

Analýza symetrických tabulek

BLANKA ŘEHÁKOVÁ

Sociálně ekonomický ústav ČSAV,
Ústí nad Labem, pracoviště Praha

Tento článek navazuje na dvě dřívější práce [Řehák, Řeháková 1984; 1986: kap. 17 a 18], které se zabývaly porovnáním marginálních rozdělání a strukturou četností ve čtvercových tabulkách párově závislých dat (stručněji řečeno v symetrických tabulkách) a kde byly též obsáhlé popsány situace, za nichž tyto tabulky vznikají a uvedeny příklady ze sociologických výzkumů.

Cílem této práce je ukázat další úlohy pro taková data a možnosti jejich analýzy a současně poskytnout návod, jak získat výsledky pomocí procedury LOGLINEAR v systému SPSSX (u většiny uvedených úloh se bez počítače neobejdeme a SPSSX je známý systém pro analýzu dat ve společenských vědách).

Analytické úlohy v symetrických tabulkách lze rozdělit do dvou typů:

- a) porovnání marginálních rozdělání,
- b) studium struktury četností vzhledem k
 - statistické závislosti,
 - symetrii četností podle hlavní diagonály a různým modifikacím pojmu symetrie,
 - vnějšímu kritériu (funkci) vyjadřujícímu hypotézu o vztahu mezi četnostmi.

Zde bude uvedeno řešení vybraných úloh typu b) tak, jak je poskytuje metodologie log-lineárních modelů. Bude demonstrováno na datech o vzdělání manželských dvojic z výzkumu města Vamberka v r. 1988.¹

Metodika log-lineárních modelů se zatím v československé sociologii plně neprosadila. Jejimi průkopníky u nás byli T. Havránek [1980] a P. Matějů [1985]. Jednoduchým úvodem do problematiky je kniha G. Upton [1978], přehled metod a možností poskytují dvě knihy S. J. Habermana [1978, 1979], matematická teorie je například v monografii autorů Bishopové, Fienberga a Hollanda [1977], nebo Habermana [1974].

V dalším textu bude vždy $n(i, j)$ značit pozorovanou četnost v poli (i, j) dvourozměrné tabulky, $m(i, j)$ očekávanou četnost za určitého modelu, $\hat{m}(i, j)$ odhad očekávané četnosti, $n(i+) = \sum_j n(i, j)$, $n(+j) = \sum_i n(i, j)$, $n = \sum_i \sum_j n(i, j)$.

Poznamenejme na začátek, že postup hledání modelu, který vysvětluje vztahy mezi údaji tabulky, spočívá v následných krocích:

- a) formulace modelu: $\log m(i, j) =$ součet určených faktorů,
- b) odhad parametrů odpovídajících jednotlivým faktorům,
- c) odhad očekávaných četností za platnosti modelu,
- d) test dobré shody mezi očekávanými a pozorovanými četnostmi,
- e) rozbor struktury reziduálních odchylek² v jednotlivých polích, byl-li model zamítnut, a formulace vhodnějšího modelu s návratem k a).

Redakce upozorňuje, že matematické vzorce jsou místo kurzívou vysazeny písmem stojatým

(1) Vzdělání bylo zjišťováno původně na sedmibodové stupnici. Sloučení do čtyř kategorií bylo provedeno pro účely tohoto článku autorkou.

(2) Reziduální odchylkou v poli (i, j) je zde míněno adjustované (upravené) roziduum $r(i, j) =$

f) interpretace modelu, byl-li přijat; zde je nutná opatrnost, protože statistická shoda může být nalezena u několika i dost odlišných modelů.

Jedním z problémů, který vzniká při logaritmování četností, je přítomnost nulových četností v tabulce. Připomeňme si, že ty mohou být buď strukturální, nebo výběrové. Nuly strukturální plynou z povahy dat: určitá pole nemohou být obsazena a model taková pole musí vyloučit. Nuly výběrové jsou důsledkem variability výběru a malé pravděpodobnosti obsazení daného pole: v praxi se výběrové nuly odstraňují například přičtením velmi malého kladného čísla do polí s výběrovými nulami, nebo přičtením čísla 0.5 (či jiného malého kladného čísla) ke všem četnostem. Jinou možností je metoda pseudobayesovského odhadu [Bishop, Fienberg, Holland: 401–433, spec. 432] (ta byla použita na data tabulky 1, k odstranění výběrové nuly v poli (1, 4)). V dalším budeme pracovat s tabulkou 2, kde už jsou všechny četnosti kladné.

Tabulka 1. Vzdělání manželských dvojic

A	B	1	2	3	4
1		39	21	4	0
2		85	163	47	5
3		8	32	46	2
4		1	8	11	9

$A(B)$ = nejvyšší dosažené vzdělání manžela (manželky),

1 = základní (i neúplné), 2 = vyučení nebo střední škola bez maturity nebo střední odborné učiliště s maturitou, 3 = střední odborná nebo všeobecná škola s maturitou, 4 = vysoká škola.

