PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1140 — 2006
Praktyka statystyki

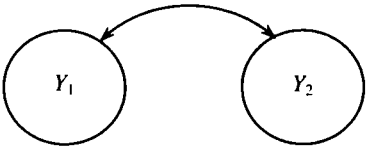
Anna Bartkowiak

Akademia Ekonomiczna we Wrocławiu

MODEL POMIARU CECHY UKRYTEJ NA PODSTAWIE WSKAŹNIKÓW DYCHOTOMICZNYCH

1. Wstęp

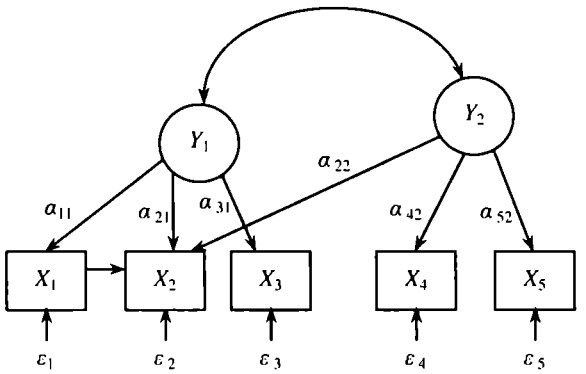
Badania ekonomiczne czy społeczne dotyczą wielu różnych zjawisk. Zjawiska te mogą mieć charakter ilościowy lub jakościowy. Wśród cech jakościowych pojawiają się cechy, które posiadają pewne kategorie, np. w lokalizacji firmy można wyróżnić następujące kategorie: centrum miasta, peryferie miasta, poza miastem. Poza takimi cechami, których określenie kategorii lub ilości jest możliwe od razu ze względu na ich charakter, występują również cechy, dla których to określenie nie jest takie oczywiste i w związku z tym nie można ich bezpośrednio zmierzyć. Do takich cech należą np. satysfakcja, frustracja, poziom ekonomiczny, atrakcyjność, zadowolenie z usługi czy produktu. Cechy bezpośrednio niemierzalne nazywa się ukrytymi lub latentnymi. Cechy te mogą być również powiązane między sobą. Wyróżnia się trzy typy powiązań: przyczynowe (jedna cecha wpływa na drugą, np. frustracja może wpływać na agresję); cykliczne (obie cechy wpływają na siebie, np. niezadowolenie z życia wpływa na depresję, a większa depresja pogłębia niezadowolenie); oraz nieprzyczynowe. Wektor zmiennych latentnych wraz z ich powiązaniami nazywamy strukturą ukrytą. Przykłady powiązań między dwoma zmiennymi ukrytymi Y_1 i Y_2 przedstawiono na rys. 1.



Rys. 1. Powiązanie nieprzyczynowe zmiennych ukrytych

Źródło: opracowanie własne.

Do analizy takich cech wykorzystuje się modele struktur ukrytych (latentnych) ze zmiennymi obserwowalnymi. W związku z tym, że cechy latentne nie są bezpośrednio mierzalne, należy znaleźć pewne wskaźniki mierzalne, które będą odzwierciedlać te cechy. Zmienne obserwowalne są wskaźnikami zmiennych ukrytych, tzn. zmienne latentne wpływają na zmienne obserwowalne. Oczywiście wskaźniki nie są wybierane losowo. Wybór ten oparty jest na wcześniej określonej przez badacza definicji cechy ukrytej. Liczba zmiennych obserwowalnych jest zatem z góry określona przez badacza i się nie zmienia. Zmienne obserwowalne oznaczmy przez X, a wpływ zmiennych ukrytych na zmienne obserwowalne przez α_{ij}, gdzie pierwszy indeks i oznacza numer zmiennej obserwowalnej, a j – numer zmiennej ukrytej, która na nią wpływa. Błąd pomiarowy zmiennej obserwowalnej przyjęto oznaczać przez ε. Przykład grafu modelu dwóch nieprzyczynowo powiązanych zmiennych latentnych wraz ze zmiennymi obserwowalnymi oraz błędami pomiarowymi przedstawiono na rys. 2.



