Analiza danych jakościowych

Andrzej Dąbrowski

Spis treści

1 Dane **7**
1.1 Skale 8

2 Statystyczne modele danych jakościowych **11**
2.1 Rozkłady prawdopodobieństwa dla liczności w tablicach 13
2.2 Testowanie zgodności modelu z danymi 15
2.3 Testowanie jednorodności 19
2.4 Test niezależności χ^2 22
2.5 Iloraz krzyżowy 24

3 Modele logitowe **33**
3.1 Modele logitowe dla zmiennych liczbowych 34
3.2 Regresja logitowa ze zmiennymi nominalnymi 35
3.3 Regresja logitowa ze zmiennymi porządkowymi 37

4 Modele logarytmiczno-liniowe **41**
4.1 Modele hierarchiczne 50

A Skale dla prawdopodobieństw **63**

B Metoda IPF **67**

C Ćwiczenia **71**
C.1 Zadania na ćwiczenia w laboratorium 72
C.2 Zadania egzaminacyjne 75
C.2.1 Egzamin poprawkowy 77

Wstęp

Skrypt ten zawiera zapis wykładów z analizy danych jakościowych, wygłoszonych przeze mnie na Uniwersytecie Wrocławskim w semestrze zimowym roku akademickiego 2002003.

Wykład ten rozszerza w istotny sposób wykłady ze statystyki, które na ogół zawierają opis metod dla danych ilościowych. Praktyczne zastosowania statystyki w naukach biologicznych, medycznych czy w naukach społecznych wymagają wiedzy z tego szczególnego działu statystyki.

Andrzej Dąbrowski
luty 2003

Rozdział 1

Dane

Dane są efektem pomiarów i obserwacji, dokonywanych w doświadczeniach planowanych i takich, które polegają na zebraniu informacji o badanym zjawisku. Temu samemu obiektowi mogą być przypisane różne dane. Na przykład, danymi, kóre mogą być przypisane choremu są: diagnoza, stopień zaawansowania choroby, wiek, ciśnienie krwi, temperatura.

1.1 Skale

Dane wyrażają swoje wartości w różnych skalach.

Skala nominalna. Skalę nominalną stosuje się w celu klasyfikacji (nazwania) obiektów w populacji. Każdej klasie nadaje się odrębne oznaczenie (nazwę) w ten sposób, aby różne klasy miały różne oznaczenia. Często te oznaczenia będziemy nazywać *poziomami*. Na przykład w skali nominalnej wyrażona może być diagnoza (*grypa*, *katar*), stopień zaawansowania choroby (*lekko chory*, *ciężko chory*, *bardzo ciężko chory*), temperatura (poniżej 37°, między 38° a 40°), temperatura (37°,38°,40°). Struktura skali nominalnej nie zmieni się, jeśli dokonamy zmiany oznaczeń za pomocą przekształcenia różnowartościowego. Na przykład, diagnoza może być zapisana za pomocą numeru statystycznego choroby[1], stan chorego jako A,B,C itp.

Skala porządkowa. Jest to szczególny rodzaj skali nominalnej. Pozwala ona uporządkować klasy według stopnia intensywności opisywanej cechy. Na przykład, stopień zaawansowania choroby (*lekko chory*, *ciężko chory*, *bardzo ciężko chory*), temperatura (poniżej 37°, między 38° a 40°), temperatura (37°,38°,40°) wyrażają się w skali porządkowej, natomiast diagnoza (*grypa*, *katar*) nie jest wyrażona w skali porządkowej. Struktura skali porządkowej zachowa się, gdy dokonamy zmiany oznaczeń przez przekształcenie, zachowujące porządek. Tradycyjnie, jeśli skalę porządkową koduje się za pomocą liczb, to porządek naturalny tych liczb[2]odzwierciedla porządek skali. Podobnie, kodując za pomocą liter alfabetu $A,B,...$ porządek skali odzwierciedla się w porządku alfabetycznym. I tak system ocen: niedostateczny, dostateczny, dobry bardzo dobry wyrażający się w skali *porządkowej* koduje się[3] w Polsce za pomocą liczb 2,3,4,5. Analogiczny system ocen w USA koduje się za pomocą liter alfabetu $A,B,...$

Skala przedziałowa. Skala ta pozwala nie tylko klasyfikować i porządkować obiekty ale i porównywać je ilościowo. Wymaga ona ustalenia jednostki pomiaru

[1] ale wtedy pełni on *wyłącznie* funkcje opisową

[2] ale nie ich wartość!

[3] co nie oznacza, że oceny mają jakakolwiek wartość liczbową

i punktu zerowego skali. W tej skali naturalną operacją porównania jest różnica. Skala zachowuje się tak samo przy przekształceniach afinicznych $x' = ax+b$ ($a > 0$), których efektem jest zmiana jednostek. Na przykład temperatura (37°,38°,40°) jest wyrażona w skali przedziałowej a jednostki, w których jest wyrażona to skala Celsjusza. Przejście do skali Fahrenheita odbywa się przez przekształcenie $F = \frac{9}{5}C + 32$. Zero skali Fahrenheita jest w punkcie, odpowiadającym $-17.778°C$.

Skala ilorazowa. Różni się ona od skali przedziałowej tym, że występuje w niej *absolutny początek skali* (absolutne zero). W skali ilorazowej wyraża się wiele parametrów biologicznych (*wzrost, waga ciała, ciśnienie krwi*). Struktura skali nie zmieni się, jeśli zastosujemy przekształcenie $x' = ax$ ($a > 0$). Na przykład, wagę ciała możemy wyrazić w gramach, ale również w kilogramach, funtach itp. Naturalną operacją porównania dla skali ilorazowej jest iloraz dwóch wielkości.

Skale: nominalna i porządkowa opisują charakterystyki jakościowe danych i *dane*, wyrażone w takich skalach nazywają się *jakościowymi.* Dane, wyrażone w skalach: przedziałowej i ilorazowej nazywamy *danymi ilościowymi*.

Materiał, przedstawiony w dalszej części skryptu, dotyczyć będzie metod statystycznych związanych z analizą danych jakościowych.

Rozdział 2

Statystyczne modele danych jakościowych

Przypuśćmy, że dana jest zmienna nominalna lub porządkowa X o wartościach $x_1, x_2, ..., x_I$. Prawdopodobieństwo, że $X = x_i$ oznaczymy przez p_i.

Dane wynikające z obserwacji w n-elementowej próbce, powstającej z niezależnego losowawania wartości cechy X, będziemy zapisywać w *tablicy kontyngencji*

x_1	x_2	...	x_I
n_1	n_2	...	n_I

(2.1)

Parametr n_i określa, ile razy zaobserwowano w próbce wartość x_i.

Problemem, z jakim możemy się spotkać w przypadku takich danych, to sprecyzowanie rozkładu prawdopodobieństwa zmiennej X, czyli układu liczb $\{p_1, p_2, ...p_I$ spełniających warunki

$$\sum_{i=1}^{I} p_i = 1,\ p_i \geqslant 0\ i = 1, 2, ...I$$

Rozkładem, związanym z jednowymiarową tablicą (2.1) jest rozkład zmiennej losowej N_i określającej, ile wyników cechy X na poziomie x_i wystąpi w próbce. Rozkład ten zależy od rozkładu prawdopodobieństwa zmiennej X.

Jeżeli każdemu obiektowi przypisujemy dwie lub więcej zmiennych nominalnych albo porządkowych $X, Y, Z, ...$ to dane, uzyskane z obserwacji tych zmiennych zapisuje się w postaci tablicy kontyngencji. Tablica kontyngencji dla pary zmiennych (X, Y) o wartościach $X = \{x_1, x_2, ...x_I\}$ i $Y = \{y_1, y_2, ...y_J\}$ ma postać:

	y_1	y_2	...	y_J
x_1	n_{11}	n_{12}	...	n_{1J}
x_2	n_{21}	n_{22}	...	n_{2J}
...	...	...	...	...
x_I	n_{I1}	n_{I2}	...	n_{IJ}

,

gdzie n_{ij} jest liczbą obserwacji w n-elementowej próbce takich, że $X = x_i$ oraz $Y = y_j$. N_{ij} niech będzie zmienną, określajacą ile wystąpiło w próbce wyników zmiennej X na poziomie x_i i jednocześnie wyników zmiennej Y na poziomie y_j. Prawdopodobieństwo $P(X = x_i, Y = y_j)$ oznaczymy symbolem p_{ij}. Prawdopodobieństwa p_{ij} spełniają warunki

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1,\ p_{ij} \geqslant 0$$

Podobnie, tablica kontyngencji dla trójki zmiennych (X, Y, Z) o wartościach $X = \{x_1, x_2, ...x_I\}$, $Y = \{y_1, y_2, ...y_J\}$ i $Z = \{z_1, z_2, ...z_K\}$ ma postać:

		z_1	z_2	...	z_K
x_1	y_1	n_{111}	n_{112}	...	n_{11K}
	y_2	n_{121}	n_{122}	...	n_{12K}
	...	...	...	...	...
	y_J	n_{1J1}	n_{1J2}	...	n_{1JK}
...	...	...	...	...	...
x_I	y_1	n_{I11}	n_{I12}	...	n_{I1K}
	y_2	n_{I21}	n_{I22}	...	n_{I2K}
...	...	...	...	...	...
	y_J	n_{IJ1}	n_{IJ2}	...	n_{IJK}

Oznaczenia użyte w ostatniej tablicy są analogiczne do użytych w opisie tablicy dwuwymiarowej: n_{ijk} jest liczbą obserwacji w próbce takich, że $X = x_i$, $Y = y_j$ i $Z = z_k$, natomiast liczba p_{ijk} jest prawdopodobieństwem tego zdarzenia, a N_{ijk} zmienną o wartościach n_{ijk}.

Analogiczne sposoby zapisu danych i oznaczenia są używane dla układu więcej niż trzech zmiennych.

Oznaczenie 2.1 *Zastąpienie symbolem + w indeksie zmiennej oznacza operację sumowania po tym indeksie. Na przykład*

$$\begin{aligned} n_{+j} &= \sum_i n_{ij}, \; n_{++} = \sum_{i,j} n_{ij}, \\ n_{i+k} &= \sum_{,j} n_{ijk} \end{aligned}$$

2.1 Rozkłady prawdopodobieństwa dla liczności w tablicach

Różne sposoby uzyskania informacji w próbce mają wpływ na rozkład zmiennych losowych N_i, N_{ij}, N_{ijk}.

Rozkład dwumianowy (Bernoullego) $B(p)$

Powtarzamy n-krotnie eksperyment, polegający na wykonaniu n_0 niezależnych powtórzeń zmiennej o dwóch poziomach: *sukces, porażka* z prawdopodobieństwem

sukcesu p. Zmienna X mierzy liczbę sukcesów w n_0 powtórzeniach, natomiast n_i jest liczbą eksperymentów w której wystąpiło x_i sukcesów.

$$P\left(N_1 = n_1, N_2 = n_2, ..., N_I = n_I\right) = \prod_{i=1}^{I} \left(n_0 x_i p^{x_i} \left(1-p\right)^{n_0 - x_i}\right)^{n_i}$$

Rozkład Poissona $P\left(\lambda\right)$

Rozkład Poissona jest przypadkiem granicznym w rozkładzie dwumianowym[1]. Wystąpi on w tej sytuacji, gdy n-krotnie, niezależnie powtarzamy pewien eksperyment o wynikach *sukces, porażka* z małym prawdopodobieństwem sukcesu i oczekiwaną liczbą sukcesów λ w jednym eksperymencie. Przypuśćmy, że w tablicy (2.1) poziom x_i oznacza liczbę sukcesów w jednym eksperymencie, a n_i liczbę eksperymentów w której wystąpiło x_i sukcesów.

$$\begin{aligned} P\left(N_1 = n_1, N_2 = n_2, ..., N_I = n_I\right) &= \prod_{i=1}^{I} \exp\left(-\lambda n_i\right) \left(\frac{\lambda^{x_i}}{x_i!}\right)^{n_i} \\ &= \exp\left(-\lambda n\right) \prod_{i=1}^{I} \left(\frac{\lambda^{x_i}}{x_i!}\right)^{n_i} \end{aligned} \tag{2.2}$$

Rozkład wielomianowy $W\left(p_1, p_2, ..., p_I\right)$

Przypuśćmy, że zmienna X ma poziomy $x_1, x_2, ..., x_I$, prawdopodobieństwo, że X jest na poziomie x_i jest równe p_i. Elementy próbki utworzone są z n niezależnych obserwacji zmiennej X .

$$P\left(N_1 = n_1, N_2 = n_2, ..., N_I = n_I\right) = n_+! \prod_{i=1}^{I} \frac{p_i^{n_i}}{n_i!} \tag{2.3}$$

Stwierdzenie 2.2 *Rozkład wielomianowy ma następujące własności*

1. $N_i \sim B\left(p_i\right)$,

2. $\left(N_1, N_2, ..., N_r, N_0\right) \sim W\left(p_1, p_2, ..., p_r, p_0\right)$*, gdzie*

$$N_0 = \sum_{i=r+1}^{I} N_i,\ p_0 = \sum_{i=r+1}^{I} p_i$$

Rozkład produktowo-wielomianowy $V\left(p_{11}, p_{12}, ..., p_{IJ}\right)$

[1]jeżeli liczba powtórzeń n_0 jest duża a prawdopodobieństwo sukcesu jest małe; parametr λ jest oczekiwaną liczbą sukcesów

Niezależne zmienne X_i mają poziomy $x_{i1}, x_{i2}, ..., x_{iJ}$, prawdopodobieństwo, że X_i jest na poziomie x_{ij} jest równe p_{ij}. Powtarzamy n_{i+}-krotnie niezależnie eksperyment obserwacji zmiennej X_i i tą operację, niezależnie powtarzamy dla $i = 1, 2, ..., I$. Wielkość n_{ij} oznacza liczbę powtórzeń, kiedy osiągnięto poziom x_{ij}.

$$\begin{aligned} P\left(N_{11} = n_{11}, N_{12} = n_{12}, ..., N_{IJ} = n_{IJ}\right) &= \prod_{i=1}^{I} n_{i+}! \prod_{j=1}^{J} \frac{p_{ij}^{n_{ij}}}{n_{ij}!}, \qquad (2.4) \\ p_{i+} &= \sum_{j=1}^{J} p_{ij} = 1 \end{aligned}$$

Stwierdzenie 2.3 *Dla każdego $i = 1, 2, ..., I$ wektory losowe $(N_{i1}, N_{i2}, ..., N_{iJ})$*

1. *są niezależne,*
2. *mają rozkłady wielomianowe $W(p_{i1}, p_{i2}, ..., p_{iJ})$*

2.2 Testowanie zgodności modelu z danymi

Definicja 2.4 *Odchyleniem danych $\{n_{1,}n_2, ..., n_I\}$ od modelu M nazywamy liczbę*

$$G^2(M) = 2\sum_{i=1}^{I} n_i \ln \frac{n_i}{\widehat{n}_i},$$

gdzie $\widehat{n}_i = n\widehat{p}_i$ oraz $\widehat{p}_i$ jest estymatorem największej wiarygodności p_i w modelu M

Definicja 2.5 *Odległością χ^2 Pearsona[2] danych $\{n_{1,}n_2, ..., n_I\}$ od modelu M nazywamy liczbę*

$$\chi^2(M) = \sum_{i=1}^{I} \frac{\left(n_i - \widehat{n}_i\right)^2}{\widehat{n}_i},$$

gdzie $\widehat{n}_i = n\widehat{p}_i$ oraz $\widehat{p}_i$ jest estymatorem największej wiarygodności p_i w modelu M,

[2]Odległość ta została zaproponowana przez Karla Pearsona w artykule z 1900 pod tytułem *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it Can be Reasonably Supposed to Have Arisen from Random Sampling.* Motywacją tego artykułu było sprawdzenie m.in. jednorodności pojawiania się wyników ruletki w Monte Carlo.

Twierdzenie 2.6 *Odległość $\chi^2(M)$ Pearsona jest, pomnożonym przez n, oczekiwanym kwadratowym błędem względnym danych względem modelu M :* [3]

$$\begin{aligned} \chi^2(M) &= n\sum_{i=1}^{I}\widehat{p}_i\left(\frac{n_i-\widehat{n}_i}{\widehat{n}_i}\right)^2, \\ \widehat{p}_i &= \frac{\widehat{n}_i}{n} \end{aligned}$$

Twierdzenie 2.7 *Odległość $\chi^2(M)$ Pearsona jest asymptotycznie, przy $n\to\infty$ równa odchyleniu $G^2(M)$*

Twierdzenie 2.8 *Dla modelu M Poissona, dwumianowego lub wielomianowego (również produktowo-wielomianowego) odchylenie G^2 jest proporcjonalne do podwojonego logarytmu ilorazu wiarygodności hipotezy zgodności z modelem M przeciwko hipotezie niezgodności z tym modelem.*

Twierdzenie 2.9 *Zmienne losowe $G^2(M)$ i $\chi^2(M)$ mają asymptotycznie, przy $n\to\infty$ rozkład χ^2. Liczba stopni swobody tego rozkładu jest różnicą liczby stopni swobody hipotezy H_1orzekającej, że do danych* nie *można stosować modelu M i liczby stopni swobody hipotezy H_0orzekającej, że do danych można stosować model M.*

Twierdzenie 2.10 *Wartości*

$$d_i=\frac{n_i-\widehat{n}_i}{\sqrt{\widehat{n}_i}}, i=1,2,...,I$$

mają asymptotycznie, przy $n\to\infty$ rozkład standardowy normalny.

Uwaga 2.11 (praktyczna) *Na poziomie istotności $\alpha=0.05$ istotnie różne od 0 są te komórki tabeli dla których $|d_i|>1.96$ ($d_i^2>3.84$); na poziomie istotności $\alpha=0.01$ istotnie różne od 0 są te komórki tabeli dla których $|d_i|>2.58$ ($d_i^2>6.66$)*

Uwaga 2.12 (praktyczna) *Dobre przybliżenie dla zgodności z rozkładem χ^2 uzyskuje się dla odległości $G^2(M)$ gdy wszystkie wartości $\widehat{n}_i$ są nie mniejsze niż 1. Analogiczny warunek dla $\chi^2(M)$ jest wyrażony przez nierówność $\widehat{n}_i\geqslant 5$*

[3] Oczekiwany błąd względny danych względem modelu nazywany jest *inercją*

Lemat 2.13 *Problem maksymalizacji*

$$\begin{aligned} \sum_i c_i \ln q_i &= \max, \\ \sum_i q_i &= 1 \end{aligned}$$

ma rozwiązanie

$$\widehat{q_i} = \frac{c_i}{\sum_i c_i}$$

Przykład 2.14 (dane von Bortkiewicza) *Statystyk niemiecki Ladislaus von Bortkiewicz przytoczył w 1898 dane, dotyczące rocznej liczby wypadków śmiertelnych, spowodowanych kopnięciem przez konia wśród żołnierzy 10 korpusów armii pruskiej w ciągu 20 lat:*

Liczba wypadków w roku	*0*	*1*	*2*	*3*	*4*
Liczba korpusów i lat	*109*	*65*	*22*	*3*	*1*

Sprawdzimy, czy dane te mogą być opisane rozkładem Poissona.

