Ewa Genge
Uniwersytet Ekonomiczny w Katowicach

ROLA KOBIET W POLSKIM SPOŁECZEŃSTWIE – ANALIZA EMPIRYCZNA Z WYKORZYSTANIEM MODELI KLAS UKRYTYCH DLA DANYCH JAKOŚCIOWYCH

Wprowadzenie

Modele klas ukrytych (ang. *latent class models*), zwane również analizą klas ukrytych (ang. *latent class analysis*) należą do tzw. modeli ze zmiennymi ukrytymi (ang. *latent variable models*), w których ukrytą zmienną jest klasa. Modele te można zaliczyć również do tzw. podejścia modelowego w taksonomii (ang. *model-based clustering*), gdzie wykorzystywana jest idea mieszanek rozkładów (zob. Domański, Pruska, 2000; Witek, 2009). W odróżnieniu od heurystycznych metod taksonomicznych (tj. metod hierarchicznych, iteracyjno-aglomeracyjnych), w których podstawą klasyfikacji obiektów do klas są różnego rodzaju miary odległości, w podejściu modelowym obiekty klasyfikowane są na podstawie prawdopodobieństw.

Istotą modelowania klas ukrytych jest badanie związków między kategoriami zmiennych nominalnych i porządkowych. Wykorzystuje ona dane zawarte w tablicy kontyngencji. Metoda ta została wprowadzona przez Lazarsfelda (1950) w latach 50. XX w., a w kolejnych latach rozwijana przez Goodmana (1970), który przyczynił się do rozwinięcia algorytmu pozwalającego otrzymać parametry funkcji największej wiarygodności, oraz Habermana (1979), który pokazał związek pomiędzy modelami klas ukrytych oraz modelami logarytmiczno-liniowymi. Metoda ta nadal cieszy się dużym zainteresowaniem i rozwijana jest m.in. przez uczonych, takich jak Hagenaars (2002), Vermunt (2010), Linzer i Lewis (2011).

1. Model klas ukrytych – definicja

Rozważa się zbiór n obiektów, charakteryzowanych za pomocą zmiennych dychotomicznych lub politomicznych, zwanych zmiennymi obserwowanymi (ang. *manifest variables*) o wielu kategoriach $l_1, \ldots, l_m$ (zob. Bąk, 2011). Zbiór

wszystkich obiektów można więc zapisać za pomocą wektora $\mathbf{x}_i = (x_{ijh}; j = 1, \ldots, m; h = 1, \ldots, l_j; i = 1, \ldots, n)$, gdzie $x_{ijh} = 1$ oznacza *i*-tą obserwację na *j*-tej zmiennej o *h*-tej kategorii. Przyjmując, że liczba wszystkich kategorii jest równa $l = \sum_{j=1}^{m} l_j$, zbiór określany jest za pomocą macierzy o wymiarach $n \times m$.

Model klas ukrytych dla danych jakościowych można zapisać jako mieszankę rozkładów wielomianowych, w której zakłada się, że każda obserwacja $\mathbf{x}_i$ pochodzi z mieszanki wielowymiarowych rozkładów wielomianowych (ang. *mixture of multivariate multinominal distributions*), określonej jako:

$$f(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s f_s(x_i|\boldsymbol{\Theta}_s), \tag{1}$$

gdzie:

f_s – funkcja gęstości ukrytej klasy P_s, (*s*-tego rozkładu składowego mieszanki),

$\mathbf{x}_i$ – wektor realizacji zmiennych obserwowanych $\mathbf{x}_i = [x_{i1,\ldots,}x_{im_1}]$,

$\boldsymbol{\Theta}_s$ – wektor parametrów ukrytej klasy P_s,

$\boldsymbol{\Theta}$ – wektor wszystkich parametrów mieszanki rozkładów, $\boldsymbol{\Theta} = (\tau_s, \boldsymbol{\Theta}_s)$,

τ_s – prawdopodobieństwo *a priori* – wartość prawdopodobieństwa, że dana obserwacja należy do klasy P_s $(\tau_s \geq 0 \wedge \sum_{s=1}^{u} \tau_s = 1), \boldsymbol{\Theta}_\mathbf{s} \neq \boldsymbol{\Theta}_l \forall s \neq l.$

Rozkłady składowe można zaś zapisać jako:

$$f_s(\mathbf{x}_i|\boldsymbol{\Theta}_s) = \prod_{j=1}^{m} \prod_{h=1}^{l_j} (\Theta_{sjh})^{x_{ijh}}, \tag{2}$$

gdzie $\boldsymbol{\Theta}_s = (\Theta_{sjh}; j = 1, \ldots, m; h = 1, \ldots, l_j)$. Równanie (2) rozumiane jest jako iloczyn m niezależnych rozkładów wielomianowych o parametrach $\boldsymbol{\Theta}_{sj}$. Parametry mieszanki oznaczone są za pomocą wektora $\boldsymbol{\Theta} = (\tau_{1,\ldots,u}, \boldsymbol{\Theta}_{1,\ldots,u})$.