Rys. 2. Model struktury ukrytej ze zmiennymi obserwowalnymi

Źródło: opracowanie własne.

Znanych jest kilka metod analizy takich modeli [Ballen 1989; Moosmuller 1989]. Jedna z metod analizy modelu opiera się na powiązaniu funkcyjnym między zmiennymi.

Zmienne obserwowalne są powiązane ze zmiennymi ukrytymi, zatem w celu analizy tych powiązań oraz pomiaru zmiennych latentnych na podstawie wskaźników należy znaleźć powiązanie funkcyjne rozkładów obu typów zmiennych. Wykorzystuje się funkcję gęstości wektora losowego $\underline{X}$, którego rozkład jest znany jako brzegowa gęstość rozkładu łącznego wektorów $\underline{X}$ i $\underline{Y}$:

$$f_{\underline{X}}(\underline{x}) = \int_{D\underline{Y}} f_{\underline{X}/\underline{Y}}(\underline{x}/\underline{y}) g_{\underline{Y}}(\underline{y}) d\underline{y}.$$

Funkcja gęstości wektora losowego $\underline{Y}$ jest oczywiście nieznana. W omawianym modelu zmienne ukryte Y_i są niezależne i są zmiennymi losowymi typu ciągłego, określone na pewnym ograniczonym odcinku, skali, dlatego wybiera się dla funkcji $g(y_i)$ funkcję rozkładu jednostajną na [0; 1].

W przedstawianym modelu zakłada się, że

$$Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j, Cov(Y_i, \varepsilon_i) = 0, E(\varepsilon_i) = 0, i = 1,...,n.$$

Rozpatruje się również i takie modele, w których nie zakłada się braku korelacji między błędami. Jednakże w omawianym tu modelu takie założenie jest niezbędne.

Jeśli błędy są skorelowane, występuje powiązanie między zmiennymi obserwowalnymi. Oznacza to, że istnieją inne zmienne ukryte, które wpływają na dane zmienne obserwowalne, i takie, które takie powiązanie usuwają. Zatem należy znaleźć najmniejszą liczbę zmiennych ukrytych tak, by nie występowały powiązania między zmiennymi obserwowalnymi. Formalnie oznacza to, że należy znaleźć m takie, by spełniony był warunek

$$f_{\underline{X}/\underline{Y}}(\underline{x}/\underline{y}) = \prod_{i=1}^{n} f_{X_i/\underline{Y}}(x_i/\underline{y}), \ \underline{y} = (y_1, y_2,..., y_m). \qquad (1)$$

Wyrażenie to jest znane jako warunkowa lub lokalna niezależność, która tkwi u podstaw wszystkich modeli struktur ukrytych. Oznacza to, że zależność wśród X_i jest całkowicie wyrażona przez ich zależność z Y_i. Uwzględniając wyrażenie (1) oraz założenie, że $g(y_i)$ jest funkcją gęstości rozkładu jednostajnego na [0; 1], otrzymuje się

$$f_{\underline{X}}(\underline{x}) = \int_0^1 ... \int_0^1 \prod_{i=1}^{n} f_{X_i/\underline{Y}}(x_i/\underline{y}) d\underline{y} = E\prod_{i=1}^{n} f_{X_i/\underline{Y}}(x_i/\underline{y}). \qquad (2)$$

Dalsza konstrukcja modelu obejmuje właściwe dobranie postaci funkcji $\{ f_{X_i/\underline{Y}}(x_i/\underline{y}) \}$.