Wyznaczymy najpierw estymator największej wiarygodności dla parametru λ. Logarytm funkcji wiarygodności (2.2) ma postać

$$\begin{aligned} \ln(L) &= \ln\left(\exp(-\lambda n)\prod_{i=1}^{I}\left(\frac{\lambda^{x_i}}{x_i!}\right)^{n_i}\right) = \\ &= -\lambda n + \sum n_i\left(x_i \ln\lambda - \ln(x_i!)\right) \\ 0 &= \frac{\partial \ln(L)}{\partial\lambda} = -n + \sum n_i \frac{x_i}{\lambda} \Longleftrightarrow \\ \widehat{\lambda} &= \frac{1}{n}\sum n_i x_i \end{aligned}$$

co w naszym przypadku daje wartość estymatora

$$\widehat{\lambda} = \frac{1}{200}\left(0*109 + 1*65 + 2*22 + 3*3 + 4*1\right) = 0.61$$

Przygotujemy tabelę do obliczeń statystyki testowej G^2 (lub χ^2)

x_i	*0*	*1*	*2*	*3*	*4*
n_i	109	65	22	3	1
$\widehat{p_i} = \exp\left(-\widehat{\lambda}\right)\frac{\lambda^{x_i}}{x_i!}$	.543 35	.331 44	.101 09	.02056	.00313
$\widehat{n_i} = n\widehat{p_i}$	108. 67	66. 29	20. 22	4. 11	.63

W ostatniej kolumnie oczekiwana liczebność wynosi $\widehat{n_i} = .63$*, co wskazuje na to, że szukanie poziomu krytycznego rozkładu* χ^2 *może być niedokładne (zbyt mała wartość - patrz Uwaga 2.12). W takich przypadkach zaleca się łączenie sąsiednich kategorii, tak aby wartość* $\widehat{n_i}$ *była dostatecznie duża. Po połączeniu dwóch ostatnich kategorii otrzymamy tablicę, dla której możemy obliczyć wartość* G^2[4]

x_i	*0*	*1*	*2*	*3 lub 4*
n_i	109	65	22	4
$\widehat{p_i} = \exp\left(-\widehat{\lambda}\right)\frac{\lambda^{x_i}}{x_i!}$	.543 35	.331 44	.101 09	.0 236 9
$\widehat{n_i} = n\widehat{p_i}$	108. 67	66. 29	20. 22	4. 74
$n_i \ln \frac{n_i}{\widehat{n_i}}$	.330 5	$-1.277\,4$	1. 856 1	$-.678\,97$

Wartość $G^2 = .460\,46$*. Hipoteza* H_1 *ma 3 stopnie swobody, gdyż nieznanymi parametrami są* p_0, p_1, p_2, p_3*, oznaczające prawdopodobieństwa wartości* x_i*, spełniające jedno równanie*

$$\sum_{i=0}^{3} p_i = 1$$

Hipoteza H_0 *ma 1 stopień swobody, gdyż* λ *jest jedynym nieznanym parametrem.* G^2 *ma więc rozkład* χ^2 *z 2 stopniami swobody. Poziom krytyczny dla modelu Poissona wynosi więc*

$$P\left(G^2 > .460\,46\right) = 0.79435$$

Wynika stąd, że z dużym przekonaniem możemy przyjąć model Poissona dla danych von Bortkiewicza.

Przykład 2.15 (listy federalistów) *W historii Stanów Zjednoczonych ważną rolę odegrało ustalenie autorstwa tzw "Listów federalistów". Zazwyczaj w takich przypadkach charakteryzuje się styl autora poprzez podanie rozkładu prawdopodobieństwa występowania charakterystycznych słów danego języka. Zbadano 262 bloki tekstu, zawierające po 200 słów każdy. Zbadamy, czy słowo "may"*[5] *może być opisane modelem Poissona. Zmienna X podaje liczbe wystąpień tego słowa w bloku.*

Liczba wystąpień słowa "may"	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Liczba fragmentów	*156*	*63*	*29*	*8*	*4*	*1*	*1*

Wartość estymatora parametru λ *wynosi*

$$\widehat{\lambda} = \frac{1}{262}\left(0 * 156 + 1 * 63 + 2 * 29 + 3 * 8 + 4 * 4 + 5 * 1 + 6 * 1\right) = .656\,49$$

[4] Ale nie χ^2!

[5] Mające dwa znaczenia: miesiąc *maj* lub czasownik *może* (od *móc*)

Tabela do obliczeń statystyki testowej G^2 (lub χ^2)

x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
n_i	156	63	29	8	4	1	1
$\widehat{p_i} = \exp\left(-\widehat{\lambda}\right)\frac{\lambda^{x_i}}{x_i!}$	.518 67	.340 5	.111 77	.02 446	.00401	.00053	.000
$\widehat{n_i} = n\widehat{p_i}$	135. 89	89. 21	29. 28	6. 41	1. 05	. 14	.0

Po połączeniu trzech ostatnich poziomów otrzymamy tablicę

x_i	*0*	*1*	*2*	*3*	*4,5,6*
n_i	156	63	29	8	6
$\widehat{n_i} = n\widehat{p_i}$	135. 89	89. 21	29. 28	6. 41	1. 21
$n_i \ln \frac{n_i}{\widehat{n_i}}$	21. 53	−21. 915	−. 278 66	1. 772 7	9. 606 8

Wartość $G^2 = 21.432$. Hipoteza H_1 ma 4 stopnie swobody, H_0 ma 1 stopień swobody. G^2 ma więc rozkład χ^2 z 3 stopniami swobody. Poziom krytyczny dla modelu Poissona wynosi więc

$$P\left(G^2 > 21.432\right) = 0.00009$$

Wynika stąd, że z dużym przekonaniem możemy odrzucić model Poissona dla tych danych. Otwartym zagadnieniem pozostaje, jakim rozkładem można opisać te dane.

2.3 Testowanie jednorodności

Gdy dane, zawarte w tabeli kontyngencji dla pary zmiennych (X, Y) można opisać rozkładem produktowo-wielomianowym, to naturalnym pytaniem o relację między X i Y jest hipoteza jednorodności. Rozkład produktowo-wielomianowy narzuca interpretację roli, jaką odgrywają zmienne X i Y:

- zmienna X jest grupująca, to znaczy na każdym poziomie x_i tej zmiennej obserwujemy niezależnie wartości zmiennej Y,
- zmienna Y jest wynikowa, co oznacza, że interesujemy się jej wartościami w zależności od różnych konfiguracji przyczyn (tu pogrupowania poprzez zmienną X)

Hipoteza jednorodności głosi, że rozkład zmiennej Y jest taki sam w każdej grupie, odpowiadającej innemu poziomowi zmiennej X.

Tłumacząc to na język rozkładu produktowo-wielomianowego:

$$H_0: \quad \forall_{j=1,2,...,j} \quad p_{1j} = p_{2j} = ... = p_{Ij} \stackrel{def}{=} q_j$$

Twierdzenie 2.16 *Test hipotezy*

$$H_0: \quad \forall_{j=1,2,...,J} \quad p_{1j} = p_{2j} = ... = p_{Ij} = q_j$$

jest oparty na statystyce testowej G^2

$$G^2 = 2\sum_{ij} n_{ij} \ln \frac{n_{ij}}{\widehat{n_{ij}}}$$

lub χ^2

$$\chi^2 = \sum_{ij} \frac{\left(n_{ij} - \widehat{n_{ij}}\right)^2}{\widehat{n_{ij}}}$$

gdzie

$$\widehat{n_{ij}} = \frac{n_{i+}n_{+j}}{n_{++}}$$

Statystyki te mają asymptotycznie rozkład χ^2 *z* $(I-1)(J-1)$ *stopniami swobody.*

Dowód. Estymatory największej wiarygodności dla nieznanych parametrów q_j uzyskamy minimalizując logarytm funkcji wiarygodności (2.4):

$$\begin{aligned} \ln\left(\prod_{i=1}^{I} n_{i+}! \prod_{j=1}^{J} \frac{p_{ij}^{n_{ij}}}{n_{ij}!}\right) &= \ln\left(\prod_{i=1}^{I} n_{i+}! \prod_{j=1}^{J} \frac{q_j^{n_{ij}}}{n_{ij}!}\right) = \\ &= c + \sum_{ij} n_{ij} \ln q_j = c + \sum_{j} n_{+j} \ln q_j \end{aligned}$$

przy warunku

$$\sum_j q_j = 1$$

Korzystając z lematu 2.13 otrzymamy rozwiązanie

$$\begin{aligned} \widehat{q_j} &= \frac{n_{+j}}{\sum_j n_{+j}} = \frac{n_{+j}}{n_{++}}, \\ \widehat{n_{ij}} &= n_{i+}\widehat{q_j} = \frac{n_{i+}n_{+j}}{n_{++}} \end{aligned}$$

Liczba stopni swobody dla hipotezy H_1 wynosi $IJ-I$, gdyż mamy IJ nieznanych parametrów, ale I dodatkowych warunków $p_{i+} = 1, i = 1, 2, ..., I$. Liczba stopni swobody dla hipotezy H_0 wynosi $J-1$, gdyż w tym przypadku nieznanymi parametrami są q_j, $j = 1, 2, ..., J$ z jednym warunkiem $\sum_j q_j = 1$. Liczba stopni swobody dla rozkładu χ^2, zgodnie z twierdzeniem 2.9, wynosi

$$DF(H_1) - DF(H_0) = IJ - I - (J-1) = (I-1)(J-1)$$

■

Przykład 2.17 (preferencje klientów) *(źródło [[4], str. 447]). Mieszkańcy południowej dzielnicy pewnego miasta zostali podzieleni na 4 grupy: mieszkających na północy dzielnicy (N), południu (S), wschodzie (E) i zachodzie (W). Z każdej z tych grup wylosowano niezależnie po 100 osób i każdej osobie zadano pytanie, czy w ciągu ostatniego tygodnia odwiedzili centrum handlowe, umieszczone w środku osiedla. Celem tej ankiety było rozstrzygnięcie, czy klienci w jednakowym stopniu korzystają z centrum dzielnicowego.*

Zmienna grupująca X o poziomach N, S, W, E wskazuje, skąd pochodzą ankietowani mieszkańcy dzielnicy. Zmienna Y ma dwa poziomy: T (tak, odwiedziłem centrum handlowe), N (nie odwiedziłem centrum handlowego). Wyniki ankiety umieszczone są w tablicy kontyngencji:

	T	N
N	28	72
S	56	44
W	43	57
E	34	66

Zgodnie z twierdzeniem 2.16 musimy wyznaczyć tablicę liczności oczekiwanych i wartości χ^2:

$\widehat{n_{ij}}$	T	N	$\widehat{n_{i+}}$
N	40.25	59.75	100
S	40.25	59.75	100
W	40.25	59.75	100
E	40.25	59.75	100
$\widehat{n_{+j}}$	161	239	400

χ^2_{ij}	T	N	χ^2_{i+}
N	3.728	2.512	6.240
S	**6.163**	**4.152**	10.305
W	.188	.125	.313
E	.970	.654	1.624
χ^2_{ij}	11.049	7.433	18.482

*Ponieważ liczebności oczekiwane są większe od 5, użyliśmy statystyki χ^2. Liczba stopni swobody wynosi 3*1=3. Poziom krytyczny wyliczamy z dystrybuanty rozkładu χ^2 z 3 stopniami swobody wynosi*

$$p = P\left(\chi^2 > 18.482\right) = .00035$$

co jest zdecydowanym argumentem za odrzuceniem hipotezy jednorodności. Spojrzenie na tablicę wartości χ^2pokazuje, gdzie realizuje się to odchylenie od jednorodności - w grupie S, gdzie wartości χ^2_{ij} są większe od 3.84, co oznacza istotnie duże (na poziomie 0.05) odchylenie od hipotezy jednorodności. Liczba odpowiedzi T (tak, korzystam z centrum handlowego) są zdecydowanie wyższe niż liczba odpowiedzi T, gdyby wszyscy odpowiadali tak samo. Podobnie, liczba odpowiedzi N (nie korzystam z centrum) jest zdecydowanie mniejsza. Można to interpretować tak, że mieszkańcy południowej części dzielnicy chętniej korzystają z centrum, usytuowanego w kierunku ich przejazdu do centrum miasta.

2.4 Test niezależności χ^2

Drugim ważnym problemem, który dotyczy dwuwymiarowych tablic kontyngencji jest testowanie niezależności. Naturalnym rozkładem, który występuje w tym zagadnieniu jest rozkład wielomianowy.

Test niezależności jest szczególnym przypadkiem twierdzenia 2.9.

Twierdzenie 2.18 *Test hipotezy niezależności*

$$H_0: \quad \forall_{i=1,2,\ldots,I} \forall_{j=1,2,\ldots,J} \quad p_{ij} = p_{i+} p_{+j}$$

jest oparty na statystyce testowej G^2

$$G^2 = 2 \sum_{ij} n_{ij} \ln \frac{n_{ij}}{\widehat{n_{ij}}}$$

lub χ^2

$$\chi^2 = \sum_{ij} \frac{\left(n_{ij} - \widehat{n_{ij}}\right)^2}{\widehat{n_{ij}}}$$

gdzie

$$\widehat{n_{ij}} = \frac{n_{i+} n_{+j}}{n_{++}}$$

Statystyki te mają asymptotycznie rozkład χ^2 z $(I-1)(J-1)$ stopniami swobody[6].

[6]Pearson w swojej oryginalnej pracy z 1900 błędnie podawał liczbe stopni swobody jako $IJ-1$. Dopiero Fisher wyjaśnił w 1922 poprawnie, na gruncie geometrii , pojęcie stopni swobody i podał reguły ich obliczania.

Dowód. Estymatory największej wiarygodności dla nieznanych parametrów p_{i+}, p_{+j} uzyskamy minimalizując logarytm funkcji wiarygodności (2.3):

$$\begin{aligned}\ln\left(n_{++}!\prod_{i,j}\frac{p_{ij}^{n_{ij}}}{n_{ij}!}\right) &= \ln\left(n_{++}!\prod_{i,j}\frac{p_{i+}^{n_{ij}}p_{+j}^{n_{ij}}}{n_{ij}!}\right)\\ &= c+\sum_{ij}n_{ij}\ln\left(p_{i+}p_{+j}\right)\\ &= c+\sum_{i}n_{i+}\ln p_{i+}+\sum_{j}n_{+j}\ln p_{+j}\end{aligned}$$

przy warunku

$$\sum_i p_{i+}=1, \sum_j p_{+j}=1$$

Korzystając z lematu 2.13 otrzymamy rozwiązanie

$$\begin{aligned}\widehat{p_{i+}} &= \frac{n_{i+}}{\sum_i n_{i+}}=\frac{n_{i+}}{n_{++}},\\ \widehat{p_{+j}} &= \frac{n_{+j}}{\sum_j n_{+j}}=\frac{n_{+j}}{n_{++}},\\ \widehat{n_{ij}} &= n_{++}\,\widehat{p_{i+}}\,\widehat{p_{+j}}=n_{++}\frac{n_{i+}n_{+j}}{\left(n_{++}\right)^2}=\frac{n_{i+}n_{+j}}{n_{++}}\end{aligned}$$

Liczba stopni swobody dla hipotezy H_1 wynosi $IJ-1$, gdyż mamy IJ nieznanych parametrów, ale 1 dodatkowy warunek $\sum_{ij}p_{ij}=1$. Liczba stopni swobody dla hipotezy H_0 wynosi $I-1+J-1=I+J-2$, gdyż w tym przypadku nieznanymi parametrami są p_{i+}, $i=1,2,...,I$ z jednym warunkiem $\sum_i p_{i+}=1$ oraz p_{+j}, $j=1,2,...,J$ z jednym warunkiem $\sum_j p_{+j}=1$. Liczba stopni swobody dla rozkładu χ^2, zgodnie z twierdzeniem 2.9, wynosi

$$DF\left(H_1\right)-DF\left(H_0\right)=IJ-1-(I+J-2)=(I-1)\left(J-1\right)$$

■

Przykład 2.19 (artretyzm, terapia, płeć) *(źródło [[3]]), Tabela przedstawia wyniki obserwacji 84 pacjentów, chorych na artretyzm. Cechy, obserwowane w eksperymencie to:*

W: wyniki leczenia (z - żadne, u - umiarkowane, l - lepsze);

P: płeć (k - kobieta, m - mężczyzna),

T: zastosowana terapia (a - aktywna, p - placebo).

n_{ijk}		W		
P	T	z	u	l
k	a	6	5	16
	p	19	7	6
m	a	7	2	5
	p	10	0	1

Zbadamy, czy zastosowana terapia miała wpływ na wyniki leczenia. Łącząc dane dla kobiet i mężczyzn, otrzymamy tabelę

n_{ij}	W		
T	z	u	l
a	13	7	21
p	29	7	7

Zbudujemy tabelę liczebności oczekiwanych i odległości χ^2

$\widehat{n_{ij}}$	W			
T	z	u	l	n_{i+}
a	20.5	6.83	13.67	41
p	21.5	7.17	14.33	43
n_{+j}	42	14	28	84

χ^2_{ij}	W			
T	z	u	l	χ^2_{i+}
a	2.744	.0042	**3.930**	6.678
p	2.616	.0040	3.749	6.369
χ^2_{+j}	5.360	.0082	7.679	13.047

*Liczba stopni swobody wynosi 1*2=2 a poziom krytyczny*

$$p = P\left(\chi^2 > 13.047\right) = .0015$$

co pozwala na odrzucenie hipotezy o niezależności wyników od zastosowanej terapii. Pogrubione pole w tablicy χ^2_{ij} *pokazuje na istotną różnicę w liczbie lepszych wyników przy zastosowanej aktywnej terapii w stosunku do hipotetycznej liczby, odpowiadającej niezależności.*

2.5 Iloraz krzyżowy

Inna koncepcja opisania związku między cechami opiera się na pojęciu stosunku szans.

Definicja 2.20 (stosunek szans) *Prawdopodobieństwo zajścia zdarzenia A jest równe p.* Stosunkiem szans *dla tego zdarzenia nazywamy iloraz*

$$\varpi = \varpi(A) = \frac{p}{1-p}$$

Dobrym estymatorem stosunku szans jest wielkość

$$\widehat{\varpi} = \widehat{\varpi}(A) = \frac{n(A)}{n - n(A)} = \frac{n(A)}{n(A')},$$

gdzie $n(A)$ jest liczbą obserwacji w próbie, dla których zaszło zdarzenie A, n jest wielkością próby. Gdy próba nie jest wielka zaleca się stosowanie nieco innego estymatora

$$\widetilde{\varpi} = \widetilde{\varpi}(A) = \frac{n(A) + 0.5}{n - n(A) + 0.5} = \frac{n(A) + 0.5}{n(A') + 0.5}$$

Przykład 2.21 *Dane o wykształceniu i dochodzie rocznym zebrano wśród 300 osób:*

	dochód niski	*dochód wysoki*
wykształcenie średnie	*70*	*30*
wykształcenie wyższe	*80*	*120*

Niech A będzie zdarzeniem, że osoba ma wykształcenie średnie, B - że ma niski dochód. Gdy ograniczymy się do osób z niskim dochodem to stosunek szans dla zdarzenia A można oszacować, jako

$$\widehat{\varpi}(A\,|B) = \frac{70}{80} = .875$$

co oznacza, że wśród osób z niskim dochodem jest prawie taka sama liczba osób o wykształceniu średnim i wyższym z lekką przewagą liczby osób z wykształceniem wyższym.