Dla danych estymatorów $\hat{\tau}_s$ i Θ_{sjh} prawdopodobieństwa *a posteriori* przynależności obiektów do poszczególnych klas mogą być obliczone za pomocą wzoru Bayes'a:

$$P(s|\mathbf{x}_i, \boldsymbol{\Theta}) = \frac{\hat{\tau}_s f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_s)}{\sum_{q=1}^{u} \hat{\tau}_q f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_q)}. \tag{3}$$

Należy zauważyć, że liczba szacowanych niezależnie parametrów modelu klas ukrytych wzrasta wraz z liczbą klas, zmiennych i ich kategorii. Liczba szacowanych parametrów wynosi $u\sum_j (l_j - 1) + (u - 1)$. Jeżeli liczba ta przekroczy liczebność zbioru lub łączną liczbę komórek w tablicy kontyngencji dla zmiennych obserwowanych, wtedy model klas ukrytych stanie się modelem nieidentyfikowalnym.

2. Model klas ukrytych z zmiennymi towarzyszącymi

Model klas ukrytych oprócz zmiennych obserwowanych może zawierać jeszcze tzw. zmienne towarzyszące (ang. *covariates* lub *concomitant variables*), mające wpływ na przynależność obiektów do klas – wpływ na prawdopodobieństwa *a priori* (zob. np. Dayton i Macready, 1988; Hagenaars i McCutcheon, 2002). Zmienne towarzyszące wraz ze zmiennymi $X_1, \ldots, X_m$ biorą udział w szacowaniu parametrów modelu klas ukrytych, na podstawie którego będzie można dokonać klasyfikacji nowych obiektów bez udziału zmiennych obserwowanych. Zmienne towarzyszące wykorzystywane są często w badaniach marketingowych, ekonomicznych, psychologicznych, w których pozyskanie zmiennych obserwowanych jest bardzo kosztowne (por. Witek, 2011).

Najczęściej parametry zmiennych towarzyszących szacowane są wraz z pozostałymi parametrami modelu klas ukrytych (jednocześnie). Ten sposób estymacji zwany jest jednokrokową techniką estymacji parametrów zmiennych towarzyszących (ang. *one-step technique for estimating the effects of covariates*) (zob. np. Dayton i Macready 1988; Hagenaars i McCutcheon, 2002). Alternatywnym sposobem estymacji parametrów zmiennych towarzyszących jest tzw. podejście trzykrokowe (ang. *three-step approach*), w którym szacowane są parametry klasycznego modelu klas ukrytych (1), następnie obliczane są prawdopodobieństwa *a posteriori* (3). W kroku trzecim szacowane są parametry równania regresji, gdzie prawdopodobieństwa te traktowane są jako zmienne zależne, a zmienne towarzyszące jako zmienne objaśniające. Jednakże Bolck, Crown i Hagenaars (2004) udowodnili, że w wyniku szacunku parametrów trzykrokową metodą estymacji, estymatory parametrów takiego modelu są obciążone.

Włączając do modelu klas ukrytych zmienne towarzyszące, zakładamy, że mają one wpływ na prawdopodobieństwa *a priori*. W klasycznym modelu klas ukrytych (bez zmiennych towarzyszących) zakładamy, że każda obserwacja ma takie samo prawdopodobieństwo przynależności do klasy ukrytej.

W przypadku gdy zmienne towarzyszące mają wpływ na prawdopodobieństwa przynależności obiektów do klas (τ_s), model klas ukrytych zapisać można jako:

$$f(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\Theta}) = \sum_{s=1}^{u} \tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) f_s(x_i | \boldsymbol{\Theta}_s), \tag{4}$$

gdzie: $\mathbf{z}_i$ – wektor realizacji zmiennych towarzyszących, $\mathbf{z}_i = [z_{i1,\ldots,}x_{im_2}]$.

Nadal jednak spełniony musi być warunek, że $(\tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) \geq 0 \wedge \sum_{s=1}^{u} \tau_s(\mathbf{z}_i, \boldsymbol{\alpha}) = 1), \boldsymbol{\Theta}_\mathbf{s} \neq \boldsymbol{\Theta}_l \forall s \neq l.$ Wpływ zmiennych towarzyszących na prawdopodobieństwa *a priori* wyrażany jest za pomocą wielomianowej funkcji logitowej (zob. Agresti, 2002).