W dalszych rozważaniach zmienne obserwowalne są zmiennymi binarnymi, tzn. $x_i = 1$ lub zero. Jeśli zmienne X_i są zmiennymi losowymi o rozkładzie Bernoulliego, to dla $i = 1,..., n$ mamy

$$f_{X_i/\underline{Y}}(x_i/\underline{y}) = \{\pi_i(\underline{y})\}^{x_i} \{1 - \pi_i(\underline{y})\}^{1-x_i}.$$

Na przykład funkcja $\pi_i(\underline{y})$ niech określa prawdopodobieństwo udzielenia pozytywnej odpowiedzi ($x_i = 1$) przez respondenta, którego ukryta postawa jest dana

przez $\underline{y}$. Funkcja ta nazywa się *funkcją odpowiedzi.* Problemem jest teraz wybór odpowiedniej parametrycznej postaci funkcji $\pi_i(\underline{y})$. Wybór ten odróżnia jeden model od drugiego.

W literaturze (por. [Bartholomew 1980; 1983]) podaje się pewne własności, które funkcja $\pi_i(\underline{y})$ powinna posiadać. Wartości tej funkcji należą do przedziału [0; 1], gdyż jest ona prawdopodobieństwem. Funkcja ta powinna także być niezmienna ze względu na permutacje zmiennych, co odzwierciedla przypadkowość ustawienia, w którym zmienne ukryte są mierzone. Spektrum polityczne można mierzyć np. od lewej do prawej lub odwrotnie. Istotny jest również warunek pozwalający na przypadkowość uporządkowania kategorii. Podaje się także mniej formalne własności, takie jak np., by liczba parametrów funkcji $\pi_i(\underline{y})$ była mała oraz by wartość oczekiwana (2) była prosta do określenia.

Wyrażenie funkcji $\pi(\underline{y})$ jako funkcji liniowej parametrów nie jest możliwe, gdyż nie spełnia ona wówczas warunku (*i*). Zamiast tego tworzy się klasę funkcji następującej postaci:

$$F\{\pi_i(\underline{y})\} = \alpha_{i0} + \sum_{j=1}^{m} \alpha_{ij} G(y_j), (i = 1, 2, ..., n). \tag{3}$$

W praktyce wybór tych funkcji jest ograniczony. Powszechnie jako funkcje F i G stosuje się funkcje logitowe $(F(v) = \text{logit}(v) = \log v/(1-v))$ lub probitowe $(F(v) = \text{profit}, (v) = \Phi^{-1}(v)$, gdzie Φ jest dystrybuantą standardowego rozkładu normalnego).

We wzorze (3) parametr α_{i0} określa prawdopodobieństwo odpowiedzi dla *i*-tej zmiennej jawnej. Jeżeli funkcje F i G są funkcjami logitowymi, to parametr α_{i0} wynosi

$$\alpha_{i0} = \text{log}it\pi_i, \text{ gdzie } \pi_i = \pi(1/2,1/2,...,1/2).$$

Stąd otrzymujemy $\pi_i = (1 + e^{-\alpha_{i0}})^{-1}$. Ponadto z wzoru (3) otrzymuje się:

$$\begin{aligned} \text{logit}\,\pi_i(\underline{y}) &= \log\{\pi_i(\underline{y})/(1-\pi_i(\underline{y}))\} = \\ &= \log \pi_i/(1-\pi_i) + \sum_{j=1}^{m} \alpha_{ij} \log y_j/(1-y_j) \end{aligned} \tag{4a}$$

lub

$$\pi_i(\underline{y}) = \pi_i \prod_{j=1}^{m} y_j^{\alpha_{ij}} / \{\pi_i \prod_{j=1}^{m} y_j^{\alpha_{ij}} + (1-\pi_i \prod_{j=1}^{m} (1-y_j)^{\alpha_{ij}}\}. \quad (4b)$$

Nazywane jest to modelem logitowym.

Parametr π_i ma użyteczną interpretację. Określa on wartość funkcji $\pi_i(\underline{y})$ dla $y_j = 1/2$ dla każdego j. Tak więc jest prawdopodobieństwem pozytywnej odpowiedzi dla indywidualności, która zajmuje środkową pozycję.

Z wzorów (4ab) otrzymuje się n parametrów π_i oraz mn parametrów α_{ij}, które należy wyestymować. Jedną z metod estymacji tych parametrów jest metoda największej wiarygodności. Dla n zmiennych binarnych dane mogą być przedstawione w tabeli kontyngencji o 2^n komórkach, w których umieszczone są całkowite częstości. Metoda największej wiarygodności jest stosowana do wszystkich komórek częstości.