Gdy ograniczymy się do osób z wyższym dochodem to stosunek szans dla zdarzenia A można oszacować, jako

$$\widehat{\varpi}(A\,|B') = \frac{30}{120} = .25$$

co oznacza, że wśród osób z wysokim dochodem jest mała liczba osób o wykształceniu średnim a duża z wyższym (4 razy większa).

Z kolei, gdy ograniczymy się do osób z wykształceniem średnim to stosunek szans dla zdarzenia B można oszacować, jako

$$\widehat{\varpi}(B\,|A) = \frac{70}{30} = 2.33$$

a wśród osób z wykształceniem wyższym

$$\widehat{\varpi}(B\,|A') = \frac{80}{120} = .67$$

Zauważmy, że

$$\frac{\widehat{\varpi}\left(A\left|B\right.\right)}{\widehat{\varpi}\left(A\left|B'\right.\right)}=\frac{\widehat{\varpi}\left(B\left|A\right.\right)}{\widehat{\varpi}\left(B\left|A'\right.\right)}=\frac{70*120}{30*80}=3.5$$

Pierwszy stosunek mówi, że iloraz szans dla średniego wykształcenia jest 3.5 raza większy w grupie zarabiających mało od takiego ilorazu w grupie zarabiających dużo. Drugi stosunek mówi, że iloraz szans dla niskiego dochodu jest 3.5 raza większy w grupie osób o średnim wykształceniu od takiego ilorazu dla osób z wyższym wykształceniem. Podsumowując, jest silny związek między niskim wykształceniem a niskim dochodem. Liczba 3.5 jest miarą siły tego związku.

Z poprzedniego przykładu wynika potrzeba zdefiniowania nowego pojęcia.

Definicja 2.22 (iloraz krzyżowy) *Dana jest para cech binarnych* (X,Y). Ilorazem krzyżowym *dla tych cech nazywamy liczbę*

$$\begin{aligned}\theta &= \theta\left(X,Y\right)=\frac{p_{11}p_{22}}{p_{12}p_{21}},\\ gdziep_{ij} &= P\left(X=x_i,Y=y_j\right),i,j=1,2\end{aligned}$$

Estymator ilorazu krzyżowego z tablicy kontyngencji

	y_1	y_2
x_1	n_{11}	n_{12}
x_2	n_{21}	n_{22}

będzie postaci

$$\widehat{\theta}=\widehat{\theta}\left(X,Y\right)=\frac{n_{11}n_{22}}{n_{12}n_{21}}$$

lub, gdy dysponujemy małą liczba obserwacji

$$\widetilde{\theta}=\widetilde{\theta}\left(X,Y\right)=\frac{\left(n_{11}+0.5\right)\left(n_{22}+0.5\right)}{\left(n_{12}+0.5\right)\left(n_{21}+0.5\right)}$$

Twierdzenie 2.23 *Niech dana będzie para cech binarnych* (X,Y). *Oznaczmy:*

$$\begin{aligned}p_{ij} &= P\left(X=x_i,Y=y_j\right),i,j=1,2\\ A &= \left\{X=x_1\right\},B=\left\{Y=y_1\right\}\end{aligned}$$

Zachodzą wtedy równości:

1. $\theta=\frac{\varpi(A|B)}{\varpi(A|B')}=\frac{\varpi(B|A)}{\varpi(B|A')}=\frac{\varpi(A'|B')}{\varpi(A'|B)}=\frac{\varpi(B'|A')}{\varpi(B'|A)}$

2. Niech $p^*_{1j} = c_1\ p_{1j},\ p^*_{2j} = c_2\ p_{2j},\ c_1\ p_{1+} + c_2\ p_{2+} = 1$. *Wtedy* p^*_{ij} *jest rozkładem prawdopodobieństwa dla pary* (X, Y) *takim, że odpowiadający mu iloraz krzyżowy*

$$\theta^* = \frac{p^*_{11}p^*_{22}}{p^*_{12}p^*_{21}}$$

jest równy iloczynowi krzyżowemu θ.

3. Dla każdego θ *istnieje układ prawdopodobieństw* $p_{ij}(\theta)$ *taki, że*

$$\begin{aligned} p_{1+}(\theta) &= \frac{1}{2},\ p_{2+}(\theta) = \frac{1}{2}, \\ p_{+1}(\theta) &= \frac{1}{2},\ p_{+2}(\theta) = \frac{1}{2} \end{aligned}$$

oraz

$$\frac{p_{11}(\theta)\, p_{22}(\theta)}{p_{12}(\theta)\, p_{21}(\theta)} = \theta$$

Układ taki nazywamy standardową reprezentacją ilorazu krzyżowego θ

Reprezentacja standardowa jest wyznaczona jednoznacznie ze wzoru

$$\begin{aligned} p_{12}(\theta) &= p_{21}(\theta) = \frac{1}{2\left(1+\sqrt{\theta}\right)}, \\ p_{11}(\theta) &= p_{22}(\theta) = \frac{1}{2} - p_{12}(\theta) \end{aligned}$$

Reprezentacja standardowa przedstawia sytuację, gdyby doświadczenie wykonano tak, że zarówno cecha X jak i Y mają swoje wartości reprezentowane z taką samą częstością (nie preferujemy żadnych wartości tych cech). Wtedy prawdopodobieństwa występujące w tablicy standardowej odzwierciedlają siłę związku między tymi cechami.

Reprezentacja standardowa dla estymatora ilorazu krzyżowego $\widehat{\theta}$ wynika z powyższych wzorów:

$$\begin{aligned} p_{12}\left(\widehat{\theta}\right) &= p_{21}\left(\widehat{\theta}\right) = \frac{1}{2\left(1+\sqrt{\widehat{\theta}}\right)}, \\ p_{11}\left(\widehat{\theta}\right) &= p_{22}\left(\widehat{\theta}\right) = \frac{1}{2} - p_{12}\left(\widehat{\theta}\right) \end{aligned}$$

Przykład 2.24 *Cecha* X *wskazuje, czy osoba jest czy nie jest chora na rzadko występującą chorobę a* Y *czy występuje, czy nie występuje u badanej osoby spadek*

wagi ciała. Ze względu na małe prawdopodobieństwa spadku czy braku spadku wagi wśród osób u których występuje ta choroba, moglibyśmy nie zauważyć rzeczywistych rozmiarów wzajemnych relacji między wartościami tych cech. Wady tej jest pozbawiona reprezentacja standardowa.

Przypuśćmy, że udało nam się zebrać dane tylko od 18 osób chorych na tą chorobę

	spadek wagi	*brak spadku wagi*
chory	10	8
nie chory	300	600

$$\widehat{\theta} = \frac{10 * 600}{8 * 300} = 2.5$$

Reprezentacja standardowa tej tabeli ma postać

	spadek wagi	*brak spadku wagi*
chory	.306	.194
nie chory	.194	.306

co ujawnia, że gdyby chorych było tyle samo, co zdrowych to iloraz szans dla spadku wagi byłby równy 1.58 (= .306194) a nie 1.25 jak to było w naszej z trudem zebranej próbie.

Wartość ilorazu krzyżowego θ $(\widehat{\theta})$ można przedstawić za pomocą wykresu kołowego, czy kwadratowego, pozwalającego zobrazować siłę związku między cechami, reprezentowaną przez iloraz krzyżowy. Na osi pionowej, odpowiadajacej osobom chorym i osi poziomej, odpowiadającej spadkowi wagi rysujemy kwadrat[7] o boku $p_{11}\left(\widehat{\theta}\right)$, na osi pionowej, odpowiadajacej osobom chorym i osi poziomej, odpowiadającej brakowi spadku wagi rysujemy kwadrat o boku $p_{12}\left(\widehat{\theta}\right)$ itd. Stosunek sumy pól kwadratów lewy- górny, prawy-dolny do sumy pól prawy-górny, lewy_dolny wynosi

$$\frac{\left(p_{11}\left(\widehat{\theta}\right)\right)^2 + \left(p_{22}\left(\widehat{\theta}\right)\right)^2}{\left(p_{12}\left(\widehat{\theta}\right)\right)^2 + \left(p_{21}\left(\widehat{\theta}\right)\right)^2} = \frac{2\left(p_{11}\left(\widehat{\theta}\right)\right)^2}{2\left(p_{12}\left(\widehat{\theta}\right)\right)^2} = \frac{p_{11}\left(\widehat{\theta}\right) p_{22}\left(\widehat{\theta}\right)}{p_{12}\left(\widehat{\theta}\right) p_{21}\left(\widehat{\theta}\right)} = \widehat{\theta}$$

[7]Możo to być ćwiartka koła o tym promieniu

Zgodnie z teorią percepcji oglądając obiekty na płaszczyźnie porównujemy ich wielkości poprzez porównanie pól. Tak więc nasz wykres, poprzez porównanie pól kwadratów, dobrze ilustruje wielkość ilorazu krzyżowego.

dtbpF220.5625pt208.375pt0ptFigure

Kiedy obliczamy estymator $\widehat{\theta}$ ilorazu krzyżowego θ interesować nas musi rozkład prawdopodobieństwa tego estymatora. Pozwoli nam to na zbudowanie przedziału ufności, co umożliwi testowanie hipotezy o prawdziwej wartości ilorazu krzyżowego.

Twierdzenie 2.25 *W tablicy kontyngencji dla binarnych cech (X, Y) o rozkładach dwumianowym, Poissona lub wielomianowym, zmienna losowa $\ln\left(\widehat{\theta}\right)$ ma, asymptotycznie przy $n \to \infty$ rozkład $\mathcal{N}\left(\ln\left(\theta\right), \widehat{\sigma}\right)$, gdzie*

$$\widehat{\sigma} = \sqrt{\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)}$$

Wniosek 2.26 *Przedział ufności na poziomie $1 - \alpha$ dla $\ln\left(\theta\right)$ ma postać:*

$$\left(\ln\left(\widehat{\theta}\right) - z\left(1 - \frac{\alpha}{2}\right)\widehat{\sigma}, \ln\left(\widehat{\theta}\right) + z\left(1 - \frac{\alpha}{2}\right)\widehat{\sigma}\right),$$

gdzie $z\left(1 - \frac{\alpha}{2}\right)$ jest kwantylem rzędu $1 - \frac{\alpha}{2}$ dla standardowego rozkładu normalnego[8]*.*

Stwierdzenie to jest równoważne temu, że przedział ufności dla θ jest postaci

$$\left(\widehat{\theta}\exp\left(-z\left(1 - \frac{\alpha}{2}\right)\widehat{\sigma}\right), \widehat{\theta}\exp\left(z\left(1 - \frac{\alpha}{2}\right)\widehat{\sigma}\right)\right)$$

Przykład 2.27 *(kontynuacja przykładu 2.24).*

Wartość $\widehat{\sigma}$ obliczamy ze wzoru

$$\begin{aligned}\widehat{\sigma} &= \sqrt{\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)} = \\ &= \sqrt{\left(\frac{1}{10} + \frac{1}{8} + \frac{1}{300} + \frac{1}{600}\right)} = .\,479\,58\end{aligned}$$

[8]Dla $\alpha = 0.05$ kwantyl ten wynosi 1.96 a dla $\alpha = 0.01$ kwantyl ten wynosi 2.58

Przedział ufności dla θ *na poziomie* 0.95 *będzie miał postać:*

$$\begin{aligned} & \left(\widehat{\theta}\exp\left(-z\left(1-\frac{\alpha}{2}\right)\widehat{\sigma}\right), \widehat{\theta}\exp\left(z\left(1-\frac{\alpha}{2}\right)\widehat{\sigma}\right)\right) \\ = & \left(2.5\exp\left(-1.96 * .479\,58\right), 2.5\exp\left(1.96 * .479\,58\right)\right) \\ = & \left(.976\,59, 6.399\,8\right) \end{aligned}$$

Wskazuje to na olbrzymi zakres możliwych wartości ilorazu krzyżowego. Odpowiedzialne za to są nadzwyczaj małe ilości obserwacji związanych z osobami chorymi.

Niezależność i jednorodność cech można łatwo wyrazić poprzez iloraz krzyżowy.

Twierdzenie 2.28 *Cechy* X *o poziomach* $\{x_1, x_2, ..., x_I\}$ *i* Y *o poziomach* $\{y_1, y_2, ...$ *mających łączny rozkład prawdopodobieństwa*

$$p_{ij} = P\left(X = x_i, Y = y_j\right),\ i = 1, 2, ..., I,\ j = 1, 2, ..., J$$

są niezależne wtedy i tylko wtedy, gdy każdy iloraz krzyżowy

$$\theta\left(i, j; i\prime, j'\right) = \frac{p_{ij}p_{i'j'}}{p_{i'j}p_{ij'}}, i, i' = 1, 2, ..., I,\ j, j' = 1, 2, ..., J$$

jest równy 1.

Sprawdzenie niezależności za pomocą ilorazów krzyżowych wymaga więc sprawdzenia $(IJ)^2$ warunków. Uciążliwość tej procedury można znacząco zmniejszyć.

Twierdzenie 2.29 *Cechy* X *i* Y *są niezależne wtedy i tylko wtedy, gdy każdy iloraz krzyżowy*

$$\theta\left(1, 1; i, j\right) = \frac{p_{11}p_{ij}}{p_{1j}p_{i1}}, i = 2, 3, ..., I, j = 2, 3, ..., J$$

jest równy 1.

W szczególności, gdy X *i* Y *są cechami binarnymi to ich niezależność jest równoważna temu, że ich iloraz krzyżowy jest równy 1.*

Analogiczne wyniki dotyczą jednorodności rozkładów

Twierdzenie 2.30 *Cecha X o poziomach $\{x_1, x_2, ..., x_I\}$ jest grupująca. Rozkład cechy Y o poziomach $\{y_1, y_2, ..., y_J\}$, ma rozkład prawdopodobieństwa*

$$p_{ij} = P\left(Y = y_j | \, X = x_i, \right), \; i = 1, 2, ..., I, \; j = 1, 2, ..., J$$

Rozkład cechy Y jest jednorodny względem X to znaczy taki, że

$$\forall_{j=1,2,...,J} \; p_{1j} = p_{2j} = ... = p_{Ij}$$

wtedy i tylko wtedy, gdy każdy iloraz krzyżowy

$$\theta\left(i, j; i\prime, j'\right) = \frac{p_{ij} p_{i'j'}}{p_{i'j} p_{ij'}}, i, i' = 1, 2, ..., I, \; j, j' = 1, 2, ..., J$$

jest równy 1.

Twierdzenie 2.31 *Rozkład cechy Y jest jednorodny względem X wtedy i tylko wtedy, gdy każdy iloraz krzyżowy*

$$\theta\left(1, 1; i, j\right) = \frac{p_{11} p_{ij}}{p_{1j} p_{i1}}, i = 2, 3, ..., I, j = 2, 3, ..., J$$

jest równy 1.

Iloraz krzyżowy estymujemy na podstawie tablicy kontyngencji. W takim razie ważny jest problem, czy estymator ilorazu krzyżowego wskazuje na danym poziomie istotności, że prawdziwa wartość tego ilorazu jest równa 1. Odpowiedź na to pytanie wynika natychmiast z twierdzenia 2.25.

Twierdzenie 2.32 *Statystyka testowa do testowania hipotez*

$$\begin{aligned} H_0 &: \quad \theta = 1, \\ H_1 &: \quad \theta \neq 1 \; (\theta < 1) \; (\theta > 1) \end{aligned}$$

oparta jest na statystyce testowej

$$z = \frac{\ln \widehat{\theta}}{\widehat{\sigma}}$$

mającej asymptotycznie standardowy rozkład normalny.

Hipotezę H_0 odrzucamy na rzecz hipotezy H_1 gdy zachodzą odpowiednie nierówności

$$\begin{aligned} |z| &> z\left(1-\frac{\alpha}{2}\right), \\ z &< -z\left(1-\alpha\right), \\ z &> z\left(1-\alpha\right) \end{aligned}$$

gdzie $z\left(u\right)$ jest kwantylem rzędu u standardowego rozkładu normalnego.

Przykład 2.33 *(kontynuacja przykładu 2.24)*

Zbadamy, czy zachorowanie na analizowaną chorobę i spadek wagi są od siebie niezależne. Obliczyliśmy, że estymator ilorazu krzyżowego ma w tym przypadku wartość $\widehat{\theta} = 2.5$, $\widehat{\sigma} = .479\,58$. Wartość statystyki z jest równa

$$z = \frac{\ln \widehat{\theta}}{\widehat{\sigma}} = \frac{\ln 2.5}{.479\,58} = 1.910\,6$$

Poziom krytyczny dla hipotez

$$\begin{aligned} H_0 &: \quad \theta = 1, \\ H_1 &: \quad \theta \neq 1 \end{aligned}$$

jest równy

$$p = P\left(|Z| > 1.910\,6\right) = .0561$$

co prowadzi do konkluzji, że dysponujemy słabymi argumentami za odrzuceniem hipotezy zerowej a więc słabymi argumentami za uznaniem zależności między zachorowaniem na analizowaną chorobę i spadkiem wagi, mimo *wydawałoby się* dużej wartości $\widehat{\theta}$.

Rozdział 3

Modele logitowe

W dwóch kolejnych rozdziałach będziemy rozważać modele prawdopodobieństw lub liczebności zdarzeń jako funkcji innych zmiennych. Stworzenie takich modeli jest o tyle kłopotliwe, że zastosowanie klasycznej teorii regresji z błędami modelu, mającymi rozkład normalny nie jest w tym przypadku możliwe. Prawdopodobieństwa bowiem ograniczone są do przedziału $(0, 1)$ a wartości bliskie krańcom skali mają szczególne znaczenie. Znacznie trudniej jest uzyskać wzrost prawdopodobieństwa o 0.01 gdy obserwujemy zdarzenie o prawdopodobieństwie 0.95 niż wtedy, gdy obserwujemy zdarzenie o prawdopodobieństwie 0.6. Rozwiązanie tego zagadnienia może ułatwić przedstawienie prawdopodobieństwa w innej skali(patrz Dodatek A)

3.1 Modele logitowe dla zmiennych liczbowych

Modele logitowe są modelami regresyjnymi, opisującymi relację między zmienną wynikową *dychotomiczną*[1] a zmiennymi objaśniającymi. W modelu tym interesuje nas regresja, najlepiej liniowa, między prawdopodobieństwem sukcesu, wyrażonym w skali logitowej a zmiennymi objaśniającymi[2].