Jeżeli w szacowaniu parametrów modelu klas ukrytych biorą udział zmienne towarzyszące, zazwyczaj pierwsza z klas jest tzw. klasą referencyjną. Zakłada się wtedy, że iloraz szans prawdopodobieństw *a priori* dla klas ukrytych, w porównaniu do tej klasy (klasy referencyjnej) jest liniową funkcją zmiennych towarzyszących. Dla m_2 zmiennych towarzyszących, wektor parametrów tych zmiennych $\boldsymbol{\alpha}_s$ ma długość $m_2 + 1$ (dla każdej zmiennej towarzyszącej i wyrazu wolnego). Ponieważ pierwsza klasa jest klasą referencyjną, z definicji $\boldsymbol{\alpha}_1 = 0$. Wtedy:

$$\ln(\tau_{2i} / \tau_{1i}) = \mathbf{z}_i \boldsymbol{\alpha}_2 \tag{5}$$

$$\ln(\tau_{3i} / \tau_{1i}) = \mathbf{z}_i \boldsymbol{\alpha}_3 \tag{6}$$

$$\vdots$$

$$\ln(\tau_{ui} / \tau_{1i}) = \mathbf{z}_i \boldsymbol{\alpha}_u \tag{7}$$

W wyniku kilku przekształceń otrzymujemy:

$$\tau_{si} = \tau_s(\mathbf{z}_i; \boldsymbol{\alpha}) = \frac{e^{\mathbf{z}_i \boldsymbol{\alpha}_s}}{\sum_{q=1}^{u} e^{\mathbf{z}_i \boldsymbol{\alpha}_q}}. \tag{8}$$

W modelu klas ukrytych z udziałem zmiennych towarzyszących, szacowanych jest więc $u-1$ wektorów $\boldsymbol{\alpha}_s$, a także warunkowych prawdopodobieństw przynależności obiektów do klas ukrytych. Mając dane estymatory $\hat{\boldsymbol{\alpha}}_s$ i Θ_{sjh}, prawdopodobieństwa *a posteriori* i przynależności obiektów do klas uzyskiwane są poprzez zastąpienie τ_s w równaniu (3) funkcją $\tau_s(\mathbf{z}_i; \boldsymbol{\alpha})$ z równania (8):

$$P(s | \mathbf{x}_i, \mathbf{z}_i) = \frac{\hat{\tau}_s(\mathbf{z}_i; \hat{\boldsymbol{\alpha}}) f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_s)}{\sum_{q=1}^{u} \hat{\tau}_q(\mathbf{z}_i; \hat{\boldsymbol{\alpha}}) f(\mathbf{x}_i, \hat{\boldsymbol{\Theta}}_q)}. \tag{9}$$

Liczba szacowanych parametrów takiego modelu klas ukrytych jest równa $u\sum_j (l_m - 1) + (s+1)(u-1)$.

3. Estymacja parametrów

Estymacja modelu klas ukrytych polega m.in. na oszacowaniu liczby i wielkości poszczególnych klas. Metodą największej wiarygodności szacowane są parametry modelu klas ukrytych (4). Funkcja największej wiarygodności określona jest wzorem:

$$\ln L = \sum_{i=1}^{n} \ln \sum_{s=1}^{u} \tau_s(\mathbf{z}_i;\boldsymbol{\alpha}) \prod_{j=1}^{m} \prod_{h=1}^{l_j} (\Theta_{sjh})^{x_{ijh}} . \tag{10}$$

Popularną metodą szacowania parametrów największej wiarygodności jest algorytm EM (Dempster et al., 1977). W pakiecie poLCA wykorzystywana jest zmodyfikowana wersja algorytmu EM (zob. Bandeen-Roche et al., 1977). Proces estymacji zapoczątkowany jest przez wartości startowe dla $\widehat{\boldsymbol{\alpha}}'_s$ i Θ'_{sjh}, dzięki którym wyznaczone są prawdopodobieństwa *a posteriori* $P(s|\mathbf{x}_i, \mathbf{z}_i)$ dane wzorem (9). Parametry zmiennych towarzyszących szacowane (i uaktualniane) są zgodnie z formułą:

$$\widehat{\boldsymbol{\alpha}}_s = \widehat{\boldsymbol{\alpha}}'_s + (-\mathbf{D}_\alpha^2 \log L)^{-1} \mathbf{D}_\alpha \log L, \tag{11}$$

gdzie $\widehat{\boldsymbol{\alpha}}'_s$ to wektor estymatorów parametrów zmiennej towarzyszącej, $\mathbf{D}_\alpha$ to gradient, zaś $\mathbf{D}_\alpha^2$ hesjan macierzy z parametrem $\boldsymbol{\alpha}$. Nowe wartości parametrów Θ_{sjh} wyznaczane są za pomocą formuły:

$$\Theta_{sj} = \frac{\sum_{i=1}^{n} \mathbf{x}_{ij} P(s|\mathbf{x}_i, \mathbf{z}_i)}{\sum_{i=1}^{n} P(s|\mathbf{x}_i, \mathbf{z}_i)}. \tag{12}$$

