^{-1} = \max(m_i)$.

Parametr μ_0 określający punkt, w którym przybliżenie zastosowane we wzorach (3) i (4) jest najlepsze, ustala się na poziomie $\mu_0 = \sum m_i y_i \big/ \sum m_i$.

3. Algorytm *reversible jump*

Ogólne założenia algorytmu *reversible jump* opisane zostały np. w [Biskup 2006]. Obecnie przedstawiona zostanie adaptacja tego algorytmu na potrzeby wyboru zmiennych oraz funkcji łączącej w uogólnionym modelu liniowym (por. [Ntzoufras

i in. 2003]). Pojedyncza iteracja algorytmu wymaga aktualizowania parametrów $\boldsymbol{\beta}_{\gamma L}$, γ oraz L i składa się z następujących etapów:

1. Wygenerowanie poszczególnych elementów wektora $\boldsymbol{\beta}_{\gamma L}$ z warunkowego rozkładu *a posteriori* $p\left(\beta_{\gamma Li}\middle|\boldsymbol{\beta}_{\gamma L\setminus i},\gamma,L,\mathbf{y}\right)$. Stosowany jest zatem algorytm Gibbsa. Ponieważ nie jest możliwe analityczne wyznaczenie takiego rozkładu warunkowego, można zastosować jedną z metod adaptacyjnych.

2. Wylosowanie jednej ze zmiennych $j\in\{1,2,\ldots,k\}$ i dodanie jej do modelu lub jej usunięcie z modelu z prawdopodobieństwem $1/k$. Generowany zostaje zatem nowy wektor γ', który różni się od aktualnego wektora γ o jedną zmienną, która zostaje albo dodana, albo usunięta.

3. Jeżeli następuje dodanie nowej zmiennej, to pojawia się dodatkowy parametr β'_j, którego wartość losowana jest z rozkładu $q_j\left(\beta'_j\middle|L\right)$. Wartości pozostałych parametrów się nie zmieniają. Akceptacja nowego wektora parametrów $\boldsymbol{\beta}'_{\gamma' L}$ następuje z prawdopodobieństwem

$$\min\left\{1,\frac{p\left(\mathbf{y}\middle|\boldsymbol{\beta}'_{\gamma' L},\gamma',L\right)p\left(\boldsymbol{\beta}'_{\gamma' L}\middle(\gamma',L\right)p(\gamma',L)}{p\left(\mathbf{y}\middle|\boldsymbol{\beta}_{\gamma L},\gamma,L\right)p\left(\boldsymbol{\beta}_{\gamma L}\middle|\gamma,L\right)p(\gamma,L)q\left(\beta'_j\middle|L\right)}\right\}.$$

Jeśli nastąpi akceptacja, to wartości g i $\boldsymbol{\beta}_{\gamma L}$ zostają zastąpione przez γ' i $\boldsymbol{\beta}'_{\gamma' L}$. W przeciwnym wypadku pozostają one bez zmian.

4. Jeżeli następuje usunięcie zmiennej j, to pozostawione parametry zachowują swoje wartości. Akceptacja nowego wektora parametrów $\boldsymbol{\beta}'_{\gamma' L}$ następuje z prawdopodobieństwem

$$\min\left\{1,\frac{p\left(\mathbf{y}\middle|\boldsymbol{\beta}'_{\gamma' L},\gamma',L\right)p\left(\boldsymbol{\beta}'_{\gamma' L}\middle|\gamma',L\right)p\left(\gamma',L\right)q_j\left(\beta_j\middle|L\right)}{p\left(\mathbf{y}\middle|\boldsymbol{\beta}_{\gamma L},\gamma,L\right)p\left(\boldsymbol{\beta}_{\gamma L}\middle|\gamma,L\right)p(\gamma L)}\right\}.$$

Jeśli nastąpi akceptacja, to wartości γ i $\boldsymbol{\beta}_{\gamma L}$ zostają zastąpione przez γ' i $\boldsymbol{\beta}'_{\gamma' L}$. W przeciwnym wypadku pozostają one bez zmian.