Przykład 3.1 (Ciśnienie) *(źródło, [1] str. 93)*

Mieszkańcy Framingham (Massachusetts), mężczyźni w wieku 40-60 lat, byli obserwowani przez 6 kolejnych lat. Notowano, czy w tym czasie zachorowali na wieńcową chorobę serca. Zbadamy, jaki wpływ na prawdopodobieństwo zachorowania może mieć poziom ciśnienia krwi

ciśnienie	*chorzy*	*zdrowi*	*probit*
112	*3*	*153*	$\ln \frac{3}{153} = -3.93$
122	*17*	*235*	$\ln \frac{17}{235} = -2.63$
132	*12*	*272*	$\ln \frac{12}{272} = -3.12$
142	*16*	*255*	$\ln \frac{16}{255} = -2.77$
152	*12*	*127*	$\ln \frac{12}{127} = -2.36$
162	*8*	*77*	$\ln \frac{8}{77} = -2.26$
177	*16*	*83*	$\ln \frac{16}{83} = -1.65$
192	*8*	*35*	$\ln \frac{8}{35} = -1.48$

Regresja liniowa okazała się dobrym modelem relacji ciśnienie - logit:dtbpF4.619i

[1] tzn, mającą dwie wartości; jedna z nich tradycyjnie nazywa się *sukcesem*

[2] Dla niektórych danych zamiast skali logitowej trzeba użyć innej skali prawdopodobieństw, na przykład probitowej czy też podwójnie logarytmicznej.

Współczynnik determinacji modelu wynosi 0.8572 *co wskazuje na dobre jego dopasowanie do danych. Jak widać z wykresu, jedynie dwa punkty, odpowiadające dwom najniższym wartościom ciśnienia odbiegają istotnie od prostej logitowej. Model, który uzyskaliśmy ma postać*

$$lgt = -6.503 + 0.0237\,c$$

gdzie c oznacza ciśnienie krwi. Wzrost tego ciśnienia o 1 jednostkę powoduje wzrost logitu o 0.0237 *co oznacza, że iloraz krzyżowy dla zachorowania i dla danego ciśnienia przy jego wzroście o 1 jednostkę wynosi* $\exp(0.0237) = 1.024$. *Zwiększenie ciśnienia o 1 jednostkę powoduje zwiększenie ilorazu szans zachorowania o 2%.*

Mając model logitowy odwracając skalę możemy narysować relację między ciśnieniem a prawdopodobieństwem zachorowaniadtbpF4.619in2.8833in0ptreglogpr.wm

Moglibyśmy w tej sytuacji zastosować regresję probitową. Jest ona nawet nieco lepiej dopasowana do danych (współczynnik determinacji jest równy 0.8781*). Praktyczna jednak łatwość wykorzystania regresji logitowej rekompensuje nieco lepszy model probitowy. Dla ilustracji pokażemy relację między ciśnieniem a prawdopodobieństwem, uzyskanym z modelu probitowego.dtbpF4.619in2.8833in0ptregprob*

Twierdzenie 3.2 *W regresji logitowej liczba stopni swobody w teście zgodności* G^2 *lub* χ^2 *jest równa liczbie występujących w danych logitów minus liczba parametrów w modelu regresyjnym.*

Dowód. Zgodnie z techniką wyznaczania liczby stopni swobody w testach zgodności, jest ona równa liczbie wolnych parametrów w hipotezie konkurencyjnej minus liczba wolnych parametrów w hipotezie zerowej. W naszym przypadku w hipotezie konkurencyjnej jest tyle parametrów, ile jest logitów do oszacowania. W hipotezie zerowej, opisującej dane za pomocą równania regresji jest tyle parametrów, ile występuje w tym równaniu. ■

3.2 Regresja logitowa ze zmiennymi nominalnymi

Regresja logitowa może znaleźć zastosowanie również wtedy, gdy niektóre zmienne objaśniające są nominalne. Każdej zmiennej nominalnej przyporządkujemy tyle zmiennych indykatorowych, ile różnych wartości ma dana zmienna. Po wprowadzeniu takich zmiennych budujemy zwykły model regresji logitowej

Definicja 3.3 *Niech zmienna nominalna X ma wartości $\{x_1, x_2, ..., x_I\}$. Zmiennymi indykatorowymi, odpowiadającymi X, nazywamy zmienne liczbowe $X^{(1)}, X^{(2)}, ..$ $X^{(I-1)}$ o wartościach $\{0, 1\}$, takie, że $X^{(i)} = 1 \Longleftrightarrow X = x_i$*

Przykład 3.4 *(kontynuacja przykładu 2.19)*

Interesuje nas jak prawdopodobieństwo uzyskania lepszego wyniku zależy od płci i zastosowanej terapii. Przekształćmy tabelę tak, aby przygotować dane do obliczeń

n_{ijk}		$prawdop \quad \lg t$	$P^{(k)}$	$T^{(a)}$
P	T	p_{ij}		
k	a	$\frac{21}{27} = .778 \quad \ln \frac{21}{6} = 1.253$	1	1
	p	$\frac{13}{32} = .406 \quad \ln \frac{13}{19} = -.379$	1	0
m	a	$\frac{7}{14} = .500 \quad \ln \frac{7}{7} = .000$	0	1
	p	$\frac{1}{11} = .091 \quad \ln \frac{1}{10} = -2.303$	0	0

Równanie regresji logitowej będzie miało postać

$$lgt\,(p_{ij}) = \alpha + \beta^{(P)} P_{ij}^{(k)} + \beta^{(T)} T_{ij}^{(a)}$$

Po zastosowaniu metody najmniejszych kwadratów otrzymamy następujące estymatory

$$\widehat{\alpha} = -1.9037,\ \widehat{\beta^{(P)}} = 1.4687,\ \widehat{\beta^{(T)}} = 1.7817 \tag{3.1}$$

Z tych estymatorów możemy oszacować logity i prawdopodobieństwa oraz oczekiwane liczebności

		$\widehat{lgt}$	$\widehat{prawdop}$
P	T		$\widehat{p_{ij}}$
k	a	$-1.9037 + 1.4687 + 1.7817 = 1.346\,7$	$\frac{1}{1+\exp(-1.346\,7)} = .794$
	p	$-1.9037 + 1.4687 = -.435$	$\frac{1}{1+\exp(.435)} = .393$
m	a	$-1.9037 + 1.7817 = -.122$	$\frac{1}{1+\exp(.122)} = .470$
	p	$-1.9037 = -1.903\,7$	$\frac{1}{1+\exp(1.903\,7)} = .130$

$\widehat{n_{ijk}}$		W	
P	T	z	l
k	a	$27 - 21.438 = 5.562$	$27 * .794 = 21.438$
	p	$32 - 12.576 = 19.424$	$32 * .393 = 12.576$
m	a	$14 - 6.58 = 7.42$	$14 * .470 = 6.58$
	p	$11 - 1.43 = 9.57$	$11 * .130 = 1.43$

n_{ijk}		W	
P	T	z	l
k	a	*6*	*2*
	p	*19*	*1*
m	a	*7*	*7*
	p	*10*	*1*

G^2		W	
P	T	z	l
k	a	$6 \ln \frac{6}{5.562} = .454\,81$	$21 \ln \frac{21}{21.438} = -.433\,49$
	p	$19 \ln \frac{19}{19.424} = -.419\,34$	$13 \ln \frac{13}{12.57} = .437\,27$
m	a	$7 \ln \frac{7}{7.42} = -.407\,88$	$7 \ln \frac{7}{6.58} = .433\,13$
	p	$10 \ln \frac{10}{9.57} = .439\,52$	$1 \ln \frac{1}{1.43} = -.357\,67$

$G^2 = .292\,7$. *Dla 1 stopni swobody* $(1 = 4 - 3)$ *poziom krytyczny, odpowiadający* $G^2 = .292\,7$ *wynosi* 0.5885 *co oznacza niezłe dopasowanie do danych.*

Parametry równania regresji 3.1 pozwalają odpowiedzieć na niektóre pytania

- *Jaki wpływ ma płeć na prawdopodobieństwo wyleczenia?*

  Różnica logitów dla kobiet i mężczyzn przy tej samej terapii wynosi $\widehat{\beta^{(P)}} = 1.4687$*, co oznacza że stosunek szans lepszego wyniku jest dla kobiet* $\exp(1.4687) = 4.3$ *raza większy niż dla mężczyzn*

- *Jaki wpływ ma terapia na prawdopodobieństwo wyleczenia?*

  Różnica logitów dla terapii aktywnej i placebo dla tej samej płci chorego wynosi $\widehat{\beta^{(T)}} = 1.7817$*, co oznacza że stosunek szans lepszego wyniku jest dla terapii aktywnej* $\exp(1.7817) = 5.9$ *raza większy niż dla placebo.*

3.3 Regresja logitowa ze zmiennymi porządkowymi

Często zmienna wynikowa ma więcej niż dwie wartości. Jeśli te wartości występują w skali porządkowej, to do opisania ich zależnosci stosuje się *model proporcjonalnych szans.*

Model ten jest serią modeli logitowych, uporządkowanych według stopnia narastania intensywności cechy wynikowej. Na przykład, gdy cecha wynikowa X ma wartości *mały, średni, duży, olbrzymi* uporządkowane to modele logitowe byłyby utworzone według narastających poziomów dychotomicznych: *małyięcej niż mały; co najwyżej średniięcej niż średni;co najwyżej dużyięcej niż duży; mniej niż olbrzymiólbrzymi*

Proporcjonalność szans polega na tym, że wszystkie te modele tworzą *równoległe* hiperpłaszczyzny regresji. Oznacza to taki sam wpływ zmiennych objaśniających w każdej klasie intensywności cechy wynikowej. Zmiany prawdopodobieństw cechy wynikowej w tych klasach są niezależne od cech objaśniających.

Działanie modelu proporcjonalnych szans wyjaśnimy na przykładzie.

Przykład 3.5 *(kontynuacja przykładu 2.19) Przypomnimy dane:*

n_{ijk}		W		
P	T	z	u	i
k	a	6	5	16
	p	19	7	6
m	a	7	2	5
	p	10	0	1

Rozbijemy tę tablicę na dwie, zawierające dychotomiczne podziały zmiennej W: $z, -u$, gdzie l oznacza wyniki lepsze (umiarkowane lub istotne), $-u$ wyniki co najwyżej umiarkowane.

n_{ijk}		W	
P	T	z	l
k	a	6	21
	p	19	13
m	a	7	7
	p	10	1

n_{ijk}		W	
P	T	$-u$	i
k	a	11	16
	p	26	6
m	a	9	5
	p	10	1

Napiszemy model proporcjonalnych szans dla tych tablic

$$\begin{aligned} lgt\left(p_{ij}^{(1)}\right) &= \alpha_1 + \beta^{(P)} P_{ij}^{(k,1)} + \beta^{(T)} T_{ij}^{(a,1)} \\ lgt\left(p_{ij}^{(2)}\right) &= \alpha_2 + \beta^{(P)} P_{ij}^{(k,2)} + \beta^{(T)} T_{ij}^{(a,2)} \end{aligned}$$

W tych wzorach $p_{ij}^{(1)}, p_{ij}^{(2)}$ oznaczają prawdopodobieństwa odpowiednio wyniku z i $-u$ w tablicach 1 i 2; $P_{ij}^{(k,1)}, P_{ij}^{(k,2)}$ zmienne (indykatorowe) odpowiadające płci w tablicach; $T_{ij}^{(a,1)}, T_{ij}^{(a,2)}$ zmienne odpowiadające terapii.

Wprowadzając dwie zmienne indykatorowe $C^{(1)}, C^{(2)}$ wskazujące na numer tablicy można oba równania zapisać za pomocą jednego, co umożliwia wykorzystanie standardowego oprogramowania

$$lgt\left(p_{ij}^{(r)}\right) = \alpha_1 C^{(1)} + \alpha_2 C^{(2)} + \beta^{(P)} P_{ij}^{(k,r)} + \beta^{(T)} T_{ij}^{(a,r)}$$

Dane z tablicy, które umożliwiają estymację modelu przyjmą teraz postać:

lgt			$P_{ij}^{(k,r)}$	$T_{ij}^{(a,r)}$	$C^{(1)}$	$C^{(2)}$
P	T					
k	a	-1.253	1	1	1	0
	p	.379	1	0	1	0
m	a	.000	0	1	1	0
	p	2.303	0	0	1	0
k	a	$-.375$	1	1	0	1
	p	1.466	1	0	0	1
m	a	.588	0	1	0	1
	p	2.303	0	0	0	1

Parametry wyznaczone z tych danych metodą najmniejszych kwadratów są następujące

$$\alpha_1 = 1.91575, \alpha_2 = 2.55400, \beta^{(P)} = -1.24425, \beta^{(T)} = -1.87275$$

Model regresyjny dobrze pasuje do danych - jego współczynnik determinacji wynosi 0.9502.

Co można odczytać z danych?

Dla mężczyzn leczonych placebo, iloraz szans złych do lepszych wyników wynosi $\exp(1.91575) = 6.8$, *natomiast iloraz szans wyników co najwyżej umiarkowanych do istotnych wynosi* $\exp(2.55400) = 12.9$. *Obie te wielkości należy pomnożyć przez* $\exp(-1.24425) = .29$ *gdy badaną osobą jest kobieta, a przez* $\exp(-1.87275) = .15$ *gdy zastosowano terapię aktywną. Na przykład, gdy zastosuje się terapię aktywną u mężczyzn to iloraz szans złych do lepszych wyników wynosi* $6.8 * .15 = 1.0$ *natomiast iloraz szans wyników co najwyżej umiarkowanych do istotnych wynosi* $2.9 * .15 = 1.9$, *co jak widać dobrze świadczy o zastosowanej terapii. Dla kobiet, leczonych aktywnie, te wyniki są jeszcze lepsze: w pierwszym przypadku wynoszą* $1.0 * .29 = .29$ *a w drugim* $1.9 * .29 = .55$ *co wskazuje na przewagę prawdopodobieństwa wyników lepszych nad gorszymi na każdym poziomie oczekiwań.*

Rozdział 4

Modele logarytmiczno-liniowe

W poprzednich rozdziałach rozważaliśmy sytuacje, w których interesowała nas zależność czy niezależność *pary* cech. Jeżeli do pary cech dołączy trzecia, to powstaje układ, który jest bardziej skomplikowany, niż by to się z pozoru wydawało. Jednym z przejawów tej komplikacji jest tzw *paradoks Simpsona*[1]. Paradoks ten polega na tym, że dla trzech zdarzeń A, B, C jest możliwy układ nierówności

$$\begin{aligned} P(A|B\cap C) &< P(A|B^c\cap C), P(A|B\cap C^c) < P(A|B^c\cap C^c) \\ ale P(A|B) &> P(A|B^c) \end{aligned}$$

Paradoks ten ostrzega nas, że w rozważaniu relacji zdarzeń nie wystarczy udowodnić, że dana relacja zachodzi dla wszystkich przypadków (tu C i C^c). Konkluzja, jak widać może być inna.

Przykład 4.1 (Paradoks Simpsona) *(źródło:[1] str.136)*

Obrońca	**Ofiara**	**Kara śmierci**	
		Tak	**Nie**
Biały	Biały	19	132
	Murzyn	0	9
Murzyn	Biały	11	52
	Murzyn	6	97

Tabela 4.1: Kara śmierci i rasa

Niech A="orzeczono karę śmierci", B="Obrońca jest Biały", C="Ofiarą jest Biały". Łatwo obliczyć odpowiednie prawdopodobieństwa

$$\begin{aligned} P(A|B) &= \frac{19}{160} = .119,\ P(A|B^c) = \frac{17}{166} = .102\,,\ P(A|B) > P(A|B^c) \\ P(A|B\cap C) &= \frac{19}{151} = .126,\ P(A|B^c\cap C) = \frac{11}{63} = .175, \\ P(A|B\cap C^c) &= \frac{0}{9} = 0,\ P(A|B^c\cap C^c) = \frac{6}{103} = .059, \\ P(A|B\cap C) &< P(A|B^c\cap C)\,,\ P(A|B\cap C^c) < P(A|B^c\cap C^c) \end{aligned}$$

Definicja 4.2 *Dana jest tablica wyników obserwacji trzech cech X, Y, Z:*

Niech $p_{ijk} = P(X = x_i, Y = y_j, Z = z_k)$, oraz niech $m_{ijk} = n\,p_{ijk}$ (m_{ijk}jest oczekiwaną liczbą obserwacji w komórce tabeli)

[1]Nazwa tego paradoksu pochodzi od artykułu, opublikowanego przez E.H. Simpsona w 1951, choć zjawisko to było znane wcześniej, np było omawiane przez Yule'a w 1903.

X	**Y**	**Z**	
		z_1	z_2
x_1	y_1	n_{111}	n_{112}
	y_2	n_{121}	n_{122}
x_2	y_1	n_{211}	n_{212}
	y_2	n_{221}	n_{222}

Tabela 4.2: Tablica wyników obserwacji

Definicja 4.3 (Model logarytmiczno-liniowy) *Modelem logarytmiczno-liniowy nazywamy taki, w którym*

$$\ln m_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \tag{4.1}$$

oraz

$$\begin{aligned}
\sum_i \lambda_i^X &= 0, \sum_j \lambda_j^Y = 0, \sum_k \lambda_k^Z = 0, \\
\sum_i \lambda_{ij}^{XY} &= 0, \sum_j \lambda_{ij}^{XY} = 0, \\
\sum_j \lambda_{jk}^{YZ} &= 0, \sum_k \lambda_{jk}^{YZ} = 0, \\
\sum_i \lambda_{ik}^{XZ} &= 0, \sum_k \lambda_{ik}^{XZ} = 0, \\
\sum_i \lambda_{ijk}^{XYZ} &= 0, \sum_j \lambda_{ijk}^{XYZ} = 0, \sum_k \lambda_{ijk}^{XYZ} = 0,
\end{aligned} \tag{4.2}$$

Wielkości $\lambda_i^X, \lambda_j^Y, \lambda_k^Z$ nazywamy efektami głównymi, $\lambda_{ik}^{XZ}, \lambda_{ij}^{XY}, \lambda_{jk}^{YZ}$ efektami interakcji (interakcjami) rzędu 2, λ_{ijk}^{XYZ} efektami interakcji (interakcjami) rzędu 3.

Zapis $\ln m_{ijk}$ w postaci równań 4.1 i 4.2 nazywamy *zapisem bilansowym*. Zapis bilansowy jest układem równań liniowych.

Twierdzenie 4.4 *Dla każdego układu $\{m_{ijk}\}$ istnieje dokładnie jeden zapis bilansowy.*

Definicja 4.5 *Rozróżnia się modele logarytmiczno-liniowe:*
Model $[XYZ]$ nazywa się modelem nasyconym*, model* [] - stałym[2].

[2]W modelu stałym wszystkie prawdopodobieństwa p_{ijk} są równe.

Model	$\ln m_{ijk}$
$[XYZ]$	$\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$
$[XZ][XY][YZ]$	$\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
$[XZ][YZ]$	$\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
$[XY][Z]$	$\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$
$[X][Y][Z]$	$\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$
$[]$	μ

Tabela 4.3: Modele logarytmiczno-liniowe

Modele logarytmiczno liniowe, w przeciwieństwie do modeli logitowych, nie wyróżniają żadnej z cech. Ich zadaniem jest stworzenie jak najprostszego modelu, objaśniającego związki między występującymi cechami.