W kolejnym etapie analizy modelu należy określić dobroć dopasowania modelu, tzn. czy rozkład empiryczny jest zgodny z dopasowanym rozkładem teoretycznym. W celu weryfikacji doboru dopasowania modelu stosuje się statystykę

$$\Lambda = 2\sum_i O_i \ln(O_i / E_i), \quad (5)$$

gdzie O_i i E_i są zaobserwowanymi i oczekiwanymi częstościami odpowiednio oraz sumowanie odbywa się po wszystkich komórkach tabeli. Dla tabeli kontyngencji o 2^n komórkach wartość oczekiwana (2) jest następująca:

$$f(\underline{x}) = \int_0^1 \cdots \int_0^1 \left(\pi_i(\underline{y})\right)^{x_i} \left(1-\pi_i(\underline{y})\right)^{1-x_i} \cdots \left(\pi_j(\underline{y})\right)^{x_j} \left(1-\pi_j(\underline{y})\right)^{1-x_j} \cdots$$
$$\cdots \left(\pi_n(\underline{y})\right)^{x_{nj}} \left(1-\pi_n(\underline{y})\right)^{1-x_{nj}} d\underline{y},$$

gdzie całkowanie odbywa się po m współrzędnych.

Statystyka Λ ma rozkład χ^2 z (2^n – liczba parametrów – 1) stopniami swobody.

Mając określony podstawowy rozmiar przestrzeni ukrytej oraz estymatory parametrów funkcji $\pi_i(\underline{y})$, pożądane jest znalezienie oczekiwanej lokalizacji indywidualności w przestrzeni ukrytej.

W pracy [Bartholomew 1983] wychodzi się od założonego rozkładu indywidualnej pozycji w przestrzeni $\underline{Y}$ przy określonej jej wartości wektora $\underline{x}$. Jest to podstawa do skalowania nie tylko określonej postawy osoby z próby, ale także każdej innej z tej samej populacji. Pełne określenie założonego rozkładu nie jest oczywiście efektywną metodą skalowania. Poszukiwana jest pojedyncza wartość, wartość

skali lub indeksu, którą można wykorzystać w późniejszej analizie. Jedną z metod jest obliczenie kilku miar położenia dla założonego rozkładu. Jedną z miar zastosowaną do modelu logitowego (probitowego) jest miara oparta na tzw. *y*-wartościach [Bartholomew 1980]. *Y*-wartość jest to warunkowa wartość oczekiwana $\underline{Y}$ przy określonym wektorze $\underline{X}$

$$E(Y_j / \underline{x}) = \int_0^1 \ldots \int_o^1 y_j g(\underline{y} / \underline{x}) d\underline{y}, \quad j = 1,\ldots,m,$$

gdzie $g(\underline{y} / \underline{x}) = \prod_{i=1}^{n} \{\pi_i(\underline{y})\}^{x_i} \{1 - \pi_i(\underline{y})\}^{1-x_i} / f(\underline{x})$.

Na podstawie *y*-wartości można również porangować komórki tabeli kontyngencji, czyli respondentów, pod względem badanej cechy ukrytej przy ich wektorze odpowiedzi reprezentowanym przez jedną z komórek tabeli kontyngencji.

Każda monotoniczna funkcja *y*-skali nie powoduje zmiany dopasowania modelu, więc skala nie może być jednorodna. Zmienne ukryte mogą być określane tylko na skali porządkowej. Określenie to jest łatwe dla modelu logitowego i jest to wyższość nad modelem probitowym.