Kroki algorytmu powtarzane są dopóty, dopóki przyrost funkcji wiarygodności nie będzie mniejszy niż zadana wartość graniczna lub nie zostanie osiągnięta maksymalna liczba iteracji. Wzory oraz szczegółowe informacje dotyczące gradientu $\mathbf{D}_\alpha$ oraz hesjanu $\mathbf{D}_\alpha^2$ można znaleźć w pracy Bandeen-Roche et al. (1997).

4. Wybór modelu i ocena jakości dopasowania

Jedną z głównych zalet modeli klas ukrytych jest to, że w odróżnieniu od popularnych metod taksonomicznych (tj. k-średnich, metody Warda), istnieje kilka statystycznych miar służących wyborowi i ocenie ich jakości dopasowania. Najczęściej w różnego rodzaju badaniach empirycznych na początku sprawdza się dopasowanie dla $s = 1$. W kolejnych krokach zwiększa się liczbę klas o jeden, tak długo aż model osiągnie najlepsze dopasowanie. Należy jednak pamiętać, że wraz z dodatkową liczbą klas, liczba szacowanych parametrów wzrasta o $1 + \sum_j (l_j - 1)$, dlatego najczęściej wykorzystywane są kryteria informacyjne, będące wyrazem kompromisu pomiędzy jakością dopasowania a złożonością modelu. Do najbardziej popularnych kryteriów informacyjnych zaliczane są: Bayesowskie kryterium informacyjne Schwarza BIC (*Bayesian Information Criterion*), kryterium informacyjne Akaike AIC (*Akaike Information Criterion*). Kryteria te mogą dawać niejednoznaczne wskazania co do oceny modeli klas ukrytych.

Istnieje kilka formuł zapisu wspomnianych kryteriów oceny dopasowania modeli klas ukrytych. W pakietach programu **R** najczęściej wykorzystywane są kryteria podlegające minimalizacji. Można je przestawić na pomocą następujących wzorów:

$$BIC_s = -2 \log P(\mathbf{x}_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + v_s \log(n), \tag{13}$$

$$AIC_s = -2 \log P(\mathbf{x}_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.) + 2v_s, \tag{14}$$

gdzie:

$\log P(\mathbf{x}_i \left| \hat{\boldsymbol{\Theta}}_s, M_s \right.)$ – logarytm funkcji wiarygodności dla oszacowanego wektora parametrów modelu,

M_s, v_s – liczba parametrów modelu,

n – liczba obserwacji.

Pierwsza część powyższych równań odpowiada za wybór modeli o najwyższej dobroci dopasowania, zaś część druga odrzuca modele z nadmierną liczbą parametrów. Porównania różnych kryteriów informacyjnych można znaleźć m.in. w pracach: McLachlan i Peel (2000), Biernacki et al. (1999), Bozdogan (2000). W części empirycznej pracy wykorzystano dwa najbardziej popularne kryteria, tj. BIC oraz AIC. Kryteria te stosowane są w celach porównawczych modeli o różnej liczbie klas. Im niższa wartość kryteriów, tym lepsza jakość dopasowania danego modelu.

5. Analiza empiryczna

Analizę klas ukrytych przeprowadzono na podstawie danych uzyskanych z bezpłatnej bazy danych Polskiego Generalnego Sondażu Społecznego (PGSS) 1992-2008*. W niniejszym artykule rozważano dane z 2008 r. Analiza została przeprowadzona z uwzględnieniem sześciu zmiennych i z pominięciem odpowiedzi „nie wiem" („trudno powiedzieć"). Badana próba liczyła 986 osób.

W przykładzie wykorzystano sześć zmiennych obserwowanych $X_1 - X_6$. W nawiasie podano oryginalne nazwy ze zbioru PGSS 2008.

1. X_1 (q5): Kobiety nie nadają się do polityki (1 – zgadzam się; 2 – nie zgadzam się);
2. X_2 (q6): Rządzenie krajem pozostawić mężczyznom (1 – zgadzam się; 2 – nie zgadzam się);
3. X_3 (q7a): Pracująca matka może zapewnić ciepło (1 – zgadzam się; 2 – nie zgadzam się);
4. X_4 (q7b): Żona niech zapewni mężowi karierę (1 – zgadzam się; 2 – nie zgadzam się);
5. X_5 (q7c): Praca matki szkodzi dziecku (1 – zgadzam się; 2 – nie zgadzam się)
6. X_6 (q7d): Lepiej gdy mężczyzna zarabia/kobieta w domu (1 – zgadzam się; 2 – nie zgadzam się).