Twierdzenie 4.6 *Różne modele logarytmiczno-liniowe reprezentują różne typy zależności między cechami*

Model	**Typ zależności**	p_{ijk}
$[XZ][YZ]$	$X \perp Y \mid Z$	$\frac{p_{i+k}p_{+jk}}{p_{++k}}$
$[XY][Z]$	$(X,Y) \perp Z$	$p_{ij+}p_{++k}$
$[X][Y][Z]$	$X \perp Y \perp Z$	$p_{i++}\, p_{+j+}p_{++k}$

Tabela 4.4: Modele zależności

Dowód. $[XZ][YZ]$:

$$\begin{aligned}
\ln m_{ijk} &= \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \Longleftrightarrow \\
n\, p_{ijk} &= \alpha\, \beta_i^X \beta_j^Y \beta_k^Z \beta_{ik}^{XZ} \beta_{jk}^{YZ} \\
np_{i+k} &= \alpha\, \beta_i^X \beta_k^Z \beta_{ik}^{XZ} \sum_j \beta_j^Y \beta_{jk}^{YZ}, \\
np_{+jk} &= \alpha \beta_j^Y \beta_k^Z \beta_{jk}^{YZ} \sum_i \beta_i^X \beta_{ik}^{XZ}, \\
np_{++k} &= \alpha \beta_k^Z \sum_j \beta_j^Y \beta_{jk}^{YZ} \sum_i \beta_i^X \beta_{ik}^{XZ}, \\
n\frac{p_{i+k}p_{+jk}}{p_{++k}} &= \alpha\, \beta_i^X \beta_k^Z \beta_{ik}^{XZ} \sum_j \beta_j^Y \beta_{jk}^{YZ} \frac{\alpha \beta_j^Y \beta_k^Z \beta_{jk}^{YZ} \sum_i \beta_i^X \beta_{ik}^{XZ}}{\alpha \beta_k^Z \sum_j \beta_j^Y \beta_{jk}^{YZ} \sum_i \beta_i^X \beta_{ik}^{XZ}} = \\
&= \alpha\, \beta_i^X \beta_j^Y \beta_k^Z \beta_{ik}^{XZ} \beta_{jk}^{YZ} = n\, p_{ijk}
\end{aligned}$$

$[XY][Z]$:

$$\begin{aligned}
\ln m_{ijk} &= \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} \Longleftrightarrow n\, p_{ijk} = \alpha\, \beta_i^X \beta_j^Y \beta_k^Z \beta_{ij}^{XY} \\
n\, p_{ij+} &= \alpha\, \beta_i^X \beta_j^Y \beta_+^Z \beta_{ij}^{XY},\ n\, p_{++k} = \alpha\, \beta_k^Z \sum_{ij} \beta_i^X \beta_j^Y \beta_{ij}^{XY}, \\
n &= n\, p_{+++} = \alpha\, \beta_+^Z \sum_{ij} \beta_i^X \beta_j^Y \beta_{ij}^{XY} \\
n\, p_{ij+}\ p_{++k} &= \alpha\, \beta_i^X \beta_j^Y \beta_+^Z \beta_{ij}^{XY} \frac{\alpha\, \beta_k^Z \sum_{ij} \beta_i^X \beta_j^Y \beta_{ij}^{XY}}{n} = \\
&= \alpha\, \beta_i^X \beta_j^Y \beta_+^Z \beta_{ij}^{XY} \frac{\alpha\, \beta_k^Z \sum_{ij} \beta_i^X \beta_j^Y \beta_{ij}^{XY}}{\alpha\, \beta_+^Z \sum_{ij} \beta_i^X \beta_j^Y \beta_{ij}^{XY}} = n\, p_{ijk}
\end{aligned}$$

$[X][Y][Z]$:

$$\begin{aligned}
\ln m_{ijk} &= \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \Longleftrightarrow n\, p_{ijk} = \alpha\, \beta_i^X \beta_j^Y \beta_k^Z \\
n\, p_{i++} &= \alpha\, \beta_i^X \beta_+^Y \beta_+^Z,\ n\, p_{+j+} = \alpha\, \beta_+^X \beta_j^Y \beta_+^Z,\ n\, p_{++k} = \alpha\, \beta_+^X \beta_+^Y \beta_k^Z \\
n &= n\, p_{+++} = \alpha\, \beta_+^X \beta_+^Y \beta_+^Z \\
n\, p_{i++} p_{+j+} p_{++k} &= \alpha\, \beta_i^X \beta_+^Y \beta_+^Z \frac{\alpha\, \beta_+^X \beta_j^Y \beta_+^Z}{n} \frac{\alpha\, \beta_+^X \beta_+^Y \beta_k^Z}{n} = \\
&= \alpha\, \beta_i^X \beta_+^Y \beta_+^Z \frac{\alpha\, \beta_+^X \beta_j^Y \beta_+^Z}{\alpha\, \beta_+^X \beta_+^Y \beta_+^Z} \frac{\alpha\, \beta_+^X \beta_+^Y \beta_k^Z}{\alpha\, \beta_+^X \beta_+^Y \beta_+^Z} = \alpha\, \beta_i^X \beta_j^Y \beta_k^Z = n\, p_{ijk}
\end{aligned}$$

■

Wniosek 4.7 *W modelu $[XZ][YZ]$ cechy X i Y są niezależne warunkowo, to znaczy*

$$p_{ij|k} = p_{i+|k}\, p_{+j|k}$$

Dowód.

$$p_{ij|k} = \frac{p_{ijk}}{p_{++k}} = \frac{p_{i+k} p_{+jk}}{(p_{++k})^2} = \frac{p_{i+k}}{p_{+++}} \frac{p_{+jk}}{p_{++k}} = p_{i+|k} p_{+j|k}$$

■

Wniosek 4.8 *W modelu $[XY][Z]$ zachodzą relacje: $X \perp Z$, $Y \perp Z$*

Dowód. $p_{i+k} = \sum_j p_{ijk} = \sum_j p_{ij+} p_{++k} = p_{i++}\, p_{++k}$. Podobnie, $p_{+jk} = \sum_i p_{ijk} = \sum_i p_{ij+} p_{++k} = p_{+j+}\, p_{++k}$ ■

Uwaga 4.9 *Relacja $Y \perp Z \,|X$ nie implikuje relacji $Y \perp Z$*

Dowód. Dla dowodu wystarczy podać przykład .
Tablica przedstawia prawdopodobieństwa dla układu trzech cech:
X wykształcenie $\{s$ - ścisłe, h - humanistyczne$\}$,
Y płeć $\{k$ - kobieta, m -mężczyzna$\}$
Z zarobki $\{w$ - wysokie, n - niskie$\}$

X	Y	Z	
		w	n
s	k	.08	.02
	m	.32	.08
h	k	.12	.18
	m	.08	.12

$Y \perp Z \,|X = s$ gdyż w tym przypadku tablica prawdopodobieństw sprowadza się do tablicy

Y	Z	
	w	n
k	.16	.04
m	.64	.16

,

dla której iloraz krzyżowy wynosi $\theta = \frac{.16*.16}{.64*.04} = 1$ co oznacza niezależność. Podobnie,
$Y \perp Z \,|X = h$. W tym przypadku tablica prawdopodobieństw ma postać

Y	Z	
	w	n
k	.24	.36
m	.16	.24

,

dla której iloraz krzyżowy wynosi $\theta = \frac{.24*.24}{.16*.36} = 1$ co również oznacza niezależność. Natomiast tabela prawdopodobieństw dla pary cech (Y, Z), gdy nie znamy wartości X przedstawia się następująco:

Y	Z	
	w	n
k	.20	.20
m	.40	.20

,

dla której iloraz krzyżowy wynosi $\theta = \frac{.20*.20}{.40*.20} = .50$, co oznacza, że te cechy są zależne. ■

Lemat 4.10 *Stopnie swobody dla modeli prostych:*

$$\begin{aligned} P_1 &: \ln(m_{ijk}) = \mu, \\ P_2 &: \ln(m_{ijk}) = \lambda_i^X, \\ P_3 &: \ln(m_{ijk}) = \lambda_{ij}^{XY}, \\ P_4 &: \ln(m_{ijk}) = \lambda_{ijk}^{XYZ} \end{aligned}$$

wynoszą odpowiednio: $1, I-1, (I-1)(J-1), (I-1)(J-1)(K-1)$

Dowód. Liczba wolnych parametrów w modelu P_1 wynosi 1, gdyż w tym przypadku nie ma żadnych ograniczeń na wartość μ.

W modelu P_2 liczba wolnych parametrów wynosi $I-1$ gdyż mamy jedno ograniczenie $\sum_{i=1}^{I} \lambda_i^X = 0$.

W modelu P_3 liczba wolnych parametrów może być wyznaczona z tabeli

λ_{11}^{XY}	...	λ_{1j}^{XY}	...	*	0
...	...	...	...	...	...
λ_{i1}^{XY}	...	λ_{ij}^{XY}	...	*	0
...	...	...	...	...	...
*	*	*	...	*	0
0	...	0	...	0	0

pamiętając, że suma λ_{ij}^{XY} w wierszach i kolumnach jest równa 0, skąd wynika, że wystarczy wypełnić pola w miejscach nie zaznaczonych *. Pola z * muszą byc wypełnione taką wartością, aby suma wartości λ_{ij}^{XY} w wierszach i kolumnach była równa 0. Takich pól jest $(I-1)(J-1)$.

Podobnie w modelu P_4, tylko w tym przypadku mamy tablicę trójwymiarową, z ostatnimi wierszamiolumnamiarstwami wypełnionymi *, stąd liczba stopni swobody równa $(I-1)(J-1)(K-1)$. ■

Twierdzenie 4.11 *Estymatory największej wiarygodności dla liczby obserwacji w polach tablic wielodzielczych, odpowiadających efektom w modelu M o rozkładzie wielomianowym lub Poissona są równe obserwowanej liczbie obserwacji dla efektów. Estymatory te są wyznaczone jednoznacznie.*

Dowód. Dowód przeprowadzimy na przykładzie rozkładu wielomianowego i modelu $[XY][YZ]$. Dowód w każdym innym przypadku jest analogiczny. Nasz model oznacza zachodzenie równości

$$\ln m_{ijk} = \ln(np_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$$

Funkcja logarytmu wiarygodności w rozkładzie wielomianowym z dokładnością do stałych ma postać

$$\sum_{ijk} n_{ijk} \ln p_{ijk}$$

co, z dokładnością do stałych jest równe

$$\sum_{ijk} n_{ijk} \ln np_{ijk} = \sum_{ijk} n_{ijk} \left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}\right)$$

W zagadnieniu estymacji należy obliczyć maksimum powyższej funkcji przy ograniczeniach

$$\begin{aligned} 1 &= \sum_{ijk} p_{ijk} = \frac{1}{n} \sum_{ijk} m_{ijk}, \\ \sum_i \lambda_i^X &= 0, \sum_j \lambda_j^Y = 0, \sum_k \lambda_k^Z = 0, \\ \sum_i \lambda_{ij}^{XY} &= 0, \sum_j \lambda_{ij}^{XY} = 0, \sum_j \lambda_{jk}^{YZ} = 0, \sum_k \lambda_{jk}^{YZ} = 0 \end{aligned}$$

Potraktujemym_{ijk} jako funkcję zmiennych $\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{jk}^{YZ}$. Niech u będzie jedną z tych zmiennych. Wtedy

$$\begin{aligned} \frac{\partial m_{ijk}}{\partial u} &= \frac{\partial \exp\left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}\right)}{\partial u} = \\ & m_{ijk} \frac{\partial \left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}\right)}{\partial u} \end{aligned}$$

Wyrażenie $\frac{\partial \left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}\right)}{\partial u}$ jest równe 1 lub 0 w zależności od tego, czy u występuje, czy też nie występuje wśród $\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{jk}^{YZ}$.

Używając metody mnożników Lagrange'a należy znaleźć maksimum funkcji

$$\begin{aligned} F &= \sum_{ijk} n_{ijk} \left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}\right) + \\ &+ \alpha \sum_{ijk} m_{ijk} + \\ &+ \beta_1^X \sum_i \lambda_i^X + \beta_1^Y \sum_j \lambda_j^Y + \beta_1^Z \sum_k \lambda_k^Z + \\ &+ \sum_j \beta_{2j} \sum_i \lambda_{ij}^{XY} + \sum_i \beta_{3i} \sum_j \lambda_{ij}^{XY} + \\ &+ \sum_k \beta_{4k} \sum_j \lambda_{jk}^{YZ} + \sum_j \beta_{5j} \sum_k \lambda_{jk}^{YZ} \end{aligned}$$

Obliczamy pochodne względem nieznanych parametrów i przyrównamy je do 0

$$\begin{aligned} 0 &= \frac{\partial F}{\partial \mu} = \\ &= \sum_{ijk} n_{ijk} + \alpha \sum_{ijk} m_{ijk} = \\ &= n + \alpha \sum_{ijk} (np_{ijk}) = n(\alpha + 1) \Longrightarrow \alpha = -1 \end{aligned}$$

Dla λ_i^X

$$\begin{aligned} 0 &= \frac{\partial F}{\partial \lambda_i^X} = \\ &= \sum_{jk} n_{ijk} + \alpha \sum_{ijk} m_{ijk} + \beta_1^X = \\ &= n_{i++} - m_{i++} + \beta_1^X \end{aligned}$$

Dodając stronami po i powyższą równość, otrzymamy

$$\begin{aligned} 0 &= \sum_i \left(n_{i++} - m_{i++} + \beta_1^X \right) = n - \sum_i (np_{i++}) + n\beta_1^X = n\beta_1^X \\ \Longrightarrow & \quad \beta_1^X = 0 \end{aligned}$$

Stąd otrzymamy, że dla efektu λ_i^X zachodzi równość[3] $\widehat{n_{i++}} = n_{i++}$.

Podobnie,dla efektu λ_j^Y zachodzi równość $\widehat{n_{+j+}} = n_{+j+}$,dla efektu λ_k^Z zachodzi równość $\widehat{n_{++k}} = n_{++k}$

Analogiczne rachunki przeprowadzimy dla efektu λ_{ij}^{XY}

$$\begin{aligned} 0 &= \frac{\partial F}{\partial \lambda_{ij}^{XY}} = \sum_k n_{ijk} + \alpha \sum_k m_{ijk} + \beta_{2j} + \beta_{3i} = \\ &= n_{ij+} - m_{ij+} + \beta_{2j} + \beta_{3i} \end{aligned} \tag{4.3}$$

Sumując jak powyżej, najpierw po i, potem po j otrzymamy

$$\begin{aligned} 0 &= n_{+j+} - m_{+j+} + I\beta_{2j} + \beta_{3+} = I\beta_{2j} + \beta_{3+}, \\ 0 &= n_{i++} - m_{i++} + \beta_{2+} + J\beta_{3i} = \beta_{2+} + J\beta_{3i} \end{aligned} \tag{4.4}$$

[3]Zawsze symbolem $\widehat{\theta}$ oznaczać będziemy estymator parametru θ, uzyskany z maksymalizowania funkcji wiarygodności

Sumując teraz najpierw po j, potem po i otrzymamy

$$0 = I\beta_{2+} + J\beta_{3+}, \tag{4.5}$$

Z równań 4.4 mnożonych: pierwsze przez J, drugie przez I oraz dodanych stronami uzyskamy

$$IJ\,(\beta_{3i} + \beta_{2j}) + I\beta_{2+} + J\beta_{3+} = 0,$$

co w połączeniu z 4.5 daje, że $\beta_{2j} + \beta_{3i} = 0$ oraz, że w 4.3 zachodzi równość $\widehat{n_{ij+}} = n_{ij+}$.

W analogiczny sposób można pokazać, że dla efektu λ_{jk}^{YZ}, $\widehat{n_{+jk}} = n_{+jk}$ ■

Wniosek 4.12 *W modelu nasyconym estymatory największej wiarygodności $\widehat{n_{ijk}}$ spełniają równość*

$$\widehat{n}_{ijk} = n_{ijk}$$

dla każdego i, j, k.

Wniosek 4.13 *Zachodzą następujące implikacje:*

$$\begin{aligned}
\forall_{i,j,k}\ (\widehat{n}_{ijk} = n_{ijk}\,) \quad &\Longrightarrow \quad \forall_{i,j}\widehat{n}_{ij+} = n_{ij+,}\forall_{i,k}\widehat{n}_{i+k} = n_{i+k}, \forall_{j,k}\widehat{n}_{+jk} = n_{+jk}, \Longrightarrow \\
&\Longrightarrow \quad \forall_i\,\widehat{n}_{i++} = n_{i++,}\forall_j\,\widehat{n}_{+j+} = n_{+j+,}\forall_k\,\widehat{n}_{++k} = n_{++k,} \Longrightarrow \\
&\Longrightarrow \quad \widehat{n}_{+++} = n_{+++,}
\end{aligned}$$

Dowód. Oczywisty ■

4.1 Modele hierarchiczne

Niech M_1 będzie danym modelem logarytmiczno liniowym.

Definicja 4.14 *Model M_2 nazwiemy hierarchicznie podporządkowanym modelowi M_1 (w skrócie - podporządkowanym M_1, $M_2 \prec M_1$) gdy zbiór efektów w modelu M_2 jest podzbiorem zbioru efektów M_1.*

Definicja 4.15 *Odchyleniem modelu M_2 od M_1 nazywamy liczbę*

$$G^2\,(M_2\,|M_1\,) = 2\sum_i\sum_j\sum_k \widehat{n}_{ijk}^{(1)} \ln \frac{\widehat{n}_{ijk}^{(1)}}{\widehat{n}_{ijk}^{(2)}},$$

gdzie $\widehat{n}_{ijk}^{(r)}$ jest estymatorem największej wiarygodności n_{ijk} w modelu M_r $(r = 1, 2)$.

Zauważmy, że odchylenie danych od modelu logarytmiczno-liniowego jest równe odchyleniem tego modelu od modelu nasyconego.

Twierdzenie 4.16 *Gdy model M_1 jest prawdziwy to*

$$G^2\left(M_2 \,|M_1\right) = G^2\left(M_2\right) - G^2\left(M_1\right)$$

Co więcej,

$$DF\left(G^2\left(M_2 \,|M_1\right)\right) = DF\left(G^2\left(M_2\right)\right) - DF\left(G^2\left(M_1\right)\right)$$

Wniosek 4.17 *Jeżeli dany jest ciąg hierarchicznie podporządkowanych modeli*

$$M_0 \succ M_1 \succ ... \succ M_{k-1} \succ M_k$$

gdzie M_0 jest modelem nasyconym oraz modele $M_0, M_1, ..., M_{k-1}$ są prawdziwe, to zachodzi wzór

$$G^2\left(M_k\right) = \sum_{r=1}^{k} G^2\left(M_r \,|M_{r-1}\right)$$

z liczbą stopni swobody równą

$$DF\left(G^2\left(M_k\right)\right) = \sum_{r=1}^{k} DF\left(G^2\left(M_r \,|M_{r-1}\right)\right)$$

Dowód twierdzenia. Dowód przeprowadzimy w szczególnym przypadku, gdy

$$\begin{aligned} \ln\left(m_{ijk}^{(1)}\right) &= \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}, \\ \ln\left(m_{ijk}^{(2)}\right) &= \mu + \lambda_j^Y + \lambda_{ik}^{XZ} \end{aligned}$$

Wtedy

$$\begin{aligned} & G^2\left(M_2 \,|M_1\right) = 2\sum_{i,j,k} \widehat{n}_{ijk}^{(1)} \ln \frac{\widehat{n}_{ijk}^{(1)}}{\widehat{n}_{ijk}^{(2)}} \qquad (4.6) \\ = \; & 2\sum_{i,j,k} \widehat{n}_{ijk}^{(1)} \left(\left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}\right) - \left(\mu + \lambda_j^Y + \lambda_{ik}^{XZ}\right)\right) \\ = \; & 2\sum_{i,j,k} \widehat{n}_{ijk}^{(1)} \left(\lambda_i^X + \lambda_{ij}^{XY}\right) = 2\sum_i \widehat{n}_{i++}^{(1)} \lambda_i^X + 2\sum_{i,j} \widehat{n}_{ij+}^{(1)} \lambda_{ij}^{XY}. \end{aligned}$$

Z twierdzenia 4.11 wynika, że gdy model M_1 jest prawdziwy to estymatory największej wiarygodności dla liczby obserwacji, odpowiadających efektom λ_i^X oraz λ_{ij}^{XY} są równe obserwowanej liczbie obserwacji. Stąd $\widehat{n}_{i++}^{(1)} = n_{i++}$ oraz $\widehat{n}_{ij+}^{(1)} = n_{ij+}$ dla dowolnych i, j.