5. Wylosowanie nowej funkcji łączącej $L'\neq L$ z prawdopodobieństwem $j(L,L)=1/\left(\left|\mathcal{L}\right|-1\right)$. Obliczenie nowych wartości parametrów $\boldsymbol{\beta}'_{\gamma' L}$ przy użyciu wzorów (1) i (2). Akceptacja nowej funkcji łączącej z prawdopodobieństwem

$$\min\left\{1, \frac{p\left(\mathbf{y}|\boldsymbol{\beta}'_{\gamma' L}, \gamma', L\right) p\left(\boldsymbol{\beta}'_{\gamma' L}\middle|\gamma', L\right) j(L'L)}{p\left(\mathbf{y}|\boldsymbol{\beta}_{\gamma L}, \gamma, L\right) p\left(\boldsymbol{\beta}_{\gamma L}\middle|\gamma, L\right) p(\gamma, L) j(L, L')} \left|\frac{\partial \boldsymbol{\beta}'_{\gamma' L}}{\partial \boldsymbol{\beta}'_{\gamma' L}}\right|\right\}$$

gdzie $\left|\frac{\partial \boldsymbol{\beta}'_{\gamma' L}}{\partial \boldsymbol{\beta}'_{\gamma' L}}\right| = \left(\frac{g'_{L'}(\mu_0)}{g'_{L}(\mu_0)}\right)^{d(\gamma)}$, a $d(\gamma)$ oznacza liczbę parametrów modelu.

Jeśli nastąpi akceptacja, to wartości L i $\boldsymbol{\beta}_{\gamma L}$ zostają zastąpione przez L' i $\boldsymbol{\beta}'_{\gamma' L}$. W przeciwnym wypadku pozostają one bez zmian.

4. Ocena ankietowa jakości życia

4.1. Opis danych

W przykładzie analizie poddane zostaną dane zgromadzone w ramach projektu „Diagnoza społeczna. Warunki i jakość życia Polaków" (por. [Czapiński, Panek 2007]). Analizowaną zmienną zależną będzie odpowiedź na postawione w ankiecie pytanie „Biorąc wszystko razem pod uwagę, jak oceniłby Pan swoje życie w tych dniach – czy mógłby Pan powiedzieć, że jest:". Respondent miał do wyboru odpowiedzi: bardzo szczęśliwy, dosyć szczęśliwy, niezbyt szczęśliwy, nieszczęśliwy. Na potrzeby analizy dokonano pogrupowania odpowiedzi na dwie kategorie w ten sposób, że odpowiedzi „bardzo szczęśliwy" i „dosyć szczęśliwy" zakodowano jako 1, natomiast pozostałe odpowiedzi jako 0. Zastosowane zostaną dwa zbiory potencjalnych zmiennych objaśniających. Zbiór pierwszy obejmować będzie następujące zmienne:

X_1 – odpowiedź na pytanie „Jak często przeciętnie w ciągu miesiąca bierze Pan udział w nabożeństwach lub innych spotkaniach o charakterze religijnym?". Możliwe kategorie odpowiedzi to: 0, 1-3, ≥ 4.

X_2 – odpowiedź na pytanie „Czuł Pan, że Pana źródło dochodów jest niestałe i niepewne?:". Możliwe kategorie odpowiedzi to: często, zdarzyło się, nigdy.

X_3 – odpowiedź na pytanie „Był Pan traktowany niesprawiedliwie przez innych w pracy?". Możliwe kategorie odpowiedzi to: często, zdarzyło się, nigdy.

X_4 – odpowiedź na pytanie „Czy czuje się Pan kochany i darzony zaufaniem?". Możliwe kategorie odpowiedzi: tak, nie.

X_5 – płeć (mężczyzna – 1, kobieta – 0).

X_6 – wiek. Możliwe kategorie odpowiedzi to: do 24 lat, 25-34 lata, 35-44 lata, 45-59 lat, 60-64 lata, 65 i więcej lat.

X_7 – miejsce zamieszkania. Możliwe kategorie odpowiedzi to: miasto, wieś.

X_8 – wykształcenie. Możliwe kategorie odpowiedzi to: nie dotyczy (osoba w wieku 0-12 lat), podstawowe, średnie i wyższe.