Tabulka 2. Vzdělání manželských dvojic – vyhlazená data

A	B	1	2	3	4
1		37,72	21,53	4,62	0,13
2		84,88	161,59	48,23	5,30
3		8,98	32,54	44,42	2,06
4		1,42	8,33	10,73	8,52

Pozn.: Vyhlazení dat bylo provedeno pseudobayesovskou metodou.

$(n(i, j) - \hat{m}(i, j)) / \hat{c}(i, j)^{1/2}$, kde $\hat{c}(i, j)$ je odhad rozptylu náhodné veličiny $n(i, j) - \hat{m}(i, j)$. Za platnosti modelu má $r(i, j)$ asymptoticky standardní normální rozdělení. Asymptotika je pravděpodobně v pořádku, je-li každé $m(i, j)$ větší než 25. Není-li tato podmínka splněna, pak reziduální odchylky poskytují jen hrubé indikace velikosti chyb, ale i tak dávají informaci o struktuře odchylek od modelu. Některé programy tisknou také (nebo jen) standardizovaná rezidua $(n(i, j) - \hat{m}(i, j)) / \hat{m}(i, j)^{1/2}$, která mají za platnosti modelu asymptoticky normální rozdělení s nulovým průměrem a rozptylem v (i, j) , který je obecně různý od jedničky. Proto, pokud je to možné, měli bychom jako pravidlo používat adjustovaná rezidua.

Model úplné nezávislosti

Hypotéza nezávislosti, obdobně jako u běžné asociační analýzy v libovolných kontingenčních tabulkách, je dána některou (z ekvivalentních) formulací pro všechna pole tabulky:

$$\begin{aligned}n(i, j) &= n(i+) n(+j) / n, \\m(i, j) &= a(i) b(j).\end{aligned}\quad (1)$$

Druhá formulace říká, že očekávaná četnost v poli (i, j) vzniká jako součin dvou nezávislých faktorů — řádkového $a(i)$ a sloupcového $b(j)$. Hypotézu nezávislosti můžeme testovat známým Pearsonovým testem chí-kvadrát a odchylky od ní například znaménkovým schématem (viz [Řehák, Řeháková 1978]).

Log-lineární model nezávislosti poskytuje paralelní test; je formulován rovnicemi

$$\begin{aligned}\log m(i, j) &= \lambda + \lambda^A(i) + \lambda^B(j), \\ \sum_i \lambda^A(i) &= \sum_j \lambda^B(j) = 0,\end{aligned}\quad (2)$$

kde λ je celkový poměr, $\lambda^A(i)$ je efekt i -tého řádku (i -té kategorie proměnné A), $\lambda^B(j)$ je efekt j -tého sloupce (j -té kategorie proměnné B). Kategorie $A(i)$ a $B(j)$ působí nezávisle jedna na druhé — vliv $A(i)$ se nemění ve sloupcích, vliv $B(j)$ se nemění v řádcích.

Statistiky testu dobré shody dat s určitým modelem, které se nejčastěji v analýze dat používají jsou dvě:

a) věrohodnostní statistika G^2

$$G^2 = 2 \sum_i \sum_j n(i, j) \log(n(i, j) / \hat{m}(i, j)), \quad (3)$$

b) Pearsonova statistika X^2

$$X^2 = \sum_i \sum_j (n(i, j) - \hat{m}(i, j))^2 / \hat{m}(i, j), \quad (4)$$

kde $\hat{m}(i, j)$ jsou odhady očekávaných četností za daného modelu.

V modelu nezávislosti je

$$\hat{m}(i, j) = n(i+) n(+j) / n \quad (5)$$

a G^2 a X^2 mají za hypotézy nezávislosti asymptoticky rozdělení chí-kvadrát s $(I-1)^2$ stupni volnosti, když I je počet řádků (sloupců) tabulky.

Pro data z tabulky 2 je $G^2 = 116.35$, $X^2 = 156.29$, což při devíti stupních volnosti dává pro obě statistiky dosaženou hladinu významnosti $P = 0.000 (< 0.050)$. Model nezávislosti, podle kterého by byla vzájemná volba partnerů v manželství zcela nezávislá na jejich vzdělání, neplatí.

Model kvazinezávislosti

V symetrických tabulkách bývá tendence k silnějšímu obsazování diagonálních polí, tj. kombinací vzájemně si odpovídajících kategorií řádkové a sloupcové proměnné, a proto model nezávislosti platí jen vzácně. Může se stát, že po vynechání (blokování) těchto polí³, platí pro zbytek tabulky model nezávislosti.

Model kvazinezávislosti předpokládá, že

$$\begin{aligned}m(i, j) &= a(i) b(j), \text{ jestliže } i \neq j, \\ &= 0, \text{ jestliže } i = j.\end{aligned}\quad (6)$$

(3) Blokována pole nemusí být jen pole hlavní diagonály, ale též např. pole s velkou reziduální odchylkou, či pole se strukturální nulou.