Wstawiając ostatnie równości do wzoru 4.6 i zwijając ten wzór od tyłu, otrzymamy

$$\begin{aligned} & 2\sum_i \widehat{n}_{i++}^{(1)} \lambda_i^X + 2\sum_{i,j} \widehat{n}_{ij+}^{(1)} \lambda_{ij}^{XY} \\ = \; & 2\sum_i n_{i++} \lambda_i^X + 2\sum_{i,j} n_{ij+} \lambda_{ij}^{XY} \\ = \; & 2\sum_{i,j,k} n_{ijk} \left(\left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \right) - \left(\mu + \lambda_j^Y + \lambda_{ik}^{XZ} \right) \right) \\ = \; & 2\sum_{i,j,k} n_{ijk} \ln \frac{n_{ijk}}{\widehat{n}_{ijk}^{(2)}} - 2\sum_{i,j,k} n_{ijk} \ln \frac{n_{ijk}}{\widehat{n}_{ijk}^{(1)}} = G^2\left(M_2\right) - G^2\left(M_1\right). \end{aligned}$$

Liczba stopni swobody w modelu $M_2 \,|M_1$ jest równa (patrz Lemat 4.10) $(I-1)+(I-1)(J-1)$, czyli różnicy

$$1 + (I-1) + (J-1) + (I-1)(J-1) + (I-1)(K-1)$$

i

$$1 + (J-1) + (I-1)(K-1)$$

co dowodzi drugiej części tezy twierdzenia.

Dowód w każdym innym przypadku jest analogiczny. ■

Twierdzenie 4.18 *Utwórzmy ciąg hierarchicznie podporządkowanych modeli:*

$$\begin{aligned} M_0 &: [XYZ] \\ M_1 &: [XY][XZ][YZ] \\ M_2 &: [XY][YZ] \\ M_3 &: [XY][Z] \\ M_4 &: [X][Y][Z] \end{aligned}$$

Wtedy

$$\begin{aligned} DF\left(M_1 \,|M_0\right) &= (I-1)(J-1)(K-1) \\ DF\left(M_2 \,|M_1\right) &= (I-1)(K-1) \\ DF\left(M_3 \,|M_2\right) &= (J-1)(K-1) \\ DF\left(M_4 \,|M_3\right) &= (I-1)(J-1) \end{aligned}$$

gdzie I, J, K jest liczbą różnych wartości cech X, Y, Z.

Dowód. Model M_0 (nasycony) jest postaci $[XYZ]$, co oznacza, że

$$\ln\left(m_{ijk}^{(0)}\right) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

Model M_1 postaci $[XY][XZ][YZ]$ ma postać:

$$\ln\left(m_{ijk}^{(1)}\right) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

Odchylenie $G^2\left(M_1 \,|M_0\right)$ jest statystyką testową w układzie hipotez:

$$\begin{aligned} H_0 &: prawdziwyjestmodelM_1, \\ H_1 &: prawdziwyjestmodelM_0 \end{aligned}$$

Liczba stopni swobody dla takiego układu hipotez jest różnicą $DF\left(H_1\right) - DF\left(H_0\right)$.

Liczba stopni swobody modelu M_0 wynosi

$$1 + I - 1 + J - 1 + K - 1 + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) + (I-1)(J-1)(K-1)$$

Podobnie, liczba stopni swobody modelu M_1 wynosi

$$1 + I - 1 + J - 1 + K - 1 + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1).$$

Jak łatwo zobaczyć, różnica tych liczb wynosi $(I-1)(J-1)(K-1)$, czyli jest liczbą stopni swobody prostego modelu λ_{ijk}^{XYZ}, który występuje w M_0 a nie występuje w M_1. W podobny sposób można uzasadnić pozostałe wzory w tezie twierdzenia. ■

Uwaga 4.19 (praktyczna) *Liczba stopni swobody w modelu warunkowym $M_{r+1}\,|M$ liczbą stopni swobody w modelu prostym, który występuje w M_r a nie występuje w M_{r+1}.*

Twierdzenie 4.20 *Estymatory największej wiarygodności $n_{ijk}^{(r+1)}$ w modelach hierarchicznych $M_{r+1}\,|M_r$ (patrz Twierdzenie 4.18) wyrażają się wzorami*

$$\begin{aligned} n_{ijk}^{(2)} &= \frac{n_{ij+}^{(1)}\, n_{+jk}^{(1)}}{n_{+j+}^{(1)}} \\ n_{ijk}^{(3)} &= \frac{n_{ij+}^{(2)}\, n_{++k}^{(2)}}{n_{+++}^{(2)}} \\ n_{ijk}^{(4)} &= \frac{n_{i++}^{(3)}\, n_{+j+}^{(3)} n_{++k}^{(3)}}{\left(n_{+++}^{(3)}\right)^2} \end{aligned}$$

Estymatory $n^{(1)}_{ijk}$ można wyznaczyć metodą iteracyjnego oszacowania proporcjonalnego (Dodatek A)

Dowód. Model $M_2\,|M_1$,postaci $[XY][YZ]$, jest modelem warunkowej niezależności $X \perp Z\,|Y$ (Twierdzenie 4.6), co oznacza, że

$$p^{(2)}_{ik|j} = p^{(2)}_{i+|j}\, p^{(2)}_{+k|j}$$

czyli równoważnie

$$\frac{p^{(2)}_{ijk}}{p^{(2)}_{+j+}} = \frac{p^{(2)}_{ij+}}{p^{(2)}_{+j+}} \frac{p^{(2)}_{+jk}}{p^{(2)}_{+j+}}$$

Mnożąc obie strony tego równania przez $n^{(2)}_{+++}$ otrzymamy, po uproszczeniach

$$n^{(2)}_{ijk} = n^{(2)}_{ij+} \frac{p^{(2)}_{+jk}}{p^{(2)}_{+j+}}$$

Mnożąc teraz licznik i mianownik ułamka po prawej stronie przez $n^{(2)}_{+++}$, otrzymamy równość:

$$n^{(2)}_{ijk} = \frac{n^{(2)}_{ij+} n^{(2)}_{+jk}}{n^{(2)}_{+j+}}$$

Korzystając z twierdzenia4.11 mamy, że $n^{(2)}_{ij+} = n^{(1)}_{ij+}$, $n^{(2)}_{+jk} = n^{(1)}_{+jk}$, $n^{(2)}_{+j+} = n^{(1)}_{+j+}$

Analogicznie, model $M_3\,|M_2$,postaci $[XY][Z]$, jest modelem niezależności pary (X, Y) i Z. Korzystając znów z twierdzenia 4.6 mamy

$$p^{(3)}_{ijk} = p^{(3)}_{ij+}\, p^{(3)}_{++k}$$

co po analogicznych operacjach, jak wyżej (mnożenie obustronne przez $n^{(3)}_{+++}$, potem mnożenie i dzielenie po prawej stronie przez $n^{(3)}_{+++}$ i wykorzystanie twierdzenia **??**) daje

$$n^{(3)}_{ijk} = \frac{n^{(2)}_{ij+}\, n^{(2)}_{++k}}{n^{(2)}_{+++}}$$

Ostatnią równość w tezie twierdzenia uzyskuje się w analogiczny sposób. ■

Uwaga 4.21 (praktyczna) *Wyniki, uzyskane w tym punkcie możemy podsumować w tabeli*

Model
$M_0 : [XYZ]$
$M_1 : [XY][XZ][YZ]$
$M_2 : [XY][YZ]$
$M_3 : [XY][Z]$
$M_4 : [X][Y][Z]$

Model warunkowy	**Typ zależności**	Estymacja	**DF**
-	nasycony		0
$M_1 \mid M_0$	-	IPF	$(I-1)(J-1)(K-1)$
$M_2 \mid M_1$	$X \perp Z \mid Y$	$\frac{n^{(1)}_{ij+}\, n^{(1)}_{+jk}}{n^{(1)}_{+j+}}$	$(I-1)(K-1)$
$M_3 \mid M_2$	$(X,Y) \perp Z$	$\frac{n^{(2)}_{ij+}\, n^{(2)}_{++k}}{n^{(2)}_{+++}}$	$(J-1)(K-1)$
$M_4 \mid M_3$	$X \perp Y \perp Z$	$\frac{n^{(3)}_{i++}\, n^{(3)}_{+j+} n^{(3)}_{++k}}{\left(n^{(3)}_{+++}\right)^2}$	$(I-1)(J-1)$

Tabela 4.5: Dopasowanie modelu hierarchicznego

Przykład 4.22 (artretyzm, terapia, płeć) *(c.d. przykładu 2.19)*

Zbadamy strukturę tych danych, stosując model logarytmiczno-liniowy na poziomie istotności 0,05

$n^{(0)}_{ijk}$		W	
P	T	z	l
k	a	*6*	*21*
	p	*19*	*13*
m	a	*7*	*7*
	p	*10*	*1*

Oszacujemy, metodą IPF liczebności $n^{(1)}_{ijk}$ *dla modelu* $[PW][TW][PT]$

$w_{ijk}^{(0)}$		z	l
k	a	*1*	*1*
	p	*1*	*1*
m	a	*1*	*1*
	p	*1*	*1*

Najpierw dopasujemy model $[PW]$

$n_{i+k}^{(0)}$		
k	z	*25*
	l	*34*
m	z	*17*
	l	*8*

$w_{i+k}^{(0)}$		
k	z	*2*
	l	*2*
m	z	*2*
	l	*2*

α_{i+k}		
k	z	$\frac{25}{2} = 12.5$
	l	$\frac{34}{2} = 17.0$
m	z	$\frac{17}{2} = 8.5$
	l	$\frac{8}{2} = 4.0$

Po uwzględnieniu współczynnika skalującego otrzymamy nową macierz:

$w_{ijk}^{(1)}$		z	l
k	a	$1 * 12.5$	$1 * 17.0$
	p	$1 * 12.5$	$1 * 17.0$
m	a	$1 * 8.5$	$1 * 4.0$
	p	$1 * 8.5$	$1 * 4.0$

$=$

$w_{ijk}^{(1)}$		z	l
k	a	12.5	17.0
	p	12.5	17.0
m	a	8.5	4.0
	p	8.5	4.0

W drugim kroku pierwszego cyklu dopasujemy model $[TW]$

$n_{+jk}^{(0)}$		
a	z	*13*
	l	*28*
p	z	*29*
	l	*14*

$w_{+jk}^{(1)}$		
a	z	$12.5 + 8.5$
	l	$17.0 + 4.0$
p	z	$12.5 + 8.5$
	l	$17.0 + 4.0$

α_{+jk}		
a	z	$\frac{13}{21} = .619$
	l	$\frac{28}{21} = 1.333$
p	z	$\frac{29}{21} = 1.381$
	l	$\frac{14}{21} = .667$

$w_{ijk}^{(2)}$		z	l
k	a	$12.5 * .619$	$17.0 * 1.333$
	p	$12.5 * 1.381$	$17.0 * .667$
m	a	$8.5 * .619$	$4.0 * 1.333$
	p	$8.5 * 1.381$	$4.0 * .667$

$=$

$w_{ijk}^{(2)}$		z	l
k	a	7.74	22.66
	p	17.26	11.34
m	a	5.26	5.32
	p	11.74	2.67

W trzecim kroku pierwszego cyklu dopasujemy model $[PT]$

$n^{(0)}_{ij+}$		
k	a	*27*
	p	*32*
m	a	*14*
	p	*11*

$w^{(2)}_{ij+}$		
k	a	$7.74 + 22.66$
	p	$17.26 + 11.34$
m	a	$5.26 + 5.32$
	p	$11.74 + 2.67$

α_{ij+}		
k	a	$\frac{27}{30.4} = .889$
	p	$\frac{32}{28.6} = 1.119$
m	a	$\frac{14}{10.58} = 1.32$
	p	$\frac{11}{14.41} = .763$

$w^{(3)}_{ijk}$		z	l
k	a	$7.74 * .889$	$22.66 * .889$
	p	$17.26 * 1.119$	$11.34 * 1.119$
m	a	$5.26 * 1.323$	$5.32 * 1.323$
	p	$11.74 * .763$	$2.67 * .763$

$=$

$w^{(3)}_{ijk}$		z	l
k	a	6.89	20.14
	p	19.31	12.69
m	a	6.96	7.04
	p	8.96	2.04

Rozpoczynamy drugi cykl iteracji

Model [PW]

$w^{(3)}_{i+k}$		
k	z	$6.89 + 19.31$
	l	$20.14 + 12.69$
m	z	$6.96 + 8.96$
	l	$7.04 + 2.04$

α_{i+k}		
k	z	$\frac{25}{26.2} = .954$
	l	$\frac{34}{32.83} = 1.036$
m	z	$\frac{17}{15.92} = 1.068$
	l	$\frac{8}{9.08} = .881$

$w^{(4)}_{ijk}$		z	l
k	a	$6.89 * .954$	$20.14 * 1.036$
	p	$19.31 * .954$	$12.69 * 1.036$
m	a	$6.96 * 1.068$	$7.04 * .881$
	p	$8.96 * 1.068$	$2.04 * .881$

$=$

$w^{(4)}_{ijk}$		z	l
k	a	6.57	20.86
	p	18.42	13.15
m	a	7.43	6.20
	p	9.57	1.80

Model [TW]

$w^{(4)}_{+jk}$		
a	z	$6.57 + 7.43$
	l	$20.86 + 6.20$
p	z	$18.42 + 9.57$
	l	$13.15 + 1.80$

α_{+jk}		
a	z	$\frac{13}{14.0} = .929$
	l	$\frac{28}{27.06} = 1.035$
p	z	$\frac{29}{27.99} = 1.036$
	l	$\frac{14}{14.95} = .936$

$w^{(5)}_{ijk}$		z	l
k	a	$6.57 * .929$	$20.86 * 1.035$
	p	$18.42 * 1.036$	$13.15 * .936$
m	a	$7.43 * .929$	$6.20 * 1.035$
	p	$9.57 * 1.036$	$1.80 * .936$

$=$

$w^{(5)}_{ijk}$		z	l
k	a	6.10	21.59
	p	19.08	12.31
m	a	6.90	6.42
	p	9.91	1.68

Model $[PT]$

$w^{(5)}_{ij+}$		
k	a	$6.10 + 21.59$
	p	$19.08 + 12.31$
m	a	$6.90 + 6.42$
	p	$9.91 + 1.68$

α_{ij+}		
k	a	$\frac{27}{27.69} = .975$
	p	$\frac{32}{31.39} = 1.019$
m	a	$\frac{14}{13.32} = 1.051$
	p	$\frac{11}{11.59} = .949$

$w^{(6)}_{ijk}$		z	l
k	a	$6.10 * .975$	$21.59 * .975$
	p	$19.08 * 1.019$	$12.31 * 1.019$
m	a	$6.90 * 1.051$	$6.42 * 1.051$
	p	$9.91 * .949$	$1.68 * .949$

$=$

$w^{(6)}_{ijk}$		z	l
k	a	5.95	21.05
	p	19.44	12.54
m	a	7.25	6.75
	p	9.40	1.59

Obliczenia w tym modelu zatrzymujemy po dwóch cyklach[4].

Przyjmiemy więc tabelę wartościami $w^{(6)}_{ijk}$ *jako tabelę z estymatorami* $n^{(1)}_{ijk}$ *dla modelu* $[PW][TW][PT]$*:*

$n^{(1)}_{ijk}$		z	l
k	a	5.95	21.05
	p	19.44	12.54
m	a	7.25	6.75
	p	9.40	1.59

$G^2_{ijk}(M_1 \mid M_0)$		z	l
k	a	$6 \ln \frac{6}{5.95}$	$21 \ln \frac{21}{21.05}$
	p	$19 \ln \frac{19}{19.44}$	$13 \ln \frac{13}{12.54}$
m	a	$7 \ln \frac{7}{7.25}$	$7 \ln \frac{7}{6.75}$
	p	$10 \ln \frac{10}{9.40}$	$1 \ln \frac{1}{1.59}$

$\Longrightarrow G^2_{ijk}(M_1 \mid M_0) = .395\,16$

Poziom krytyczny, odpowiadający wartości $.395\,16$ *dla rozkładu* χ^2 *z 1 stopniem swobody (* $(I-1)(J-1)(K-1) = 1$ *) wynosi* $0,5296$ *co upoważnia nas do zaakceptowania modelu* M_1*.*

Oszacujemy teraz parametry modelu $M_2 \mid M_1$ *gdzie* $M_2 : [PW][TW]$*. Od razu możemy obliczyć estymatory* $n^{(2)}_{ijk}$ *w tym modelu (patrz tabela 4.5) ze wzoru* $n^{(2)}_{ijk} = \frac{n^{(1)}_{i+k} n^{(1)}_{+jk}}{n^{(1)}_{++k}}$:

[4]Kryteria stopu zależą od wybranej opcji. Może to być dokładność liczności brzegowych czy też, jak w naszym przykładzie, liczba cykli obliczeń.

$n^{(1)}_{i+k}$		
k	z	25.39
	l	33.59
m	z	16.65
	l	8.34

$n^{(1)}_{+jk}$		
a	z	13.20
	l	27.80
p	z	28.84
	l	14.13

$n^{(1)}_{++k}$	
z	42.04
l	41.93

$n^{(2)}_{ijk}$		z	l
k	a	$\frac{25.39*13.20}{42.04}$	$\frac{33.59*27.80}{41.93}$
	p	$\frac{25.39*28.84}{42.04}$	$\frac{33.59*14.13}{41.93}$
m	a	$\frac{16.65*13.20}{42.04}$	$\frac{8.34*27.80}{41.93}$
	p	$\frac{16.65*28.84}{42.04}$	$\frac{8.34*14.13}{41.93}$

$=$

$n^{(2)}_{ijk}$		z	l
k	a	7.97	22.27
	p	17.42	11.32
m	a	5.23	5.53
	p	11.42	2.81

$G^2_{ijk}(M_2 \mid M_1)$		z	l
k	a	$5.95 \ln \frac{5.95}{7.97}$	$21.05 \ln \frac{21.05}{22.27}$
	p	$19.44 \ln \frac{19.44}{17.42}$	$12.54 \ln \frac{12.54}{11.32}$
m	a	$7.25 \ln \frac{7.25}{5.23}$	$6.75 \ln \frac{6.75}{5.53}$
	p	$9.40 \ln \frac{9.40}{11.42}$	$1.59 \ln \frac{1.59}{2.81}$

$\Longrightarrow G^2_{ijk}(M_2 \mid M_1) = 2.9388 \Longrightarrow G^2_{ijk}(M_2) = G^2_{ijk}(M_2 \mid M_1) + G^2_{ijk}(M_1 \mid M_0)$ $= .39516 + 2.9388 = 3.334$

Poziom krytyczny, odpowiadający wartości 3.334 *dla rozkładu* χ^2 *z 2 stopniami swobody (* $(I-1)(J-1)(K-1) + (I-1)(K-1) = 2$ *) wynosi* $0,1888$ *co upoważnia nas do zaakceptowania modelu* M_2.