Dla pojedynczej zmiennej ukrytej jest to łatwe do pokazania. Zakłada się, że gęstość zmiennej *Y* pod warunkiem $\underline{x}$ jest dana następująco:

$$g(y / \underline{x}) \propto y^T (1 - y)^{A-T} \psi(y), \; 0 \le y \le 1, \; A = \sum_{i=1}^{n} \alpha_i, \; T = \sum_{i=1}^{n} \alpha_i x_i,$$

gdzie $\psi(y)$ jest funkcją *y* niezależną od $\underline{x}$ i *Y* jest zmienną losową o rozkładzie jednostajnym na [0; 1]. Wszystkie informacje o *Y* zawarte w X_i (i = 1,..., n) są przekazane przez statystykę $T = \sum_{i=1}^{n} \alpha_i X_i$. Wynika stąd, że wszystkie *y*-wartości wyznaczone przez $\underline{X}$ muszą być funkcją jedynie statystyki *T*. Można zastosować samą statystykę *T* (T_i oznacza wartość statystyki *T* dla *i*-tej indywidualności przy jej określonym wektorze odpowiedzi) w celu ustalenia rankingu indywidualności na skali ukrytej, ale wcześniej trzeba określić, w jakim sensie ranking ten ma być interpretowany. W ustaleniu tej interpretacji wykorzystuje się twierdzenie, które mówi, że jeżeli $T_1 < T_2$, to dla każdej wartości y_0 jest bardziej prawdopodobne, że indywidualność 1 jest poniżej tej wartości niż indywidualność 2. Większe „intuicyjne" znaczenie ma funkcja *T*, gdy odwołamy się do tego, że α mierzy dyskryminacyjną moc pytań. Te zmienne (pytania, wskaźniki), dla których czynnik α jest duży, mają większe znaczenie w określaniu statystyki *T*.

2. Dopasowanie jednoczynnikowego modelu logitowego

W modelu jednoczynnikowym mamy do czynienia z pojedynczą zmienną ukrytą Y, tak więc wszystkie wzory podane będą dla $m = 1$.

Dla modelu, w którym wykorzystano model logitowy [Bartholomew 1980], pomiarem zależności dla X_i i X_j jest tzw. *cross-product ratio*:

$$R_{ij} = \frac{E\pi_i(y)\pi_j(y)E(1-\pi_i(y))(1-\pi_j(y))}{E\pi_i(y)(1-\pi_j(y))E\pi_j(y)(1-\pi_i(y))}. \tag{6}$$

Jeżeli przyjmuje się, że funkcje F i G są funkcjami logitowymi, to

$$R_{ij} = 1+\sigma^2\alpha_i\alpha_j + o(\alpha^4) \quad i \neq j, \tag{7}$$

gdzie $\sigma^2 = 3{,}289868$.

Model polega na dopasowaniu wartości teoretycznych *cross-product ratio* do ich wartości empirycznych. Wartości empiryczne oblicza się z wzoru

$$R_{ij} = \frac{n_{00}n_{11}}{n_{01}n_{10}}, \tag{8}$$

gdzie: n_{11} – liczba osób, dla których $x_i = x_j = 1$,

n_{01} – liczba osób, dla których $x_i = 0, x_j = 1$,

pozostałe analogicznie.

Tak więc dla małych wartości α wyrażenia

$$c_{ij} = (R_{ij} - 1)/\sigma^2 \quad i \neq j, \tag{9}$$

mają taką samą formę jak kowariancje (korelacje) w standardowym modelu analizy czynnikowej. Oznacza to, że model może być dopasowany przez odpowiednią adaptację jednej z istniejących metod standardowych. W przypadku, gdy wartości α są duże, minimalizujemy odległość pomiędzy zbiorem estymowanych wartości R_{ij} a ich oczekiwanymi wartościami, stosując techniki iteracyjne.

Dla omawianego przypadku ($m = 1$) istnieje prosta metoda estymacji. Zakładając przybliżenie (7), przyrównuje się sumy $\{\hat{c}_{ij}\}$ po wierszach do ich przewidywanych wartości. Otrzymuje się w ten sposób następujące równania

$$\alpha_i\left(\sum_{j=1}^{n}\alpha_j + \alpha_i\right) = \hat{C}_i = \sum_{j=1}^{n}\hat{c}_{ij} \quad i \neq j. \tag{10}$$

Równania te dają dobre przybliżenie, ale można je jeszcze poprawić, stosując techniki iteracyjne. Parametry $\alpha_{i0}(\pi_i)$ są estymowane przez porównanie obserwowanych i oczekiwanych częstości brzegowych.