X_9 – dochód miesięczny netto (na rękę) średnio z ostatnich trzech miesięcy (w zł).

Wszystkie zmienne z wyjątkiem X_9 mają charakter kategorialny. Część z nich ma charakter binarny, niektóre mają więcej niż dwie kategorie. W przypadku tych ostatnich zmiennych wprowadzone zostaną dodatkowe, sztuczne zmienne binarne, tak aby można było dokonać ich oddzielnej interpretacji. W związku z tym wprowadzono następujące dodatkowe zmienne (zmienne binarne, które nie podlegają modyfikacji nie zostały uwzględnione w tabeli):

Tabela 1. Podział zmiennych kategorialnych

Zmienna	Kategorie	Nowe zmienne
X_1	0	
	1-3	X_{1A}
	≥ 4	X_{1B}
X_2	często	X_{2A}
	zdarzyło się	X_{2B}
	nigdy	
X_3	często	X_{3A}
	zdarzyło się	X_{3B}
	nigdy	
X_6	do 24	
	25-34	X_{6A}
	34-44	X_{6B}
	45-59	X_{6C}
	60-64	X_{6D}
	65 i więcej	X_{6E}
X_8	nie dotyczy (osoba w wieku 0-12 lat)	
	podstawowe	X_{8A}
	średnie	X_{8B}
	wyższe	X_{8C}

Źródło: opracowanie własne.

Należy zwrócić uwagę, że zmienna licząca *k* kategorii jest zawsze zastępowana poprzez $k - 1$ zmiennych. Zawsze istnieje bowiem jedna kategoria bazowa, względem której interpretuje się wartości parametrów związanych z poszczególnymi zmiennymi. Na przykład w przypadku zmiennej X_{6A} parametr z nią związany będzie interpretowany jako zmiana prawdopodobieństwa dla zmiennej *Y* związana z faktem bycia w wieku 25-34 lata w stosunku do osób, które są w wieku do 24 lat.

Po wprowadzeniu sztucznych zmiennych pełny model mieć będzie 18 zmiennych objaśniających oraz 19 parametrów (jeden związany z wyrazem wolnym modelu). Liczba potencjalnych modeli, jaka powstaje w wyniku uwzględnienia takiej liczby zmiennych, wynosi $2^{18} = 262\,144$ (zakładamy, że każdy model mieć będzie

wyraz wolny). W modelu nie będą uwzględniane interakcje pomiędzy zmiennymi, ponieważ uwzględnienie nawet tylko interakcji drugiego rzędu zwiększyłoby liczbę parametrów modelu do przeszło stu, co spowodowałoby, że liczba potencjalnych modeli byłaby zbyt duża.

Drugi zbiór zmiennych objaśniających będzie węższy, obejmować będzie jednak również interakcje pierwszego rzędu. Drugi wariant obejmować będzie więc zmienne: X_4, X_5, X_7, X_9. Po uwzględnieniu interakcji drugiego rzędu pełny model będzie miał 11 parametrów, a liczba potencjalnych modeli będzie równa $2^{10} = 1024$.

Do obliczenia prawdopodobieństw *a posteriori* wykorzystany został model opisany w poprzednich paragrafach. Dla wariantu pierwszego (18 zmiennych) przeprowadzono 2 200 000 iteracji, z których 2 000 000 wykorzystane zostały do wyznaczenia prawdopodobieństw. Dla wariantu drugiego ze względu na znacznie mniejszą liczbę potencjalnych modeli przeprowadzono 500 000 iteracji, z których wykorzystano 300 000.

Dla wariantu pierwszego dostępny zbiór danych liczył 3297 obserwcji, a dla drugiego 10 565. Różnica w liczbie obserwacji wynika z występowania brakujących danych dla poszczególnych zmiennych. W każdym przypadku wykorzystywane były dane tylko dla osób, które udzieliły odpowiedzi na wszystkie pytania ze zbioru potencjalnych zmiennych objaśniających.

Czas obliczeń dla wariantu pierwszego (2,2 mln iteracji) wyniósł ok. 40 godzin, natomiast dla wariantu drugiego ok. 6 godzin (komputer z procesorem Intel E8500, program w języku Delphi).