Oszacujemy teraz parametry modelu $M_3 \mid M_2$ *gdzie* $M_3 : [P][TW]$. *Możemy obliczyć estymatory* $n^{(3)}_{ijk}$ *w tym modelu (patrz tabela 4.5) ze wzoru*

$$n^{(3)}_{ijk} = \frac{n^{(2)}_{i++}\, n^{(2)}_{+jk}}{n^{(2)}_{+++}}$$

$n^{(2)}_{ijk}$		z	l
k	a	7.97	22.27
	p	17.42	11.32
m	a	5.23	5.53
	p	11.42	2.81

$n^{(2)}_{i++}$	
k	58.98
m	24.99

$n^{(2)}_{+jk}$		
a	z	13.20
	l	27.80
p	z	28.84
	l	14.13

$n^{(2)}_{+++}$	83.97

$n^{(3)}_{ijk}$		z	l
k	a	$\frac{58.98*13.20}{83.97}$	$\frac{58.98*27.80}{83.97}$
	p	$\frac{58.98*28.84}{83.97}$	$\frac{58.98*14.13}{83.97}$
m	a	$\frac{24.99*13.20}{83.97}$	$\frac{24.99*27.80}{83.97}$
	p	$\frac{24.99*28.84}{83.97}$	$\frac{24.99*14.13}{83.97}$

$=$

$n^{(3)}_{ijk}$		z	l
k	a	9.27	19.53
	p	20.26	9.92
m	a	3.93	8.27
	p	8.58	4.21

$G^2_{ijk}(M_3 \mid M_2)$		z	l
k	a	$7.97 \ln \frac{7.97}{9.27}$	$22.27 \ln \frac{22.27}{19.53}$
	p	$17.42 \ln \frac{17.42}{20.26}$	$11.32 \ln \frac{11.32}{9.92}$
m	a	$5.23 \ln \frac{5.23}{3.93}$	$5.53 \ln \frac{5.53}{8.27}$
	p	$11.42 \ln \frac{11.42}{8.58}$	$2.81 \ln \frac{2.81}{4.21}$

$\Longrightarrow G^2_{ijk}(M_3 \mid M_2) = 3.9628 \Longrightarrow G^2_{ijk}(M_3) = 3.9628 + 3.334 = 7.2968$

Poziom krytyczny, odpowiadający wartości 7.2968 *dla rozkładu* χ^2 *z 3 stopniami swobody (* $2 + (I-1)(K-1) = 3$*) wynosi* $0,06302$ *co upoważnia nas do zaakceptowania modelu* M_3.

Oszacujemy teraz parametry modelu $M_4 \mid M_3$ *gdzie* $M_3 : [P][T][W]$. *Estymatory* $n^{(4)}_{ijk}$ *możemy obliczyć ze wzoru*

$$n^{(4)}_{ijk} = \frac{n^{(3)}_{i++}\, n^{(3)}_{+j+}\, n^{(3)}_{++k}}{\left(n^{(3)}_{+++}\right)^2}$$

$n^{(3)}_{i++}$	
k	58.98
m	24.99

$n^{(3)}_{+j+}$	
a	41.0
p	42.97

$n^{(3)}_{++k}$	
z	42.04
l	41.93

$n^{(3)}_{+++}$	83.97

$n^{(4)}_{ijk}$		z	l
k	a	$\frac{58.98*41.0*42.04}{83.97^2}$	$\frac{58.98*41.0*41.93}{83.97^2}$
	p	$\frac{58.98*42.97*42.04}{83.97^2}$	$\frac{58.98*42.97*41.93}{83.97^2}$
m	a	$\frac{24.99*41.0*42.04}{83.97^2}$	$\frac{24.99*41.0*41.93}{83.97^2}$
	p	$\frac{24.99*42.97*42.04}{83.97^2}$	$\frac{24.99*42.97*41.93}{83.97^2}$

$=$

$n^{(4)}_{ijk}$		z	l
k	a	14.42	14.38
	p	15.11	15.07
m	a	6.11	6.09
	p	6.40	6.39

$G^2_{ijk}(M_4\,\vert M_3)$		z	l
k	a	$9.27 \ln \frac{9.27}{14.42}$	$19.53 \ln \frac{19.53}{14.38}$
	p	$20.26 \ln \frac{20.26}{15.11}$	$9.92 \ln \frac{9.92}{15.07}$
m	a	$3.93 \ln \frac{3.93}{6.11}$	$8.27 \ln \frac{8.27}{6.09}$
	p	$8.58 \ln \frac{8.58}{6.40}$	$4.21 \ln \frac{4.21}{6.39}$

$\Longrightarrow G^2_{ijk}(M_4\,\vert M_3) = 10.462$

$\Longrightarrow G^2_{ijk}(M4) = 10.462 + 7.2968 = 17.759$

Poziom krytyczny, odpowiadający wartości 17.759 *dla rozkładu* χ^2 *z 4 stopniami swobody (* $3 + (J-1)\,(K-1) = 4$*) wynosi* $0,0014$ *co upoważnia nas do odrzucenia modelu* M_4*.*

Ostatecznie możemy przyjąć, że na poziomie istotności 0.05 *modelem, opisującym dane jest* $[P][TW]$*, co oznacza , że związane ze sobą są wyniki leczenia i zastosowana terapia. Wybór pacjentów wg kryteriów płci ani nie był związany z wyborem zastosowanej terapii, ani z uzyskanymi wynikami.*

Gdybyśmy przeprowadzili rozumowanie na poziomie 0.1[5] *to ostatnim zaakceptowanym modelem byłby* $[PW][TW]$ *z poziomem krytycznym* $0,1661$*. Model taki oznacza, że przy każdych danych wynikach leczenia nie ma związku między płcią a wyborem terapii, natomiast zarówno płeć jak i terapia mogą mieć wpływ na wyniki leczenia*[6]*.*

Oszacowany przez nas model danych nie musi być jedynym. Poszliśmy jedną z możliwych ścieżek w drzewku modeli hierarchicznych. Przypuśćmy, jak to robią pakiety statystyczne, że oszacowaliśmy wszystkie dopuszczalne modele na wybranym poziomie istotności. Który z nich wybrać? Jednym z używanych w statystyce kryteriów jest kryterium AIC, podane przez Akaike czy też kryterim bayesowskie BIC. Pozwalają one wybrać ten model, który jednocześnie najlepiej pasuje do danych i jest najoszczędniejszy w swoim opisie. Wybiera się więc ten model, który ma większą wartość kryterium.Dla modeli logarytmiczno - liniowych (p.[1] str. 251) można te kryteria wyrazić wzorami

$$\begin{aligned} AIC(M) &= G^2(M) - 2DF(M), \\ BIC(M) &= G^2(M) - \ln(n_M)\, DF(M), \\ & \quad gdzien_M jestliczbobserwacjidlamodeluM \end{aligned}$$

[5] co często jest przyjmowane w programach statystycznych jako wartość domyślna (np. w programie *Statistica*)

[6] Patrz też wyniki modelu logitowego dla tych danych

W rozważanym przykładzie wartość kryterium Akaike zmieniała się następująco:

$$\begin{aligned} AIC\,(M_1) &= 0.39516 - 2 * 1 = -1.6048, \\ AIC\,(M_2) &= 3.334 - 2 * 2 = -.666 \\ AIC\,(M_3) &= 7.2968 - 2 * 3 = 1.2968 \end{aligned}$$

Dodatek A

Skale dla prawdopodobieństw

Definicja A.1 *Przypuśćmy, że obserwowana wielkość X jest wyrażona w jakiejś skali liczbowej. Skalą dla wielkości X nazywamy każdą rosnącą i ciągłą funkcję H. Wartości X w nowej skali są równe $H(X)$*

Wymóg ścisłego wzrostu skali jest zrozumiały - wartości obserwowanego zjawiska wyrażone w nowej skali powinny zachować porządek skali początkowej. Podobnie, ciągłość oznacza, że wartości bliskie w skali początkowej będą bliskie w nowej skali. Różnowartościowość funkcji H umożliwia powrót z nowej skali do skali początkowej.

Uwaga A.2 *Złożenie skal H_1 i H_2 jest skalą. W szczególności złożenie skali liniowej $H_1 = \alpha + \beta u$ $(\beta > 0)$ jest skalą. Nałożenie skali liniowej umożliwia wybór zera i jednostki każdej skali.*

Definicja A.3 *Skala prawdopodobieństw to funkcja rosnąca i ciągła*[1]

$$H : (0,1) \longrightarrow R$$

Definicja A.4 *Skala prawdopodobieństw jest symetryczna gdy $H(1-p) = -H(p)$*

Uwaga A.5 *Dla skali symetrycznej $H\left(\frac{1}{2}\right) = 0$*

Twierdzenie A.6 *Każdą skalę można zsymetryzować*

$$H^0(p) = H(p) - H(1-p)$$

Dowód. 1. H^0 jest funkcją ciągłą, bo jest różnicą funkcji ciągłych.
2. Niech $p_1 < p_2$. $H^0(p_1) = H(p_1) - H(1-p_1) < H(p_2) - H(1-p_2) = H^0(p_2)$ (funkcja $-H(1-p)$ jest rosnąca)
3. H^0 jest symetryczna: $H^0(1-p) = H(1-p) - H(1-(1-p)) = -H^0(p)$
■

Przykład A.7 (Skale kwantylowe) *Niech F będzie rosnącą i ciągłą dystrybuantą rozkładu zmiennej losowej.*

Lewostronna skala kwantylowa *oparta na F jest funkcją*

$$H_L(p) = F^{-1}(p)$$

Prawostronna skala kwantylowa *oparta na F jest funkcją*

$$H_P(p) = -F^{-1}(1-p)$$

[1]Zazwyczaj definiuje się skalę dla przedziału otwartego, wykluczając z rozważań zdarzenia niemożliwe i pewne

Uwaga A.8 *Niech F będzie rosnącą i ciągłą dystrybuantą rozkładu prawdopodobieństwa, symetrycznego w zerze. Wtedy:*

1. lewostronna i prawostronna skala kwantylowa jest symetryczna,

2. dla każdego p, $H_L(p) = H_P(p)$

Dowód. 1. Niech $H_L(p) = u$, $H_L(1-p) = v$. Wtedy $F(u) = p$, $F(v) = 1-p$. Z definicji rozkładu symetrycznego w 0 mamy, że $v = -u$. Podobnie, niech $H_P(p) = u$, $H_P(1-p) = v$. Wtedy $F(-u) = 1-p, F(-v) = p$ co implikuje równość $v = -u$.

2. Niech $H_L(p) = u$, $H_P(p) = v$. Wtedy $F(u) = p, F(-v) = 1-p$. Z tej równości i symetrii wynika, że $v = u$. ■

Definicja A.9 *Skalę kwantylową opartą na dystrybuancie Φ rozkładu normalnego standardowego*[2] *nazywamy* ***skalą probitową***

Skalę probitową stosujemy dla zjawisk o rozkładzie prawdopodobieństwa symetrycznie rozłożonym wokół wartości $\frac{1}{2}$ i niezbyt daleko odbiegającym od tej wartości.

Dla zjawisk, w których obserwujemy zjawiska ekstremalne (np. śmiertelność owadów na skutek stosowania środków chemicznych) stosuje się prawo i lewostronną skalę kwantylową opartą na rozkładzie Gumbela[3] o dystrybuancie

$$F(u) = \exp(-\exp(-u))$$

Wtedy $H_L(p) = -\ln(-\ln(p))$, $H_P(p) = \ln(-\ln(1-p))$. Takie przekształcenie nazywane jest *skalą podwójnie logarytmiczną*. Jak łatwo zauważyć skala podwójnie logarytmiczna nie jest symetryczna.

Najczęściej, ze względu na swoją prostotę i dopasowanie do często występujących w praktyce zjawisk asymetrycznych[4] jest skala logitowa.

Definicja A.10 ***Skala logitowa*** *jest symetryzacją skali logarytmicznej dla prawdopodobieństw*

$$lgt(p) = \ln(p) - \ln(1-p) = \ln\left(\frac{p}{1-p}\right)$$

[2]Dystrybuanta ta jest ciągła i rosnąca, a rozkład jest symetryczny w 0.

[3]Rozkład Gumbela jest jednym z trzech możliwych rozkładów granicznych dla wartości największej z ciągu niezależnych zmiennych losowych. To ciekawe twierdzenie udowodnił Gniedenko w 1943.

[4]występują mało prawdopodobne zjawiska, ale z jednego końca skali, np bardzo prawdopodobne są stany zdrowia i lekkiego stanu choroby a mało prawdopodobne stany ciężkiej choroby

Jak widać, skala logitowa jest równa logarytmowi stosunku szans dla zdarzenia o prawdopodobieństwie p.

Mając wartość logitu, łatwo obliczyć prawdopodobieństwo ze wzoru

$$lgt^{-1}(u) = \frac{1}{1+\exp(-u)}$$

Przykład A.11 (Kennedy i Nixon) *W rywalizacji o fotel prezydenta USA w listopadzie 1960 wygrał Kennedy. Dane przedstawiają procent poparcia dla Kennedy'ego i Nixona w listopadzie 1960 i styczniu 1962 (w połowie kadencji) wśród katolików (elektorat Kennedy'ego) i protestantów (elektorat Nixona)*

% poparcia		*Kennedy*	*Nixon*
protestanci	*XI,60*	*38*	*62*
	I,62	*59*	*41*
katolicy	*XI,60*	*78*	*22*
	I,62	*89*	*11*

Czytając bezpośrednio procenty poparcia dla Kennedy'ego widzimy, że wśród protestantów poparcie wzrosło w połowie kadencji o 21 punktów procentowych, a wśród katolików o 11 punktów procentowych. Czyżby Kennedy zasłużył sobie wśród protestantów na większy wzrost poparcia? Pamiętając, jak trudno zdobyć choć jeden procent poparcia w grupie wysokiego poziomu poparcia wyraźmy poparcie dla Kennedy'ego w skali logitowej

logit poparcia		*Kennedy*
protestanci	*XI,60*	$\ln\frac{38}{62} = -.490$
	I,62	$\ln\frac{59}{41} = .364$
katolicy	*XI,60*	$\ln\frac{78}{22} = 1.266$
	I,62	$\ln\frac{89}{11} = 2.091$

Przyrost poparcia dla Kennedy'ego w skali logitowej wynosi wśród protestantów $.854$ *a wśród katolików* $.825$*. Wskazuje to na równomierny wzrost poparcia dla Kennedy'ego w obu grupach.*

Dodatek B

Metoda IPF

Metoda iteracyjnego oszacowania proporcjonalnego (metoda **I***terative* **P***roportion* **F***itting*) została opracowana przez Deminga i Stephana w 1940 [2]. Metoda ta jest przydatna w znajdowaniu estymatorów $n^{(r)}_{ijk}$ w hierarchicznych modelach warunkowych. Procedurę tą można opisać w kilku krokach

1. Iteracja zerowa $w^{(0)}_{ijk}$ estymatorów $n^{(r)}_{ijk}$ powinna być tak wybrana, aby odpowiadała modelowi podporządkowanemu modelowi, dla którego wyznaczamy estymatory $n^{(r)}_{ijk}$. Takim modelem jest model stały, dla którego $w^{(0)}_{ijk} = 1$

2. Mnożąc przez odpowiednie współczynniki skalujące sukcesywnie dopasuj $w^{(0)}_{ijk}$ tak, aby zachowane zostały liczebności brzegowe dla efektów, występujących w estymowanym modelu; w ten sposób otrzymamy kolejne przybliżenia $w^{(1)}_{ijk}, w^{(2)}_{ijk}, w^{(3)}_{ijk}, ...$

3. Proces kontynuuj tak długo, aż różnica między liczbnościami brzegowymi $w^{(s)}_{ijk}$ i liczbnościami brzegowymi $n^{(r)}_{ijk}$ dla efektów, występujących w modelu będzie mniejsza od zadanej wartości ε.

Współczynniki skalujące są obliczane w specyficzny sposób dla każdego efektu . Przypuśćmy, że jesteśmy w $s-1$ iteracji $w^{(s-1)}_{ijk}$ i chcemy dopasować nowe wartości $w^{(s)}_{ijk}$ tak, aby zachowane były liczebności, odpowiadające efektowi λ^{XY}_{ij} z modelu M_r. Wiadomo (twierdzenie **??**), że wtedy $n^{(r)}_{ij+} = n^{(r-1)}_{ij+}$. Współczynnikiem skalującym będzie wtedy

$$\alpha_{ij} = \frac{n^{(r-1)}_{ij+}}{w^{(s-1)}_{ij+}}$$

Nowe wartości $w^{(s)}_{ijk}$ otrzymujemy ze wzoru

$$w^{(s)}_{ijk} = \alpha_{ij} w^{(s-1)}_{ijk}$$

Zauważmy, że wtedy

$$w^{(s)}_{ij+} = \sum_{k=1}^{K} w^{(s)}_{ijk} = \sum_{k=1}^{K} \alpha_{ij} w^{(s-1)}_{ijk} = \alpha_{ij} w^{(s-1)}_{ij+} = n^{(r-1)}_{ij+}$$

Analogicznie możemy wyznaczyć współczynniki skalujące dla dowolnych efektów oraz wykonać kolejne kroki iteracyjne.

Anderson, Fienberg i Haberman pokazali, że $w^{(s)}_{ijk}$ są zbieżne do estymatorów największej wiarygodności $n^{(r)}_{ijk}$.

Przykład B.1 *Dopasujmy model* $[XY][YZ]$ *do danych* $n_{ijk}^{(r-1)}$:

$n_{ijk}^{(r-1)}$		z_1	z_2
x_1	y_1	*1*	*2*
	y_2	*3*	*4*
x_2	y_1	*5*	*6*
	y_2	*7*	*8*

$w_{ijk}^{(0)}$		z_1	z_2
x_1	y_1	*1*	*1*
	y_2	*1*	*1*
x_2	y_1	*1*	*1*
	y_2	*1*	*1*

Dopasujemy macierz dla efektu λ_{ij}^{XY}*, gdyż występuje on w naszym modelu* $[XY][YZ]$

$n_{ij+}^{(r-1)}$		
x_1	y_1	*3*
	y_2	*7*
x_2	y_1	*11*
	y_2	*15*

$w_{ij+}^{(0)}$		
x_1	y_1	*2*
	y_2	*2*
x_2	y_1	*2*
	y_2	*2*

α_{ij}		
x_1	y_1	$\frac{3}{2} = 1.5$
	y_2	$\frac{7}{2} = 3.5$
x_2	y_1	$\frac{11}{2} = 5.5$
	y_2	$\frac{15}{2} = 7.5$

Po uwzględnieniu współczynnika skalującego otrzymamy nową macierz:

$w_{ijk}^{(1)}$		z_1	z_2
x_1	y_1	$1 * 1.5$	$1 * 1.5$
	y_2	$1 * 3.5$	$1 * 3.5$
x_2	y_1	$1 * 5.5$	$1 * 5.5$
	y_2	$1 * 7.5$	$1 * 7.5$

$=$

$w_{ijk}^{(1)}$		z_1	z_2
x_1	y_1	1.5	1.5
	y_2	3.5	3.5
x_2	y_1	5.5	5.5
	y_2	7.5	7.5

Teraz wyliczymy kolejne przybliżenie odpowiadające efektowi λ_{jk}^{YZ} *dla modelu* $[XY][YZ]$.