W pracy przedstawiono metodę estymacji wykorzystującą pierwszy i drugi porządek brzegowy. Model logitowy posiada własność, która często umożliwia uzyskanie prostego rozwiązania dla $m = 1$. Jak już wcześniej wspomniano, rozwiązanie to jest dobrym punktem wyjścia dla metody iteracyjnej, dzięki której rozwiązanie to można ulepszyć. Metoda ta bazuje na wartościach $\hat{c}_{ij} = \dfrac{R_{ij} - 1}{\alpha_i \alpha_j \sigma^2}$ (por. (7)).

Podstawą tej metody jest wyestymowanie parametrów α_i oraz π_i w taki sposób, by:

a) częstości *cross* produktu dla modelu były możliwie najbliższe obserwacjom,

b) brzegowe częstości modelu i danych były zgodne.

Parametry α_i są estymowane metodą iteracyjną, wychodząc z przybliżenia danego wzorem (7). Kroki iteracyjne są następujące:

1) wyznaczyć wektor $\underline{\alpha}$ tak, aby $\alpha_i \alpha_j$ były najbliższe wyestymowanym wartościom $\dfrac{R_{ij} - 1}{\sigma^2}$ dla $i, j = 1,2,...,n, \quad i \neq j$;

2) wyznaczyć wektor $\underline{\pi}$, przyrównując $E\pi_i(y) = \int_0^1 \dfrac{\pi_i y^{\alpha_i}}{\pi_i y^{\alpha_i} + (1-\pi_i)(1-y)^{\alpha_i}} dy$ do odpowiednich brzegowych częstości z wykorzystaniem wektora $\underline{\alpha}$ otrzymanego w 1);

3) poprawić estymację $\underline{\alpha}$, wykorzystując równanie (10) lub minimalizując formę $\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} (\hat{c}_{ij} - \alpha_i \alpha_j)^2$;

4) ulepszyć estymację $\underline{\pi}$, wykorzystując poprawiony wektor $\underline{\alpha}$;

5) powtarzać cykl, aż $\underline{\alpha}$ oraz $\underline{\pi}$ osiągną zbieżność.

W większości badanych przypadków, gdzie α_i są małe i ich wielkość jest tego samego rzędu, zbieżność estymacji parametrów jest szybka; jest to również prawdziwe dla $\underline{\pi}$.

Gdy wszystkie parametry są wyestymowane, należy sprawdzić dobroć dopasowania modelu.

Problem testowania dobrego dopasowania jest częściowo rozwiązany. Jeżeli 2^n jest małe w porównaniu z rozmiarem próby, można porównać zaobserwowane i

oczekiwane częstości w komórkach 2^n tablicy kontyngencji testem χ^2 lub statystyką log-wiarygodności. W przypadku, gdy n jest duże, to 2^n łatwo przekracza rozmiar próby. Tak więc z np. $n = 20, 30$ większość komórek będzie pusta i większość pozostałych będzie zawierać tylko jedną obserwację. Statystyka log-wiarygodności może być nadal stosowana, ale w tym przypadku występuje mała znajomość jej rozkładu.

Przykład 1. Problemem badawczym jest pomiar stopnia wiedzy o raku. Próbę 1729 osób sklasyfikowano według czterech pytań dotyczących wiedzy o raku. Prawidłowej odpowiedzi przyporządkowano wartość 1, a nieprawidłowej wartość zero. Zakłada się, że te cztery pytania odzwierciedlają poziom wiedzy o raku. Dane przedstawiono w tabeli kontyngencji 1.