4.2. Ocena ankietowa jakości życia – wyniki obliczeń

Dla wariantu pierwszego (18 zmiennych) w wyniku przeprowadzonych obliczeń niezerowe prawdopodobieństwa uzyskano dla 2721 modeli (ze względu na wybór zmiennych). Uzyskane wyniki można analizować z co najmniej kilku punktów widzenia, ponieważ model jest w analizowanym problemie zdefiniowany przez wybór funkcji łączącej oraz wybór zmiennych. Tabela 1 przedstawia rozkład brzegowy funkcji łączącej. Jak widać, zdecydowanie najbardziej prawdopodobna jest funkcja logit. Znaczące prawdopodobieństwo ma jeszcze tylko funkcja probit.

Tabela 1. Rozkład *a posteriori* funkcji łączącej

Funkcja łącząca	Logit	Probit	Log-log	Clog-log
p	0,797212	0,181855	0,0000045	0,020929

Źródło: opracowanie własne.

W przypadku konieczności wyboru najlepszego zestawu zmiennych sytuacja nie jest już tak jednoznaczna. Najbardziej prawdopodobny model ma prawdopodobieństwo równe tylko ok. 0,11. Jeśli jednak popatrzymy na zmienne występujące

w siedmiu najbardziej prawdopodobnych modelach, okazuje się, że znaczna część zmiennych się powtarza. Potwierdza to również tab. 4. Okazuje się, że zmienne X_9, X_4 i X_{2A} występują we wszystkich modelach (z prawdopodobieństwem 1), zmienne X_{6C} i X_{2B} mają prawdopodobieństwo bliskie jedności, zmienne X_{8A}, X_{3A}, X_{3B} mają prawdopodobieństwo ok. 0,75. Najbardziej „kontrowersyjna" okazuje się zmienna X_5, która ma prawdopodobieństwo równe ok. 0,52. Co ciekawe wszystkie wymienione zmienne (i tylko one) należą do modelu najbardziej prawdopodobnego (tab. 2). Model najbardziej prawdopodobny różni się natomiast od modelu drugiego w kolejności właśnie o zmienną X_5.

Tabela 2. Rozkład *a posteriori* zbiorów zmiennych

Wybrane zmienne	*p*
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6C}, X_{8A}, X_9	0,10910
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_{6C}, X_{8A}, X_9	0,07029
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6A}, X_{6C}, X_{8A}, X_9	0,05426
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_{6A}, X_{6C}, X_{8A}, X_9	0,04451
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6C}, X_{6E}, X_{8A}, X_9	0,02968
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_{6C}, X_9	0,02289
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6A}, X_{6C}, X_9	0,02049

Źródło: opracowanie własne.

Wydaje się więc, że do modelowania zmiennej *Y* należałoby uwzględnić model logitowy oraz zmienne: X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6C}, X_{8A}, X_9 (tab. 3). Można zatem powiedzieć, że wpływ na poczucie szczęścia spośród rozpatrywanych zmiennych mają następujące czynniki: częste lub sporadyczne poczucie niestałości i niepewności dochodów (X_{2A},X_{2B}), niesprawiedliwe traktowanie w pracy (X_{3A}, X_{3B}), poczucie, że jest się kochanym i darzonym zaufaniem (X_4), płeć (X_5), bycie w wieku od 45 do 59 lat, posiadanie wykształcenia podstawowego (X_{8A}) oraz dochód (X_9). Nie mają natomiast wpływu na szczęście m.in.: religijność, miejsce zamieszkania oraz wybrane kategorie wieku i wykształcenia.

W wariancie drugim potencjalny zestaw zmiennych objaśniających składa się ze zmiennych X_4, X_5, X_7, X_9 oraz ich interakcji X_4X_5, X_4X_7, X_4X_9, X_5X_7, X_5X_9, X_7X_9.

Prawdopodobnie ze względu na znacznie wyższą liczbę obserwacji (ponad 10 000) tym razem otrzymane wyniki są znacznie bardzie jednoznaczne, zwłaszcza w odniesieniu do wyboru funkcji łączącej.