Uwzględniono również następujące zmienne towarzyszące:

a) Z_1: płeć respondenta (1 – mężczyzna, 2 – kobieta);
b) Z_2: stan cywilny: kawaler, konkubinat, żonaty, rozwiedziony, separacja, wdowiec;
c) Z_3: wykształcenie: zawodowe (niepełne podstawowe, podstawowe, zasadnicze zawodowe), średnie (niepełne średnie, średnie ogólnokształcące, średnie zawodowe, policealne/pomaturalne, nieukończone studia wyższe), wyższe (ukończone studia licencjackie, ukończone studia magisterskie).

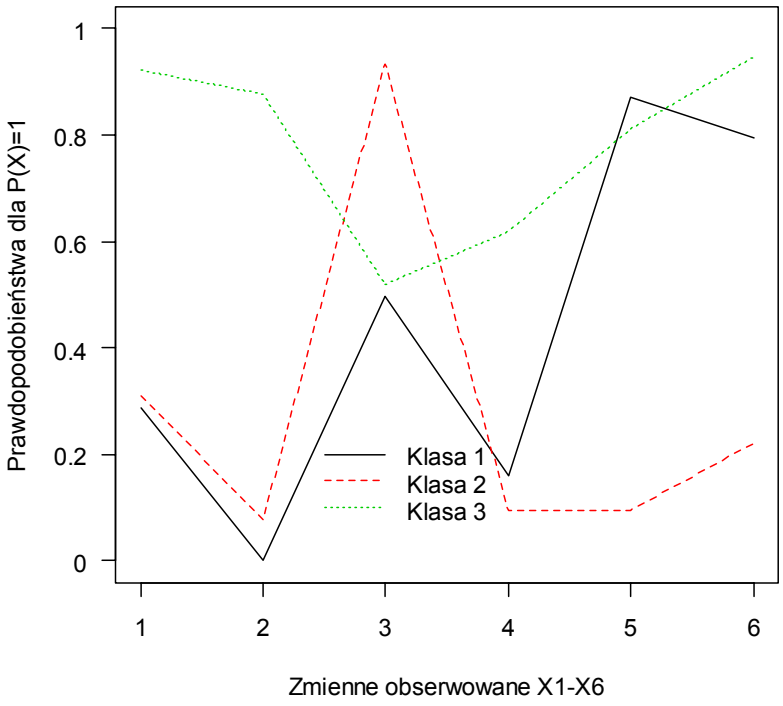
W badaniach wykorzystano pakiet poLCA programu **R**.

Aby wybrać optymalną liczbę klas ukrytych (ukrytą liczbę składowych modelu), obliczono wartości kryteriów informacyjnych AIC oraz BIC dla liczby klas $s = 1, \ldots, u$ dla tzw. modelu podstawowego, tj. bez udziału zmiennych towarzyszących (ang. *base model*), (zob. np. Collins i Lanza, 2011). W przypadku analizowanego zbioru danych kryteria wskazały minimalną wartość dla liczby klas równej cztery. Niewiele większą wartość otrzymano dla trzech klas. W takich sytu-

* Dane dostępne na stronie: http://pgss.iss.uw.edu.pl.

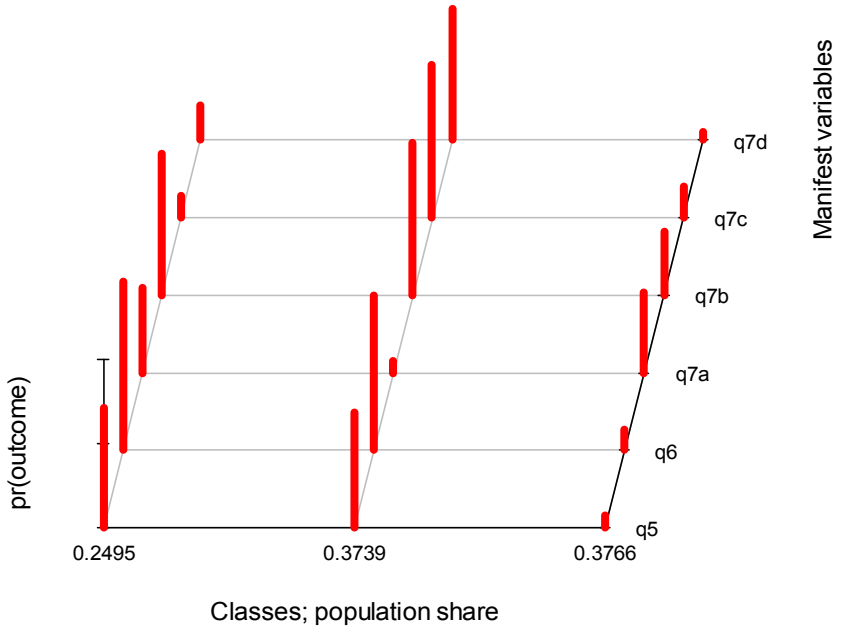
acjach często wybierane są modele mniej złożone (zob. np. Collins i Lanza, 2011), dlatego też w dalszej części pracy analizowano model o trzech klasach ukrytych.