$n_{+jk}^{(r-1)}$	z_1	z_2
y_1	6	8
y_2	10	12

$w_{+jk}^{(1)}$	z_1	z_2
y_1	7	7
y_2	11	11

α_{jk}	z_1	z_2
y_1	$\frac{6}{7} = .857$	$\frac{8}{7} = 1.143$
y_2	$\frac{10}{11} = .909$	$\frac{12}{11} = 1.091$

I kolejne przybliżenie estymatorów:

$w_{ijk}^{(2)}$		z_1	z_2
x_1	y_1	$1.5 * .857$	$1.5 * 1.143$
	y_2	$3.5 * .909$	$3.5 * 1.091$
x_2	y_1	$5.5 * .857$	$5.5 * 1.143$
	y_2	$7.5 * .909$	$7.5 * 1.091$

$=$

$w_{ijk}^{(2)}$		z_1	z_2
x_1	y_1	1.286	1.714
	y_2	3.182	3.815
x_2	y_1	4.714	6.286
	y_2	6.818	8.182

W ten sposób zakończyliśmy pierwszy cykl przybliżeń. Wartości brzegowe dla efektu λ_{ij}^{XY} *wynoszą*

$w^{(2)}_{ij+}$		
x_1	y_1	$1.286 + 1.714$
	y_2	$3.182 + 3.815$
x_2	y_1	$4.714 + 6.286$
	y_2	$6.818 + 8.182$

$=$

$w^{(2)}_{ij+}$		
x_1	y_1	3.0
	y_2	6.997
x_2	y_1	11.0
	y_2	15.0

która już jest idealnie zbliżona do $n^{(r-1)}_{ij+}$*, nie ma więc potrzeby wprowadzać poprawki na ten efekt. Trzeba jeszcze sprawdzić wartości brzegowe dla efektu* λ^{YZ}_{jk}

$w^{(2)}_{+jk}$		
y_1	z_1	$1.286 + 4.714$
	z_2	$1.714 + 6.286$
y_2	z_1	$3.182 + 6.818$
	z_2	$3.815 + 8.182$

$=$

$w^{(2)}_{+jk}$		
y_1	z_1	6.0
	z_2	8.0
y_2	z_1	10.0
	z_2	11.997

Tu też wartości brzegowe są bardzo bliskie $n^{(r-1)}_{+jk}$*, co oznacza, że znaleźliśmy estymatory największej wiarygodności dla* $n^{(r)}_{ijk}$*, równe* $w^{(2)}_{ijk}$*:*

$w^{(2)}_{ijk}$		z_1	z_2
x_1	y_1	1.286	1.714
	y_2	3.182	3.815
x_2	y_1	4.714	6.286
	y_2	6.818	8.182

Tutaj zbieżność uzyskaliśmy po dwóch iteracjach w jednym cyklu, obejmującym wszystkie efekty modelu[1]*. W przypadku ogólnym takich iteracji trzeba będzie wykonać więcej.*

[1]Nie jest to przypadek. Haberman w 1974 pokazał, że jeśli liczba nieznanych parametrów modelu nie przekracza 6, to metoda IPF jest zbieżna w jednym cyklu.

Dodatek C

Ćwiczenia

C.1 Zadania na ćwiczenia w laboratorium

Materiały na ćwiczenia:

http:/www.math.yorku.caŚCSĆoursesrcat

1. **Dopasowywanie rozkładów.**

1.1 **Wykres *poisonness***

Dane:

Dane von Bortkiewicza (1898). Liczba wypadków śmiertelnych w 10 korpusach armii pruskiej w ciągu 20 lat:

liczba wypadków	0	1	2	3	4
liczba obserwacji (korpusy x lata)	109	65	22	3	1

Listy Federalistów. Występowanie słowa *may* w 262 blokach po 200 słów.

liczba wystąpień	0	1	2	3	4	5	6
liczba bloków	156	63	29	8	4	1	1

itbpFU2.8732cm5.8496cm0cmpoischart1.gifitbpF3.7079cm5.8057cm0cmpoischart2.g

Metoda.

1.1.1 Pokaż, że gdy w n_k próbach wystąpiło k sukcesów i gdy rozkład liczby sukcesów jest rozkładem Poissona z parametrem λ to dla dużej liczby n obserwacji zachodzi w przybliżeniu równość

$$u_k \stackrel{df}{=} \ln\left(\frac{k!\, n_k}{n}\right) = -\lambda + (\ln\, \lambda)\, k$$

Wielkość u_k nazywamy *pseudolicznikiem* (ang. *count metameter*)

1.1.2. Napisz za pomocą najwygodniejszego dla ciebie narzędzia (np. *Excela*) procedurę, która rysuje wykres punktowy $\{(k, u_k) : k = 0, 1, ...\}$ oraz wpisuje w ten układ prostą regresji, oblicza jej równanie i drukuje wartość współczynnika determinacji R^2.

1.1.3. Oceń wizualnie, na podstawie sporządzonych wykresów czy można przyjąć, że *Dane von Bortkiewicza* pochodzą z rozkładu Poissona.

1.1.4. Zrób zadanie 1.1.3. Dla *Listów Federalistów*.

1.2. **W*ykresy Orda*.**

Metoda (Ord,1967) zapoznaj się z metodą w [3]

2. Sprawdź metodą Orda typ rozkładu dla poznanych przykładów. Napisz odpowiednią procedurę w znanym ci języku programowania.

3. **Własności ilorazu krzyżowego** θ

Dana jest tablica prawdopodobieństw 2×2

X	Y	
	y_1	y_2
x_1	p_{11}	p_{12}
x_2	p_{21}	p_{22}

i odpowiadający jej iloraz krzyżowy $\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$.

3.1 Pokaż, że prawdziwe są nierówności:

$$\begin{aligned}
\theta &> 1 \Longleftrightarrow P(Y = y_1 | X = x_1) > P(Y = y_1 | X = x_2), \\
\theta &> 1 \Longleftrightarrow P(X = x_1 | Y = y_1) > P(X = x_1 | Y = y_2), \\
\theta &< 1 \Longleftrightarrow P(Y = y_1 | X = x_1) < P(Y = y_1 | X = x_2), \\
\theta &< 1 \Longleftrightarrow P(X = x_1 | Y = y_1) < P(X = x_1 | Y = y_2)
\end{aligned}$$

3.2 Udowodnij, że dla każdego $\theta > 0$ i dla każdych $0 < p < 1$ i $0 < q < 1$ istnieje tablica prawdopodobieństw 2×2

X	Y	
	y_1	y_2
x_1	p_{11}	p_{12}
x_2	p_{21}	p_{22}

taka, że jej iloraz krzyżowy jest równy θ i taka, że $p_{1\cdot} \stackrel{df}{=} p_{11} + p_{12} = p$ oraz $p_{\cdot 2} \stackrel{df}{=} p_{12} + p_{22} = q$.

Wskazówka. Oznaczmy $p_{12} \stackrel{df}{=} x$. Pokaż, korzystając z własności Darboux, że równanie $f(x) = \theta$ ma zawsze rozwiązanie. Funkcja $f(x)$ jest zdefiniowana wzorem

$$f(x) = \frac{(p - x)(q - x)}{x(x + 1 - p - q)}$$

3.3 Spróbuj wyznaczyć taką tablicę dla $\theta = 1.5$, $p = 0.2$, $q = 0.6$

4. **Test** χ^2 **i test oparty na ilorazie krzyżowym** θ

4.1 Oblicz iloraz krzyżowy θ dla danych Pearsona o rozwoju umysłowym i fizycznym uczniów. Zilustruj na podstawie tych danych nierówności, opisane w zadaniu 3.1, zastępując odpowiednie prawdopodobieństwa przez ich częstości. Co te nierówności oznaczają?

4.2 Przedstaw tę tablicę w postaci standaryzowanej i narysuj odpowiadający jej wykres kołowy. Jak wygląda w tablica w postaci standaryzowanej i odpowiadający jej wykres kołowy dla przypadku niezależności i jednorodności?

4.3 Zastosuj test χ^2 i test oparty na ilorazie krzyżowym θ dla testowania hipotezy niezależności dla tych danych. Zapoznaj się z metodą obliczeń testu χ^2w programach *Excel* i *Statistica*

4.4 Znajdź 95% przedział ufności dla θ.

4.5 Dla lewego i prawego końca tego przedziału zbuduj tablice w postaci standaryzowanej i narysuj odpowiadające im wykresy kołowe. Porównaj wykresy, otrzymane w punktach 4.2 i 4.5. Jak z tych wykresów odczytać zależność (niezależność) wierszy i kolumn?

Dane: Rozwój umysłowy i fizyczny uczniów.

	Rozwój umysłowy	
Rozwój fizyczny	dobry	zły
dobry	581	561
zły	209	351

Źródło. Pearson, K., (1906) *On the relationship of inteligence to size and shape of head, and to other physical and mental characters*, Biometrica, 5, 105-146

4.4 Wykonaj to samo dla danych:

Dane: Liczba dobrze rozwiązanych zadań z matematyki

	Zadania	
Płeć	geometryczne	niegeometryczne
uczennice	21	29
uczniowie	22	32

Źródło. Wyniki matury próbnej z matematyki (poziom podstawowy) w III LO w Wałbrzychu w 2001 (informacja od nauczyciela)

5. **Test symetrii**

5.1 Próba z rozkładu wielomianowego o prawdopodobieństwie

$P(X = x_i, Y = y_j) = p_{ij}$, $(i, j = 1, 2, ..., I)$ umieszczona jest w tablicy $N = [n_{ij}]$ (n_{ij} jest liczbą obserwacji w próbie takich, że $X = x_i$ oraz takich, że $Y = y_j$).

Znajdź test χ^2 do testowania hipotezy

$$H_0: \quad p_{ij} = p_{ji}$$

dla wszystkich $i, j = 1, 2, ..., I$.

5.2 Użyj tego testu do testowania hipotezy H_0 w tablicy danych:
Dane: Porównanie wzrostu 205 par małżeńskich.

Mąż	Żona		
	wysoka	średnia	niska
wysoki	18	28	14
średni	20	51	28
niski	12	25	9

Co oznacza hipoteza H_0 dla wzrostu par małżeńskich?

Źródło. Wyniki zebrane przez Galtona, Christensen [59]

5.3 Zbadaj symetrię rozwoju umysłowego i fizycznego uczniów

6. **Eksperyment przedszkolny.** W 1962 roku przeprowadzono eksperyment, w którym wziął udział 123 dzieci z 3 i 4-letnich z ubogich rodzin w Ypsilanti w stanie Michigan. Część dzieci, wybranych losowo, uczęszczała przez dwa lata do przedszkola. Pozostałe dzieci do przedszkola nie uczęszczały.

C.2 Zadania egzaminacyjne

1. Na poniższym drzewku podane są wyniki obliczeń dla hierarchicznych model logliniowych trzech zmiennych X, Y, Z. Na krawędzi, łączącej dwa modele podane są wartości $G^2\left(M_r \,|M_{r-1}\right)$.

   Na przykład $G^2\left([XZ][YZ]\,|[XY][YZ][XZ]\right) = 8$. Początkowa wartość, nie zaznaczona na drzewku, oznaczająca $G^2\left(M_1\,|M_0\right) = G^2\left([XY][YZ][XZ]\,|[XY\right.$ wynosi 10. Liczba różnych wartości cechy X jest równa $I = 3$,cechy Y jest równa $J = 4$, cechy Z jest równa $K = 2$.

   dtbpFU13.1182cm5.4235cm0pttree342.wmfPodaj wzór na ostateczny model, wynikający z tych obliczeń.

2. Tablica zawiera prawdopodobieństwa $P\left(X = x_i, Y = y_j, Z = z_k\right)$. Wybierz, jaki typ zależności

   (a) $[XZ][YZ]$

(b) $[XY][Z]$

(c) $[X][Y][Z]$

(d) *żaden z nich*

występuje w danych. Dla ułatwienia, wystarczy sprawdzić czy warunek, określający typ zależności zachodzi dla p_{111}

dtbpF206.4375pt81.9375pt0ptFigure

3. Zmienna X ma dwie wartości: *w wysokie zarobki*, *n niskie zarobki*, zmienna Y wartości - *k kobieta*, *m mężczyzna*, Z: *s wykształcenie średnie*, *z wykształcenie wyższe*. Model logitowy, łączący te zmienne ma postać:

$$L = -1 - Y^{(m)} + 2\,Z^{(w)},$$

gdzie L jest logitem prawdopodobieństwa uzyskania wysokich zarobków, $Y^{(m)}$ jest równe 1 gdy Y ma wartość m, 0 gdy Y ma wartość k; $Z^{(w)}$ jest równe 1 gdy Z ma wartość w, 0 gdy Z ma wartość s.

   (a) Kto ma większe prawdopodobieństwo wysokich zarobków: kobieta z wykształceniem wyższym, czy mężczyzna ze średnim?

   (b) Ile to większe prawdopodobieństwo wynosi?

   (c) Oblicz iloraz krzyżowy dla par zmiennych (Y, X)

4. Napisz układ równań w modelu logitowym proporcjonalnych szans, w którym zmienna wynikowa P oznacza stosunek danej osoby do palenia: *nie pali, trochę pali*, *dużo pali*. Zmiennymi objaśniającymi są P płeć: *kobieta, mężczyzna*, R stosunek rodziców do palenia: *oboje palą, jedno z nich pali, żadne nie pali*. Jakie znaki będą miały współczynniki przy zaprojektowanych przez ciebie zmiennych objaśniających, jeśli dzieci obojga palących rodziców więcej palą niż dzieci rodziców, z których jedno pali, a ci palą więcej niż dzieci rodziców niepalących. Podobnie, jeśli mężczyźni palą więcej od kobiet?

5. Cechy X i Y są niezależne. Uzupełnij tabelę z liczebnościami

?	?	4
8	12	16
28	?	?

6. Wśród studentów ADJ uzyskano następujące wyniki

ocena	2	3	4	5
Kobiety	10	40	120	10
Mężczyźni	10	10	80	20

Czy na poziomie 0.05 można twierdzić, że wyniki z egzaminu i płeć są od siebie niezależne?

C.2.1 Egzamin poprawkowy

1. Rozpoznaj właściwy model zależności dla prawdopodobieństw:dtbpF206.4375p

   Wsk. Wybierz spośród modeli: [??][??], [??][?], [X][Y][Z]. Zamiast ? musisz wstawić odpowiednie litery X,Y,Z. Jeśli kilka modeli pasuje, wybierz jeden z nich.

2. Zbuduj metodą najmniejszych kwadratów model logitowy dla danych:

W	P	L
w	k	1
	m	0
n	k	-1
	m	-1

   gdzie L jest logitem prawdopodobieństwa dobrego samopoczucia, W wzrostem (w - wysoki, n- niski), P płcią badanego.

   Wsk. Metoda najmniejszych kwadratów dla danych (x_i, y_i) $i = 1, 2, ...n$ w modelu

$$y = f(x, \alpha, \beta, ...)$$

   gdzie $\alpha, \beta, ...$ są nieznanymi parametrami modelu, polega na ich wyznaczeniu takim, że

$$\sum_{i=1}^{n} (f(x_i, \alpha, \beta, ...) - y_i)^2$$

   osiąga minimum względem $\alpha, \beta, ...$

3. Po wykonaniu zad.2 wyznacz iloraz krzyżowy dla tablicy

	zadowoleni	niezadowoleni
kobiety		
mężczyźni		

dla każdego ustalonego poziomu wzrostu. Która para dominuje

(a) zadowolone kobiety i niezadowoleni mężczyźni, czy

(b) niezadowolone kobiety i zadowoleni mężczyźni

4. Ala, Basia i Celina rzucały po 100 razy, każda swoją monetą. Ala uzyskała 40 orłów, Basia i Celina po 30 orłów. Czy na poziomie 0.05 można twierdzić, że Ala i Basia rzucały taką samą monetą a prawdopodobieństwo wyrzucenia orła przez Celinę było dwa razy mniejsze od prawdopodobieństwa wyrzucenia orła przez Alę?

5. Na poniższym drzewku podane są wyniki obliczeń dla hierarchicznych model logliniowych trzech zmiennych X, Y, Z. Na krawędzi, łączącej dwa modele podane są wartości $G^2\left(M_r \,|M_{r-1}\right)$.

   Na przykład $G^2\left([XZ][YZ]\,|[XY][YZ][XZ]\right) = 8$. Początkowa wartość, nie zaznaczona na drzewku, oznaczająca $G^2\left(M_1\,|M_0\right) = G^2\left([XY][YZ][XZ]\,|[XY\right.$ wynosi 10. Liczba różnych wartości cechy X jest równa $I = 4$,cechy Y jest równa $J = 4$, cechy Z jest równa $K = 2$.

   dtbpF13.1182cm5.4279cm0ptFigure Znajdź wszystkie modele, zaakceptowane na poziomie 0.05.

Indeks

χ^2, 15

dane, 8
  ilościowe, 9
  jakościowe, 9

G^2, 15

hipoteza
  jednorodności, 18
  niezależności, 21

iloraz krzyżowy, 24
  reprezentacja standardowa, 25

kryterium
  Akaike, 57
  bayesowskie, 57

metoda
  IPF, 64
model
  hierarchiczny, 47
  logarytmiczno-liniowy, 40
  nasycony, 41
  proporcjonalnych szans, 36
  stały, 41

niezależność
  warunkowa, 43

odchylenie G^2, 15
odległość
  χ^2 Pearsona, 15

paradoks Simpsona, 40

regresja
  logitowa, 32
    ze zmiennymi nominalnymi, 34
    ze zmiennymi porządkowymi, 36
  probitowa, 33
rozkład
  dwumianowy, 13
  wielomianowy, 14
    produktowy, 14
rozkład
  Poissona, 13

skala
  ilorazowa, 9
  kwantylowa, 60
  logitowa, 61
  nominalna, 8
  podwójnie logarytmiczna, 61
  porządkowa, 8
  prawdopodobieństw, 60
  probitowa, 61
  przedziałowa, 8
stopnie swobody
  dla modeli prostych, 44
stosunek szans, 23

tablica
  kontyngencji, 12

zapis bilansowy, 41
zmienna
    grupująca, 18
    wynikowa, 18
zmienne
    indykatorowe, 34

Literatura

[1] Agresti, A., (1990), *Categorical Data Analysis*, New York: Wiley

[2] Deming, W.E., Stephan F.F., (1940), On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11:** 427-444

[3] Friendly, M., *Categorical Data Analysis with Graphics,* http:/www.math.yorku.caŚCSĆoursesrcat

[4] McPherson, G.,(1990), *Statistics in Scientific Investigation*, New York: Springer