Tabela 5 przedstawia rozkład brzegowy funkcji łączącej. Jak widać, zdecydowanie najbardziej prawdopodobna jest funkcja log-log. Prawdopodobieństwo to jest równe niemal 1. Z tego względu nie przedstawiono rozkładu funkcji łączącej i zbioru zmiennych objaśniających.

Tabela 3. Łączny rozkład zbioru zmiennych i funkcji łączącej

Wybrane zmienne	Logit	Probit	Log-log	Clog-log
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6C}, X_{8A}, X_9	0,08710	0,02076	0,00000	0,00124
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_{6C}, X_{8A}, X_9	0,05791	0,01071	0,00000	0,00167
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6A}, X_{6C}, X_{8A}, X_9	0,04258	0,01116	0,00000	0,00053
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_{6A}, X_{6C}, X_{8A}, X_9	0,03579	0,00798	0,00000	0,00073
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6C}, X_{6E}, X_{8A}, X_9	0,02031	0,00795	0,00000	0,00142
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_{6C}, X_9	0,01966	0,00273	0,00000	0,00050
X_{2A}, X_{2B}, X_{3A}, X_{3B}, X_4, X_5, X_{6A}, X_{6C}, X_9	0,01671	0,00327	0,00000	0,00051

Źródło: opracowanie własne.

Tabela 4. Prawdopodobieństwa występowania zmiennych w poszczególnych modelach

X_9	X_{8C}	X_{8E}	X_{8A}	X_7	X_{6E}	X_{6D}	X_{6C}	X_{6B}	X_{6A}	X_5	X_4	X_{3B}	X_{3A}	X_{2B}	X_{2A}	X_{1E}	X_{1A}
1,00	0,06	0,08	0,75	0,03	0,18	0,06	0,94	0,11	0,37	0,52	1,00	0,79	0,78	0,99	1,00	0,10	0,04

Źródło: opracowanie własne,

Tabela 5. Rozkład *a posteriori* funkcji łączącej

Funkcja łącząca	Logit	Probit	Log-log	Clog-log
p	0,00004667	0	0,99995333	0

Źródło: opracowanie własne.

Jeśli chodzi o najbardziej prawdopodobne modele ze względu na uwzględniane zmienne objaśniające, istnieją dwa modele mające zdecydowanie wyższe prawdopodobieństwa od pozostałych. Różnią się one sposobem uwzględnienia zmiennej X_5 (płci). W pierwszym przypadku występuje ona w interakcji ze zmienną X_4 (czy osoba czuje się kochana), w drugim występuje ona samodzielnie. Tabela 6 przedstawia jeszcze dwa inne modele, które zajęły trzecie i czwarte miejsce ze względu na ich prawdopodobieństwo. Pozostałe modele miały prawdopodobieństwa mniejsze niż 0,01.

Trudno jest jednoznacznie stwierdzić, który z dwóch najbardziej prawdopodobnych modeli powinien być wybrany. Nie jest tutaj również pomocna analiza uwzględniająca prawdopodobieństwa występowania poszczególnych zmiennych we wszystkich modelach (tab. 7). Wynika z niej bowiem, że interesujące nas zmienne X_4 i X_4X_5 mają niemal jednakowe prawdopodobieństwo wystąpienia.

Tabela 6. Rozkład *a posteriori* zbiorów zmiennych

Wybrane zmienne	p
X_4, X_9, X_4X_5, X_4X_9	0,423645
X_4, X_5, X_9, X_4X_9	0,402468
X_4, X_9, X_4X_9, X_5X_9	0,05679
X_4, X_9, X_4X_9	0,047153

Źródło: opracowanie własne.

Tabela 7. Prawdopodobieństwa występowania zmiennych w poszczególnych modelach

X_7X_9	X_5X_9	X_4X_9	X_5X_7	X_4X_7	X_4X_5	X_9	X_7	X_5	X_4
0,01317	0,073227	1	0,01689	0,01394	0,456978	1	0,011823	0,430075	1

Źródło: opracowanie własne.