Następnie szacowano modele klas ukrytych dla zmiennych $X_1 - X_6$ i różnych zestawach zmiennych towarzyszących (np. $Z_1 + Z_2$, $Z_1 + Z_3$). Rozważano również interakcje pomiędzy zmiennymi towarzyszącymi, ale wszystkie z nich okazały się nieistotne. Na podstawie analizy przeprowadzonych obliczeń (analiza kryteriów informacyjnych oraz badania istotności parametrów za pomocą testu t-Studenta) przyjęto ostateczny podział badanej próby respondentów na trzy klasy, z wykorzystaniem trzech zmiennych towarzyszących. Dla wybranego modelu przedstawiono prawdopodobieństwa przyjmowania przez zmienne obserwowane wartości 1 („zgadzam się") w klasie pierwszej, drugiej i trzeciej (rys. 1).



Rys. 1. Prawdopodobieństwo wyboru wartości 1 dla zmiennych $X_1 - X_6$

Na rys. 2 przedstawiono prawdopodobieństwa wyboru pierwszej kategorii dla zmiennych $X_1 - X_6$ (odpowiedź na „tak") dla każdej z klas. Wysokość słupków oznacza prawdopodobieństwa odpowiedzi „tak/zgadzam się". Widoczne są także prawdopodobieństwa *a priori* (wagi) dla poszczególnych klas.



Rys. 2. Wyniki segmentacji respondentów

W klasie pierwszej, najmniej licznej ($\tau_1 = 0{,}25$), 28% respondentów twierdzi, że kobiety nie nadają się do polityki. Bardzo mały procent (0,07%) w tej klasie stanowią osoby zgadzające się z tym, że rządzenie krajem należy pozostawić mężczyznom. Prawie 50% zgadza się z opinią, że pracująca matka może zapewnić ciepło. 16% twierdzi, że żona jest odpowiedzialna za karierę męża. Największy odsetek w tej grupie (87%) stanowią respondenci przekonani, że praca matki szkodzi dziecku. Niewiele mniej (79%) respondentów uważa, że lepiej, gdy zarabia mężczyzna.

Klasa druga jest klasą liczniejszą – należy do niej 37% wszystkich ankietowanych. W klasie tej 31% respondentów uważa, że kobiety nie nadają się do polityki, a 7% zgodziło się z opinią, że rządzenie krajem należy pozostawić mężczyznom. W klasie drugiej jest największy (w porównaniu z klasą pierwszą i trzecią) udział osób (93%), które sądzą, że pracująca matka może zapewnić ciepło. Tylko 9% ankietowanych uważa, że żona powinna zapewnić karierę mężowi. Taki sam procent stanowią osoby, które twierdzą, że praca matki szkodzi dziecku. W klasie tej jest najmniej osób (w porównaniu do klasy pierwszej i trzeciej), tj. 22%, które sądzą, iż lepiej jest, gdy o utrzymanie rodziny troszczy się mężczyzna.