Vyjádřeno jako log-lineární model:

$$\log m(i, j) = \lambda + \lambda^A(i) + \lambda^B(j), i \neq j, \quad (7)$$

$$\sum_i \lambda^A(i) = \sum_j \lambda^B(j) = 0.$$

V tomto případě nelze vyjádřit odhady očekávaných četností algebraickým výrazem, počítají se iteračním postupem. Statistiky G^2 a X^2 mají za modelu kvazinezávislosti počet stupňů volnosti $df = I(I-3) + 1$.

Data tabulky 2 neodpovídají ani tomuto modelu, neboť $G^2 = 19.10$, $X^2 = 23.62$; to dává při pěti stupních volnosti dosažené hladiny významnosti $P = 0.002$ pro G^2 a $P = 0.000$ pro X^2 . Oproti modelu nezávislosti však došlo ke zlepšení, neboť $116.35/9 = 12.9$ je mnohem větší, než $19.10/5 = 3.8$, které je stále ještě daleko od jedničky.⁴

Hypotéza kvazinezávislosti říká, že odhlédneme-li od manželství, v nichž mají partneři stejné vzdělání, neprojevuje se žádná závislost mezi volbou a vzděláním v ostatních manželstvích. Jak vidíme, ani tato hypotéza neplatí. Silný pokles v hodnotě G^2/df mezi testem nezávislosti a kvazinezávislosti ukazuje na existenci vlivného faktoru vedoucího ke shodě vzdělání v manželství. Ostatní faktory projevující se při vzniku vzdělanostní nerovnosti manželských párů a vytvářející četnostní strukturu, budou zkoumány pomocí dalších modelů a hypotéz.

Model symetrie

Symetrii v dvourozměrné symetrické tabulce rozumíme situaci, kdy očekávané četnosti polí rozložených symetricky vzhledem k diagonále jsou si rovny, tj.

$$m(i, j) = m(j, i) \text{ pro všechna } i \text{ a } j. \quad (8)$$

Tomu odpovídá log-lineární model

$$\log m(i, j) = \lambda + \lambda^A(i) + \lambda^B(j) + \lambda^{AB}(i, j),$$

$$\lambda^A(i) = \lambda^B(i) \text{ pro všechna } i,$$

$$\lambda^{AB}(i, j) = \lambda^{AB}(j, i) \text{ pro } i \leq j, \quad (9)$$

$$\sum_i \lambda^A(i) = \sum_j \lambda^B(j) = \sum_i \lambda^{AT}(i, j) = \sum_j \lambda^{AB}(i, j) = 0.$$

Parametr $\lambda^{AB}(i, j)$ se interpretuje jako interakční efekt mezi i -tou kategorií proměnné A a j -tou kategorií proměnné B. Odhady očekávaných četností jsou

$$\hat{m}(i, j) = \begin{cases} (n(i, j) + n(j, i))/2, & i \neq j, \\ n(i, i), & i = j. \end{cases} \quad (10)$$

statistiky G^2 a X^2 mají asymptoticky rozdělení chí-kvadrát s $I(I-1)/2$ stupni volnosti.

Model symetrie tedy připouští specifické interakce mezi kategoriemi proměnných. Na jeho platnost nemá vliv diagonála; marginální populační rozdělení jsou totožná.

Kdyby platil model (9) pro data tabulky 2, znamenalo by to, že interakce mezi kategoriemi vzdělání v manželských dvojicích existují, že jsou nezávislé na jejich nositelích (muž, žena) a že efekty jednotlivých kategorií vzdělání manžela jsou stejné jako odpovídající kategorie vzdělání manželky, Z tabulky 2 je však na první pohled zřejmé, že v ní model symetrie platit nebude (velmi rozdílné četnosti $n(i, j)$ a $n(j, i)$)

(4) Jedním z kritérií vhodnosti modelu je poměr G^2/df , resp. X^2/df (df = počet stupňů volnosti statistiky); vychází z toho, že očekávaná hodnota G^2 i X^2 je rovna df . Čím větší je statistika než df , tím horší je model.

a marginální četnosti $n(i+)$, $n(+i)$). Skutečně, $G^2 = 53.2$, $X^2 = 49.8$, $df = 6$, $P = 0.000$ pro obě statistiky. Model je horší, než model kvazinezávislosti ($53.2/6 = 8.9 > 3.8$), ale lepší než model nezávislosti ($8.9 < 12.9$). Rozdíl proti nezávislosti může být způsoben tím, že na platnost modelu (9) nemá vliv diagonála, jejíž obsazení je na první pohled velmi výrazné. Rozdíl proti kvazinezávislosti může být způsoben jak zřejmou rozdílností marginálních rozdělení, tak ovšem i možnou neplatností interakční symetrie.