Tabela 1. Tabela kontyngencji 1

Komórka	Liczebność	
	zaobserwowana (O_i)	oczekiwana (E_i)
0000	477	466,5
0001	12	16,1
0010	150	156,4
0011	11	8,6
0100	231	250,1
0101	13	18,9
0110	378	355,5
0111	45	44
1000	63	67,1
1001	7	3
1010	32	35,8
1011	4	2,4
1100	94	78,8
1101	12	7,5
1110	169	182,9
1111	31	34,4
Suma	1729	

Źródło: obliczenia własne na podstawie [Bartholomew 1983].

W celu estymacji parametrów funkcji $\pi_i(y)$ zastosowano iterację opisaną w poprzednim paragrafie.

Funkcja $\pi_i(y)$, określająca prawdopodobieństwo udzielenia odpowiedzi poprawnej ($x_i = 1$) przez respondenta, którego ukryta postawa dana jest przez y, określona jest wzorem (por. (4b)):

$$\pi_i(y) = \frac{\pi_i y^{\alpha_i}}{\pi_i y^{\alpha_i} + (1-\pi_i)(1-y)^{\alpha_i}}.$$

Zgodnie z opisanymi krokami estymacji parametrów funkcji $\pi_i(y)$ należy najpierw obliczyć wielkości R_{ij}, a następnie wyznaczyć wektor $\underline{\alpha}$ tak, by $\alpha_i\alpha_j$ były najbliższe wartościom $\frac{R_{ij}-1}{\sigma^2}$. Wartości te, obliczone z tabeli kontyngencji 1, przedstawione są w tab. 2.

Tabela 2. Tabela kontyngencji 2

R_{ij} $(\frac{R_{ij}-1}{\sigma^2})$	1	2	3
2	2,813 (0,551)		
3	1,683 (0,207)	5,05 (1,231)	
4	2,301 (0,395)	2,946 (0,591)	3,105 (0,64)

Źródło: obliczenia własne.

Metoda wykorzystana w tym przykładzie polega na minimalizacji funkcji

$$\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}(c_{ij} - \alpha_i\alpha_j)^2.$$

Otrzymuje się stąd pierwszą estymację parametrów α_i przedstawioną w tab. 3.

W kolejnym etapie wyznacza się wartości pierwszego estymatora parametrów π_i, przyrównując $E\pi_i(y)$ do częstości brzegowych (O_i) i wykorzystując α_i wyznaczone w punkcie 2). Oczekiwana częstość dana jest wzorem:

$$E\pi_i(y) = \int_0^1 \frac{\pi_i y^{\alpha_i}}{\pi_i y^{\alpha_i} + (1-\pi_i)(1-y)^{\alpha_i}} dy.$$

Rozwiązując układ równań:

$$E\pi_1(y) = 0{,}23828,\ E\pi_2(y) = 0{,}56275,\ E\pi_3(y) = 0{,}47426,\ E\pi_4(y) = 0{,}07808,$$

otrzymuje się pierwszą estymację parametrów π_i (tab. 3).

Postępując zgodnie z dalszymi etapami estymacji i powtarzając cykl aż do otrzymania zbieżności parametrów, otrzymano estymację, którą przedstawiono w tab. 3.

Tabela 3. Estymacje zbieżności parametrów

Parametr	Pierwsza estymacja	Ostatnia estymacja
α_1	0,444	0,445
α_2	1,228	1,55
α_3	0,864	0,86
α_4	0,506	0,456
π_1	0,213	0,212
π_2	0,604	0,62
π_3	0,461	0,464
π_4	0,057	0,061
c_{12}	1,011	
c_{13}	0,541	
c_{14}	1,761	
c_{23}	1,161	
c_{24}	0,952	
c_{34}	1,464	
Λ	23,71	19,15
Stopnie swobody	7	7

Źródło: obliczenia własne na podstawie [Bartholomew 1983].

W celu obliczenia wartości statystyki $\Lambda = 2\sum O_i \ln(\frac{O_i}{E_i})$, określającej dobroć dopasowania modelu, należy obliczyć częstości oczekiwane (tab. 1):

$$f(\underline{x}) = E\prod_{i=1}^{n} f_i(x_i / y) = \int_0^1 \prod f_i(x_i / y)dy .$$

W tym przykładzie statystyka Λ ma rozkład $\chi^2(2^4 - 8 - 1)$.