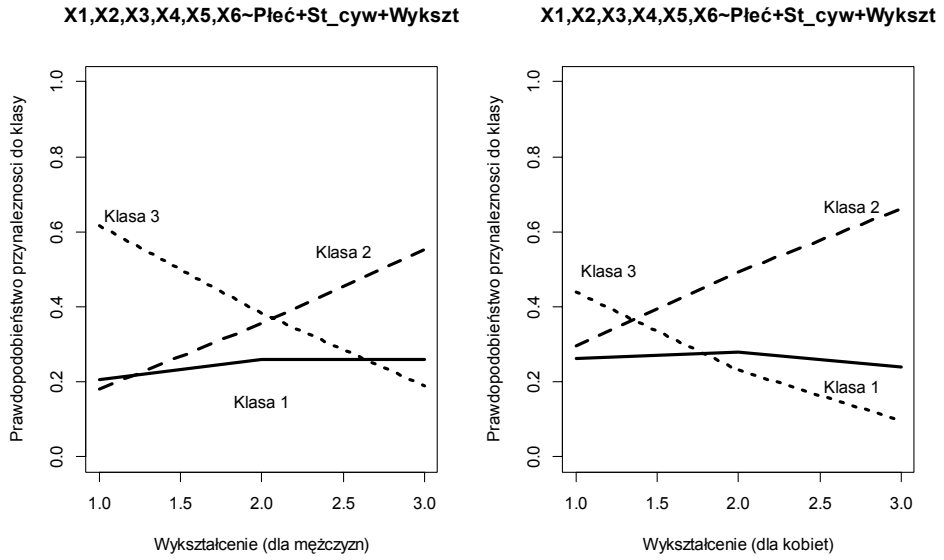
Klasa trzecia jest tak samo liczna, jak klasa druga (τ_3 = 0,37). Ponad 90% osób zgadza się z opinią, że kobiety nie nadają się do polityki. Nieco mniej (87%) uważa, że rządzenie krajem należy pozostawić mężczyznom. Ponad połowa ankietowanych tej klasy jest zdania, że pracująca kobieta może zapewnić rodzinne ciepło, ale na pytanie: „Czy praca matki szkodzi dziecku?” aż 81%

odpowiedziało twierdząco. W klasie tej aż 95% osób uważa, że lepszym rozwiązaniem dla rodziny jest tylko zawodowa praca mężczyzny.

W kolejnej części pracy dokonano analizy wpływu zmiennych towarzyszących na przynależność analizowanych obiektów do klas. Jeżeli chodzi o zmienną „płeć", okazuje się, że dla mężczyzn występuje najwyższe prawdopodobieństwo przynależności do klasy trzeciej, a najniższe w przypadku klasy drugiej. Z kolei udział kobiet w klasie drugiej jest najwyższy i wynosi prawie 50%, kolejno w klasie pierwszej oraz trzeciej.

Dokonując analizy wpływu zmiennej towarzyszącej „stan cywilny" (dla mężczyzn z średnim wykształceniem), prawdopodobieństwo przynależności do klasy pierwszej jest prawie takie samo dla osób o różnym stanie cywilnym. W klasie drugiej największe prawdopodobieństwo występuje w przypadku kawalerów, następnie panów żyjących w konkubinacie oraz żonatych (najniższe dla wdowców). Prawdopodobieństwo przynależności do klasy trzeciej („konserwatywnej") jest najwyższe dla wdowców, następnie osób żyjących w separacji i rozwiedzionych.

Jeżeli chodzi o zmienną towarzyszącą „wykształcenie", to dla mężczyzn, żonatych o wykształceniu zawodowym, najwyższe jest prawdopodobieństwo przynależności do klasy trzeciej. Prawdopodobieństwo przynależności do tej klasy spada wraz z lepszym wykształceniem respondentów. Z kolei prawdopodobieństwo przynależności do klasy drugiej wzrasta wraz z lepszym wykształceniem. Jeśli chodzi o klasę pierwszą, to prawdopodobieństwo przynależności do tej klasy jest prawie takie samo dla osób o różnym poziomie wykształcenia. Wpływ wykształcenia na przynależność do klas dla kobiet jest bardzo podobny (rys. 3). Ze względu na ograniczenia objętościowe na rys. 3 zamieszczono tylko wykres dla zmiennej towarzyszącej Z_3 (wykształcenie).



Rys. 3. Wykres przynależności kobiet (strona lewa) i mężczyzn (strona prawa) do trzech klas

Dla zmiennej towarzyszącej „wykształcenie" sporządzono oddzielne wykresy dla kobiet i mężczyzn, przyjmując, że zmienne jakościowe są równe kategorii występującej najczęściej (stan cywilny – zamężna/żonaty). W podobny sposób sporządzono wykresy i dokonano interpretacji dla zmiennej towarzyszącej „płeć" i „stan cywilny" (zob. np. Linzer i Lewis, 2011; Witek, 2011).

Podsumowanie

W artykule przedstawiono przykład zastosowania modeli klas ukrytych do oceny roli kobiet w polskim społeczeństwie. Analiza klas ukrytych umożliwiła segmentację respondentów na podstawie odpowiedzi udzielonych w badaniu Polskiego Generalnego Sondażu Społecznego. Wyodrębniono trzy klasy o podobnych wzorcach zachowań i postaw dla polskich respondentów. Dokonano również oceny wpływu zmiennych demograficznych na ich przynależność do klas.