Model kvazisymetrie

Model symetrie je velmi přísný a ve společenskovědní praxi bude platit jen zřídka. Méně omezující je model kvazisymetrie, který předpokládá, že

$$\begin{aligned} m(i,j) &= a(i) b(j) d(i,j), \\ d(i,j) &= d(j,i), \quad i \neq j, \end{aligned} \quad (11)$$

nebo alternativně

$$m(i,j) m(j,k) m(k,i) / m(j,i) m(k,j) m(i,k) = 1.$$

Log lineární vyjádření je

$$\begin{aligned} \log m(i,j) &= \lambda + \lambda^A(i) + \lambda^B(j) + \lambda^{AB}(i,j), \\ \lambda^{AB}(i,j) &= \lambda^{AB}(j,i), \quad i \leq j, \\ \sum_i \lambda^A(i) &= \sum_j \lambda^B(j) = \sum_i \lambda^{AB}(i,j) = \sum_j \lambda^{AB}(i,j) = 0. \end{aligned} \quad (12)$$

Model kvazisymetrie se liší od modelu symetrie tím, že už nepředpokládá rovnost řádkových a sloupcových efektů, tj. shodu marginálních rozložení, ale pouze symetrii interakčních efektů. Odhady očekávaných četností mimo diagonálu se nedají vyjádřit algebraickým výrazem, získávají se pomocí **iteračních** algoritmů. Pro odhady na diagonále platí $\hat{m}(i,i) = n(i,i)$ jako u modelu symetrie.⁵ Statistiky G^2 a X^2 mají asymptoticky rozdělení chí-kvadrát s $(I-1)(I-2)/2$ stupni volnosti. Závažnost modelu spočívá v tom, že odhaluje a měří skryté symetrické interakce mezi kategoriemi řádkové a sloupcové proměnné po extrakci a odečtení vlivu rozdílných marginálních efektů.

Pro tabulku 2 a model kvazisymetrie je $G^2 = 0.90$, $X^2 = 0.88$, což při třech stupních volnosti dává dosaženou hladinu významnosti 0.825 pro G^2 a 0.829 pro X^2 , a proto tento model odpovídá datům velmi dobře; to dokumentuje i tabulka 3. Všimněme si, že adjustovaná rezidua $r(i,j)$ a $r(j,i)$ se liší jen znaménkem, tak je tomu v modelu kvazisymetrie vždy.

První závěr tedy je, že existuje symetrická interakce mezi kategoriemi vzdělání u manželských dvojic přes nerovnost vzdělanostní struktury mužů a žen. Další interpretace vychází z odhadů hodnot řádkových, sloupcových a zvláště interakčních efektů.

Tyto odhady, odhady jejich rozptylů a další odvozené charakteristiky (z-skóry, intervaly spolehlivosti) jsou tím přesnější, čím větší je počet pozorování, a to nejen

(5) Modely symetrie a kvazisymetrie byly formulovány pro úplné tabulky, ale protože jsou v obou případech odhady očekávaných četností diagonály rovny příslušným pozorovaným četnostem, působí odchylky od modelů jen pole mimo diagonálu a proto se odhady očekávaných četností v nediagonálních polích, reziduální odchylky, hodnoty G^2 , X^2 , ani počty stupňů volnosti nemění, ať už pracujeme s úplnou tabulkou nebo blokujeme diagonálu, či má tabulka na diagonále strukturní nuly. Mění se však odhady parametrů λ^{AB} , neboť když vyloučíme diagonálu, vyloučíme parametry $\lambda^{AB}(i,i)$, $1 \leq i \leq I$. Poznamenejme ještě, že pokud vyjme diagonálu, je pro $I = 3$ model kvazisymetrie ekvivalentní s modelem kvazinezávislosti a pro $I > 3$ plyne kvazisymetrie z kvazinezávislosti.

Tabulka 3. Pozorované četnosti, odhady očekávaných četností a adjustovaná rezidua při modelu kvazisy-metrie pro tab. 2.

A	B	1	2	3	4
1		37,72 37,72 —	21,53 22,09 — 0,37	4,62 4,02 0,41	0,13 0,17 — 0,11
2		84,88 84,32 0,37	161,59 161,59 —	48,23 49,70 — 0,83	5,30 4,39 0,81
3		8,98 9,58 — 0,41	32,54 31,07 0,83	44,42 44,42 —	2,06 2,93 — 0,78
4		1,42 1,38 0,11	8,33 9,24 — 0,81	10,73 9,86 0,78	8,52 8,52 —

$$G^2 = 0,90, P = 0,825, X^2 = 0,88, P = 0,829, df = 3$$

celkově, ale i v každém poli tabulky. Je dosud málo známo o tom, pro jak velké čet-nosti v polích je už dosaženo dostatečné přesnosti. V datech je málo vysokoškolačků jak mezi muži, tak mezi ženami a některá pole jsou tak mizivě obsazena, že k ziska-ným výsledkům je třeba přistupovat jen velmi opatrně.