Dla ustalenia uszeregowania według stopnia wiedzy o raku respondentów, których odpowiedzi należą do określonej komórki w tabeli kontyngencji, oblicza się y-wartości. Rangują one komórki zgodnie ze stopniem wiedzy o raku. Są to wartości oczekiwane zmiennej ukrytej Y przy określonym wektorze odpowiedzi $\underline{x}$:

$$E(Y / \underline{x}) = \int_0^1 yg(y / \underline{x})dy ,$$

gdzie

$$g(y / \underline{x}) = \frac{\prod_{i=1}^{n} f_i(x_i / y)}{f(\underline{x})} .$$

Tabela 4 zawiera *y*-wartości.

Tabela 4. Wartości postaw respondentów (*y*)

Komórka	*y*-wartości	Wartości statystki *T*
0000	0,212	0
0001	0,304	0,465
0010	0,384	0,86
0011	0,475	1,316
0100	0,522	1,55
0101	0,615	2,006
0110	0,7	2,41
0111	0,797	2,886
1000	0,304	0,445
1001	0,394	0,901
1010	0,475	1,305
1011	0,567	1,761
1100	0,615	1,995
1101	0,711	2,451
1110	0,796	2,885
1111	0,889	3,311

Źródło: obliczenia własne.

Analiza *najwięcej-wartości* wskazuje, że najwięcej informacji o raku mają oczywiście te osoby, których odpowiedzi na wszystkie wskaźniki zakwalifikowane były do kategorii 1, a następnie ci, których odpowiedzi padały na komórki 1, 2 i 3 lub 2, 3 i 4. Natomiast spośród tych czterech najistotniejszy okazał się drugi wskaźnik.

Można wykorzystać metodę skalowania opartą na statystyce $T = \sum \alpha_i X_i$, w której przekazane są wszystkie informacje o Y zawarte w X_i. Ranking oparty na statystyce T oznacza, że jeżeli $T_1 < T_2$, to bardziej prawdopodobne jest to, że osoba, której odpowiedzi były zakodowane jako 0000, jest poniżej dowolnej wartości *y* niż ta, której odpowiedzi były zakodowane jako 0001. Wartości statystyki T przedstawiono w tab. 4.

Opisana metoda pozwala na ustalenie spodziewanej wartości cechy ukrytej przy określonych wskaźnikach (zmiennych obserwowalnych). Jednakże wartości te można traktować jedynie w skali porządkowej, czyli pozwala ona uporządkować respondentów pod względem badanej cechy.

Jeżeli wskaźniki nie są binarne, tylko wielotomiczne, stosuje się podobną metodę analizy modelu, jednakże wówczas nie bazuje się na wartościach R_{ij}, a na wartościach brzegowych i warunkowych wartościach oczekiwanych.

Literatura

Bartholomew D.J., *Factor Analysis for Categorical Data.* "Journal of the Royal Statistics Society" 1980 vol. 42.

Bartholomew D.J., *Latent Variable Models for Ordered Categorical Data*, „Journal of Econometrics" 1983 vol. 22.

Bollen K.A., *Structural Equations with Latent Variables*, Wiley Interscience Publication, Canada 1989.

Moosmuller G., *On the Relationship between Linear Factor Analysis (FA) and Latent Structure Analysis (LSA) from Lazarsfeld: a Classification of Models*, „Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali" 1989 vol. 11.

LATENT VARIABLE MODEL FOR BINARY DATA

Summary

Latent variable is a variable which is not directly observable. To measure this kind of variable we need a observable indicators. Latent and observable variable are bound by some function. In literature there are many methods to analyze latent variable model. This article discusses on one of this methods intended for continuous latent variable and binary data (observable variables). One-factor model logit it shown to fit as latent variable model. The method is illustrated by one empirical example.