Do klasy pierwszej zaliczono najmniej osób przeciwnych temu, by kobiety zajmowały się polityką (zarówno jeśli chodzi pełnienie różnych funkcji politycznych, jak i rządzenie krajem). W przypadku pracy zawodowej panuje tu raczej przekonanie, by kobieta została w domu. Respondenci klasy drugiej są przekonani, że kobiety jak najbardziej powinny realizować się zawodowo, a rodzina na tym nie ucierpi. Nie mają również przeciwwskazań, by kobiety pełniły funkcje polityczne. Klasa trzecia jest klasą osób „konserwatywnych", będących zdania, że kobieta po prostu powinna przebywać w domu (ani nie pracować, ani nie angażować się w życie polityczne naszego kraju).

Bibliografia

Agresti A. (2002): *Categorical Data Analysis*. John Wiley & Sons, Hoboken.

Bandeen-Roche K., Miglioretti D.L., Zeger S.L., Rathouz P.J. (1997): *Latent Variable Regression for Multiple Discrete Outcomes*. „Journal of the American Statistical Association", No. 92(40), s. 123-135.

Bąk A. (2011), *Modele klas ukrytych dla danych jakościowych*. W: *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*. Red. E. Gatnar, M. Walesiak. C.H. Beck, Warszawa, s. 204-222.

Biernacki C., Celeux G., Govaert G. (1999): *Choosing Models in Model-Based Clustering and Discriminant Analysis*. „Journal of Statistical Computation and Simulation", No. 64, s. 49-71.

Bolck A., Croon M., Hagenaars J. (2004): *Estimating Latent Structure Models with Categorical Variables: One-step Versus Three-step Estimators*. „Political Analysis", No. 12(1), s. 3-27.

Bozdogan H. (2000): *Akaike's Information Criterion and Recent Developments in Information Criterion*. „Journal of Mathematical Psychology", No. 44, s. 62-91.

Collins L.M., Lanza S.T. (2011): *Latent Class and Latent Transition Analysis with Applications in the Social, Behavioral, and Health Sciences*. John Wiley & Sons, Wiley, s. 100-103; 151, 177.

Dayton C. M., Macready G.B. (1988): *Concomitant-variable Latent-class Models*. „Journal of the American Statistical Association", No. 83(401), s. 173-178.

Dempster A.P., Laird N.P., Rubin D.B. (1977): *Maximum Likelihood for Incomplete Data Via the EM Algorithm (with discussion)*. „Journal of the Royal Statistical Society", No. 39, ser.B, s. 1-38.

Domański C., Pruska K. (2000): *Nieklasyczne metody statystyczne*. PWE, Warszawa.

Goodman L. (1970): *The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classification*. „Journal of the American Statistical Association", No. 65, s. 226-256.

Haberman S.J. (1979): *Analysis of Qualitative Data, New Developments*. Academ ic Press, New York, No 2.

Hagenaars A.J., McCutcheon A.L. (2002): *Applied Latent Class Analysis*. Cambridge University Press, Cambridge.

Lazarsfeld P.F. (1950): *The Logical* and Mathematical Foundations of Latent Structure Analysis. W: *Measurement and Prediction*. Red. S.A. Stouffer. John Wiley & Sons, New York, s. 362-412.

Linzer D., Lewis J. (2011): *poLCA: An R Package for Polytomous Variable Latent Class Analysis*. „Journal of Statistical Software", No. 42(10), s. 1-29.

McLachlan G.J., Peel D. (2000): *Finite Mixture Models*. Wiley, New York, s. 81-116.

Vermunt, J.K. (2010): *Latent Class Modeling With Covariates: Two Improved Three-step Approaches*. Political Analysis, 18, s. 450-469.

Witek E. (2009): *Analiza skupień – podejście modelowe*. W: *Statystyczna analiza danych z wykorzystaniem programu R*. Red. M. Walesiak, E. Gatnar. Wydawnictwo Naukowe PWN, Warszawa, s. 434-462.

Witek E. (2011): *Modele mieszanek dla danych jakościowych*. W: *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*. Red. E. Gatnar, M. Walesiak. C.H. Beck, Warszawa, s. 223-241.

A ROLE OF WOMEN IN POLISH SOCIETY – AN EMPIRICAL ANALYSIS WITH THE USE OF LATENT CLASS MODELS

Summary

The paper focuses on latent class models and it's application for quantitative data. Latent class modeling is one of a multivariate analysis techniques of the contingency table and can be viewed as a special case of model-based clustering, for multivariate discrete data. It is assumed that each observation comes from one of a number of subpopulations, with its own probability distribution.

We used latent class analysis for grouping and detecting inhomogeneities of Polish opinions on role of women in polish society. We analyzed data collected as part of the Polish General Social Survey (GSS) using poLCA package of R.