Každý z-skór slouží k testu hypotézy, že k němu příslušné lambda je nulové. Pokud bereme každý efekt zvlášť, pak ta lambda, která jsou v absolutní hodnotě větší než 1.96 lze považovat za nenulová (hladina významnosti $\alpha = 0.05$). Pokud budeme interpretovat více efektů najednou a zajímá nás významnost celé struktury najednou, je správné použít simultánní test, např. Holmův sekvenční postup.⁶

Hlavní efekty zde nejsou zvlášť zajímavé, lze říci, že v podstatě odráží marginální strukturu tabulky. Porovnáme-li mezi sebou $\lambda^A(i)$ a $\lambda^B(i)$ ($i=1, \dots, 4$) vidíme, proč nemohl platit model symetrie. Příslušné z-skóry jsou⁷

(6) Holmův sekvenční postup je popsán v publikaci Řehák, Řeháková [1986 : 324], v níž lze nalézt též tabulku kritických hodnot dvoustranného normálního testu pro simultánní testování hypotéz (Tabulka A).

(7) Výpočet jmenovatelů se provádí podle vzorce:
 $\text{var}(\lambda^A(i) - \lambda^B(i)) = \text{var} \lambda^A(i) + \text{var} \lambda^B(i) - 2 \text{cov}(\lambda^A(i), \lambda^B(i))$. Protože však některé programy (např. LOGLINEAR v SPSSX) netisknou $\text{var} \lambda^A(I)$, $\text{var} \lambda^B(I)$, $\text{cov}(\lambda^A(I), \lambda^B(I))$, je výpočet

$$\begin{aligned} \text{var}(\lambda^A(I) - \lambda^B(I)) \text{ poněkud nepříjemný: } & \text{var}(\lambda^A(I) - \lambda^B(I)) = \sum_{i=1}^{I-1} \text{var}(\lambda^A(i)) + \\ & + 2 \sum_{i=1}^{I-2} \sum_{j=1}^{I-1} \text{cov}(\lambda^A(i), \lambda^A(j)) + \\ & + \sum_{i=1}^{I-1} \text{var}(\lambda^B(i)) + 2 \sum_{i=1}^{I-2} \sum_{j=1}^{I-1} \text{cov}(\lambda^B(i), \lambda^B(j)) - \\ & - 2 \sum_{i=1}^{I-1} \sum_{j=1}^{I-1} \text{cov}(\lambda^A(i), \lambda^B(j)). \end{aligned}$$

Tabulka 4a. Odhady parametrů modelu kvazisymetrie pro tab. 2.

Parametr	Odhad	Std. chyba	Z-skór	95 % interval spolehlivosti	
A	$\lambda(1)$	- 0,836	0,175	- 4,777	- 1,179
	$\lambda(2)$	1,301	0,135	9,637	1,036
	$\lambda(3)$	0,216	0,153	1,412	- 0,084
	$\lambda(4)$	- 0,681	0,181	- 3,762	- 1,036
B	$\lambda(1)$	0,236	0,149	1,584	- 0,056
	$\lambda(2)$	1,035	0,154	6,721	0,733
	$\lambda(3)$	0,420	0,154	2,727	0,118
	$\lambda(4)$	- 1,691	0,248	- 6,818	- 2,177
AB	$\lambda(1,1)$	1,805	0,348	5,187	1,123
	$\lambda(2,1)$	0,472	0,142	3,324	0,194
	$\lambda(2,2)$	0,324	0,169	1,917	- 0,007
	$\lambda(3,1)$	- 0,618	0,213	- 2,901	- 1,035
	$\lambda(3,2)$	- 0,240	0,143	- 1,678	- 0,520
	$\lambda(3,3)$	0,733	0,206	3,558	0,329
	$\lambda(4,1)$	- 1,659	0,519	- 3,196	- 2,676
	$\lambda(4,2)$	- 0,556	0,216	- 2,574	- 0,979
	$\lambda(4,3)$	0,125	0,226	0,553	- 0,318
	$\lambda(4,4)$	2,090	0,412	5,073	1,282

Pozn. : A = Kategorie vzdělání muže

B = kategorie vzdělání ženy

AB = kombinace kategorií vzdělání pro muže i ženy

$$(\lambda^A(1) - \lambda^B(1))/\sqrt{\text{var}(\lambda^A(1) - \lambda^B(1))} = -1.072/0.203 = -5.277,$$

$$(\lambda^A(2) - \lambda^B(2))/\sqrt{\text{var}(\lambda^A(2) - \lambda^B(2))} = 0.266/0.147 = 1.810,$$

$$(\lambda^A(3) - \lambda^B(3))/\sqrt{\text{var}(\lambda^A(3) - \lambda^B(3))} = -0.204/0.176 = -1.159,$$

$$(\lambda^A(4) - \lambda^B(4))/\sqrt{\text{var}(\lambda^A(4) - \lambda^B(4))} = 1.010/0.285 = 3.540,$$

takže rozdíly $\lambda^A(1) - \lambda^B(1)$ a $\lambda^A(4) - \lambda^B(4)$ jsou významně odlišné od nuly ($\alpha = 0.05$). a to je v rozporu s předpoklady modelu symetrie (v souboru je statisticky významná převaha žen nad muži u základního vzdělání a naopak převaha mužů nad ženami u vysokoškolského vzdělání).