Tabela 8. Estymatory parametrów wybranego modelu

Zmienna	Opis	Estymator parametru (wartość oczekiwana rozkładu *a posteriori*)	Odchylenie standardowe rozkładu *a posteriori*
Wyraz wolny		–0,8058	0,232
X_{2A}	częste poczucie niestałości dochodu	–1,401	0,1522
X_{2B}	sporadyczne poczucie niestałości dochodu	–0,5014	0,1323
X_{3A}	częste poczucie niesprawiedliwego traktowania w pracy	–0,7113	0,2067
X_{3B}	sporadyczne poczucie niesprawiedliwego traktowania w pracy	–0,3819	0,1073
X_4	czy czujesz się kochany i darzony zaufaniem	2,622	0,164
X_5	płeć	0,2984	0,1043
X_{6C}	wiek 45-59 lat	–0,488	0,1025
X_{8A}	wykształcenie podstawowe	–0,4043	0,1068
X_9	dochód	0,0003478	0,00004219

Źródło: opracowanie własne.

Na zakończenie przedstawione zostaną wyniki estymacji parametrów modelu najbardziej prawdopodobnego dla wariantu pierwszego. Pozwolą one na stwierdzenie, jaki jest kierunek zależności pomiędzy odpowiedziami na poszczególne pytania ankiety a prawdopodobieństwem oceny swojego życia jako szczęśliwego.

Można zatem stwierdzić, że najwyższe prawdopodobieństwo bycia szczęśliwym występuje, kiedy osoba nie ma poczucia niestałości dochodów, nie jest niesprawiedliwie traktowana w pracy, czuje się kochana i darzona zaufaniem, jest mężczyzną, nie jest w wieku 45-59 lat, ma wykształcenie wyższe niż podstawowe i ma jak najwyższe dochody. Osoba tak zdefiniowana i mająca dochody w wysokości 10 000 zł ma prawdopodobieństwo 0,996, że będzie szczęśliwa.

Z kolei osoba mająca często poczucie niestałości dochodu, nierzadko niesprawiedliwie traktowana w pracy, nieczująca się kochana i darzona zaufaniem, będąca kobietą w wieku 45-59 o wykształceniu podstawowym i z miesięcznym dochodem 500 zł ma prawdopodobieństwo 0,0025, że będzie szczęśliwa. Chociaż jeśli zwiększymy dochody takiej osoby do 10 000 zł, to prawdopodobieństwo to wzrasta do 0,42.

Analiza odchyleń standardowych rozkładu *a posteriori* pozwala stwierdzić, że oszacowania wszystkich parametrów są wiarygodne – w żadnym przypadku nie są one na tyle wysokie, aby mogły sugerować zmianę kierunku zależności pomiędzy zmienną objaśnianą a objaśniającą.

Literatura

Agresti A., *Categorical Data Analysis*, John Wiley & Sons, 2002.

Biskup D., *Wybór modelu oraz zmiennych do modelu w ujęciu bayesowskim*, [w:] *Praktyka statystyki*, red. W. Miszczak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1140, AE, Wrocław 2006.

Czapiński J., Panek T., *Diagnoza społeczna 2007. Warunki i jakość życia Polaków*, Rada Monitoringu Społecznego, 2007.

Green P., *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, „Biometrika" 1995 nr 82, s. 711-732.

Kass R.E., Wasserman L., *A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion*, „Journal of American Statistical Association" 1995 nr 90, s. 928-934.

Ntzoufras I., Dellaportas P., Forster J.J., *Bayesian variable and link determination for generalized linear models*, „Journal of Statistical Planning and Inference" 2003 nr 111, s. 165-180,

APPLICATION OF BAYESIAN MODEL CHOICE PROCEDURES TO IDENTIFY FACTORS INFLUENCING THE QUALITY OF LIFE

Summary: The paper describes a bayesian procedure for variable and link function selection in the generalized linear model. The procedure uses Monte Carlo methods, specifically the reversible jump algorithm. The procedure has been used to identify factors influencing the quality of life. The subjects of the analysis was survey data in which the question was whether the respondent feels happy. It has been found that the factors determining the happiness of a human being are among others sex, income, age and education.