Všechny variance a kovariance na pravé straně už nalezneme ve výstupech z výše uvedeného programu.

Interakční efekty vyjadřují intenzitu vzájemné volby od vysoké přitažlivosti (velké kladné číslo) až k vysoké nepřitažlivosti (záporné číslo s velkou absolutní hodnotou). Simultánním testem ($\alpha = 0.05$) zjistíme, že z interakcí jsou nenulové $\lambda(1,1)$, $\lambda(2,1)$, $\lambda(3,1)$, $\lambda(3,3)$, $\lambda(4,1)$, $\lambda(4,2)$, a $\lambda(4,4)$. Kladné znaménko je u $\lambda(1,1)$, $\lambda(2,1)$, $\lambda(3,3)$ a $\lambda(4,4)$, takže kombinace vzdělání 1–1, 2–1, 3–3, 4–4 se přitahují. Záporné znaménko u $\lambda(3,1)$, $\lambda(4,1)$, $\lambda(4,2)$ říká, že kombinace 3–1, 4–1, 4–2 se ne-

Tabulka 4b. Hodnoty koeficientů $\lambda(i, j)$ v modelu kvazisymetrie pro tabulku 2

<i>i</i>	<i>j</i>	1	2	3	4
1		1,8	0,5	- 0,6	- 1,7
2		0,5	0,3	- 0,2	- 0,6
3		- 0,6	- 0,2	0,7	0,1
4		- 1,7	- 0,6	0,1	2,1

přitahují. U kombinací 2–2, 3–2, 4–3 se nenulový interakční efekt statisticky významně neprojevil. Hodnoty parametrů je nutno posuzovat relativně vzhledem k celé struktuře, neboť jsou vázány vztahy (12). Interpretace koeficientů proto musí vždy vycházet z kontrastů jejich hodnot.

Výsledky tabulky 4a tedy říkají, že při zachování symetrie volby se párují především obě nižší kategorie. Dále se tu upřesňuje informace o významnosti volby shodných kategorií: neplatí u všech stupňů vzdělání, neboť u kombinace 2–2 odpovídá počet manželství očekávané hodnotě za předpokladu nezávislého (náhodného) vzdělanostního párování. Snížení intenzity vzájemné volby v párech se týká „vzdálených“ kategorií. To je dobře vidět z tabulky 4b, z jejíž struktury je dále možno vyčíst klesající uspořádání hodnot směrem od diagonály a na jeho základě hypotézu dále specifikovat – viz další, tzv. distanční model.

Distanční modely

Parametry $\lambda^{AB}(i, j)$ v tab. 4b mají klesavý charakter ve směru od diagonály, to znamená, že čím dále jsou od sebe kategorie vzdělání, tím méně takových kombinací se vyskytuje u partnerů v manželství. Je tedy možné pokusit se o upřesnění modelu a k podmínce kvazisymetrie ještě přidat klesavý charakter interakčních faktorů ve směru od diagonály. To lze provést distančními modely, které jsou ovšem použitelné jen v situacích, kdy proměnné jsou alespoň ordinální, aby měl smysl pojem „vzdálenost kategorií proměnné“. Jednoduchý distanční model

$$\log m(i, j) = \lambda + \lambda^A(i) + \lambda^B(j) + \eta |i - j|, \quad (13)$$

$$\sum \lambda^A(i) = \sum \lambda^B(j) = 0$$

dává interakci mezi kategoriemi *i* a *j* do souvislosti s jejich vzdáleností měřenou jako $|i - j|$, přičemž váha této vzdálenosti (parametr η) je konstantní, nezávislá na *i* a *j*. Pro $\eta = 0$ přechází model do modelu nezávislosti, je-li studována celá tabulka, nebo do modelu kvazinezávislosti, je-li vyloučena diagonála.

Statistiky dobré shody dat z tabulky 2 s distančním modelem (13) jsou $G^2 = 8.81$, $X^2 = 8.14$, což při osmi stupních volnosti (obecně $I(I-2)$) dává dosažené hladiny významnosti $P = 0.358$ a $P = 0.420$, takže model nelze zamítnout. Dlužno říci, že

Tabulka 5. Pozorované četnosti, odhady očekávaných četností a adjustovaná rezidua při distančním modelu pro data z tab. 2.

A	B	1	2	3	4
1		37,72	21,53	4,62	0,13
		40,97	17,51	5,01	0,51
		- 1,19	1,34	- 0,20	- 0,55
2		84,88	161,59	48,23	5,30
		75,93	170,37	48,74	4,96
		2,10	- 2,38	- 0,14	0,20
3		8,98	32,54	44,42	2,06
		12,65	28,38	42,63	4,34
		- 1,27	1,14	0,57	- 1,33
4		1,42	8,33	10,73	8,52
		3,45	7,73	11,62	6,20
		- 1,22	0,27	- 0,36	1,41

$$G^2 = 8,81, P = 0,358, X^2 = 8,14, P = 0,420, df = 8$$

vysvětluje data hůře, než model kvazisymetrie (12), což je vidět i z porovnání adjustovaných reziduí z tabulky 3 a 5.

Největší odklon od modelu je působen druhou kategorií, a to v polích (2,1) a (2,2), která se kompenzují. Přitažlivost kombinace vzdělání 2-2 je výrazně menší než u 1-1, 3-3, 4-4 a kombinace (muž, 2) — (žena, 1) je zastoupena více, než model čeká. Parametr η je záporný a vysoce významný. V absolutní hodnotě je výraz $\eta |i-j|$ tím větší, čím jsou od sebe kategorie vzdělání partnerů vzdálenější.

V modelu kvazisymetrie působí odchylky jen pole mimo diagonálu, a to není pravda v distančním modelu. Zkusme ho tedy ještě jednou, ale na tabulku s vynechanou diagonálou. Statistiky dobré shody jsou teď $G^2 = 1.64$, $X^2 = 1.56$, což při

Tabulka 6. Odhady parametrů distančního modelu pro tab. 2.

Parametr	Odhad	Std. chyba	95 % Interv. spol.		z-skór	
A	$\lambda(1)$	- 0,485	0,122	- 0,724	- 0,247	- 3,988
	$\lambda(2)$	0,961	0,081	0,802	1,119	11,874
	$\lambda(3)$	- 0,002	0,102	- 0,202	0,197	- 0,022
	$\lambda(4)$	- 0,474	0,160	- 0,788	- 0,160	- 2,962
B	$\lambda(1)$	0,591	0,105	0,385	0,797	5,626
	$\lambda(2)$	0,570	0,094	0,385	0,755	6,032
	$\lambda(3)$	0,148	0,102	- 0,052	0,348	1,448
	$\lambda(4)$	- 1,309	0,198	- 1,697	- 0,921	- 6,611
AB	η	- 0,829	0,081	- 0,988	- 0,670	- 10,247

Tabulka 7. Pozorované četnosti, odhady očekávaných četností a adjustovaná rezidua při distančním modelu pro tabulku 2 s vynechanou diagonálou.

A	B	1	2	3	4
1		–	21,53	4,62	0,13
		–	22,28	3,69	0,31
		–	– 0,50	0,62	– 0,33
2		84,88	–	48,23	5,30
		84,13	–	50,14	4,14
		0,50	–	– 1,05	0,99
3		8,98	32,54	–	2,06
		8,82	31,72	–	3,04
		0,09	0,43	–	– 0,87
4		1,42	8,33	10,73	–
		2,33	8,40	9,75	–
		– 0,72	– 0,04	0,87	–

$$G^2 = 1,64, P = 0,801, X^2 = 1,56, P = 0,816, df = 4$$

čtyřech stupních volnosti (obecně $df = I(I-3)$) dává dosaženou hladinu významnosti $P = 0,801$ a $P = 0,816$. Je zřejmé, že nediagonální data vyhovují distančnímu modelu (13) lépe, než data úplná. Výsledky jsou shrnuty v tabulce 7 a 8. Všimněme si, že vysoké reziduum v poli (2,1) zmizelo. To potvrzuje domněnku, že v předchozím modelu vadilo zvláštní postavení pole (2,2).

Vzdálenost mezi kategoriemi i a j nemusí být v distančním modelu vyjádřena jen pomocí $|i-j|$, ale jakoukoliv jinou rozumnou funkcí, např. $(i-j)^2$, která vyjadřuje rychlejší pokles nebo růst interakce, než $|i-j|$. Pro data z tabulky 2 bez diagonály a model

$$\log m(i,j) = \lambda + \lambda^A(i) + \lambda^B(j) + \eta(i-j)^2, \quad (14)$$

$$\sum \lambda^A(i) = \sum \lambda^B(j) = 0$$

Tabulka 8. Odhady parametrů distančního modelu pro tab. 2 s vynechanou diagonálou.

Parametr	Odhad	Std. chyba	95 % interv. spol.		z-skór	
A	$\lambda(1)$	– 0,496	0,184	– 0,857	– 0,135	– 2,693
	$\lambda(2)$	1,139	0,189	0,768	1,509	6,012
	$\lambda(3)$	– 0,143	0,152	– 0,441	0,155	– 0,941
	$\lambda(4)$	– 0,499	0,203	– 0,897	– 0,101	– 2,458
B	$\lambda(1)$	0,562	0,148	0,272	0,852	3,799
	$\lambda(2)$	0,869	0,208	0,461	1,277	4,177
	$\lambda(3)$	0,044	0,156	– 0,262	0,350	0,286
	$\lambda(4)$	– 1,475	0,297	– 2,057	– 0,893	– 4,966
AB	η	– 0,973	0,227	– 1,418	– 0,528	– 4,279

Tabulka 9. Interakční členy modelu (13) pro tabulku 4×4

0	η	2η	3η
η	0	η	2η
2η	η	0	η
3η	2η	η	0

Tabulka 10. Interakční členy modelu (15) pro tabulku 4×4

0	η_1	$\eta_1 + \eta_2$	$\eta_1 + \eta_2 + \eta_3$
η_1	0	η_2	$\eta_2 + \eta_3$
$\eta_1 + \eta_2$	η_2	0	η_3
$\eta_1 + \eta_2 + \eta_3$	$\eta_2 + \eta_3$	η_3	0

dostaneme $G^2 = 0.92$, $P = 0.922$, $X^2 = 0.90$, $P = 0.924$, $df = 4$ a největší adjustované reziduum -0.85 v poli (2,3), což je ještě lepší výsledek než pro model (12). Odhad parametru η je roven -0.301 ; model tak ukazuje, že přitažlivost kategorií vzdělání při vstupu do manželství klesá velmi rychle s rozdílem mezi nimi.

Distančních modelů existuje celá řada, zde ještě ukážeme jedno zobecnění modelu (13). Tabulka 9 ukazuje, jak vypadá interakční člen u $\log m(i,j)$ pro tabulku 4×4 a model (13). Všimněme si, že interakční členy ve vedlejších diagonálách jsou vždy stejné, tj. mezi kategoriemi se předpokládá stejná vzdálenost.

Jednoduché zobecnění, které ukazuje tabulka 10 a které vyjadřuje proměnnou vzdálenost mezi kategoriemi, vede na distanční model

$$\log m(i,j) = \lambda + \lambda^A(i) + \lambda^B(j) + \sum_{k=g}^{h-1} \eta_k, \quad (15)$$

$$g = \min(i,j) < h = \max(i,j),$$

$$\log m(i,i) = \lambda + \lambda^A(i) + \lambda^B(i),$$

$$\sum \lambda^A(i) = \sum \lambda^B(j) = 0,$$

který lze aplikovat jak na celou tabulku (pak má $df = I(I-3) + 2$ stupňů volnosti), tak pouze na nediagonální pole (pak má $df = (I-2)^2$ stupňů volnosti).⁸

Pro data z tabulky 2 a model (15) dostaneme $G^2 = 6.29$, $P = 0.392$, $X^2 = 5.83$, $P = 0.442$, $df = 6$, maximální adjustované reziduum $= -2.04$ v poli (2,2). Odhady parametrů η_1 , η_2 , η_3 jsou po řadě rovné -0.695 , -0.842 , -1.179 a jsou významné na hladině $\alpha = 0.05$. Shoda modelu s daty je prakticky stejná jako pro model (13) a protože nelze prokázat, že by se parametry η_1 , η_2 , η_3 od sebe významně lišily, nepřináší model (15) oproti (13) nic nového pro interpretaci.

(8) Pro $I = 4$ modely (13) a (15) splývají, jsou-li použity pouze na nediagonální pole.

Závěr

Uvedené modely nejsou zdaleka všechny, které jsou možné pro dvourozměrné symetrické tabulky; jsou to spíše ty základní a méně rafinované, ale dobře interpretovatelné. Řadu dalších lze nalézt např. v knize autorů Bishopové, Fienberga a Hollanda [1977], kde je též provedeno zobecnění pojmu symetrie na vícerozměrné tabulky.

Dodatek: Zadání modelů v jazyce SPSSX

A. Modely nezávislosti (1), kvazinezávislosti (6) a distanční modely (13) a (14).

Zde nejsou žádné problémy, jen nepříjemnosti spojené s dopočtem závislých parametrů (z podmínek, které jsou u modelů vždy uvedeny) a ovšem též jejich standardních chyb, z-skórů a intervalů spolehlivosti, pokud chceme o parametrech provádět statistickou inferenci. Zadání ukážeme na datech z tab. 2.

```
DATA LIST LIST / A B N
VARIABLE LABELS A 'VZDEL. MUZE' B 'VZDEL. ZENY'
VALUE LABELS A B 1 'ZAKL' 2 '2 + 3 + 4'
                3 'STR. S MAT' 4 'VYSOK.'
```

```
COMPUTE ND = 1
IF (A EQ B) ND = 0
WEIGHT BY N
LOGLINEAR A B (1,4)
/ PRINT = ALL
/ *MODEL NEZAVISLOSTI
/ DESIGN = A B
/ *MODEL QUAZINEZAVISLOSTI
/ CWEIGHT = ND
/ DESIGN = A B
```

```
BEGIN DATA :
1 1 37.72
1 2 21.53
1 3 4.62
. . .
. . .
4 2 8.33
4 3 10.73
4 4 8.52
```

```
END DATA
COMPUTE D = ABS (A - B)
LOGLINEAR A B (1,4) WITH D
/ PRINT = ALL
/ *DISTANCNI MODEL (13)
/ DESIGN = A B D
/ *DISTANCNI MODEL (13) BEZ DIAGONALY
/ CWEIGHT = ND
/ DESIGN = A B D
```

Poznámky: 1) ND je proměnná blokujeící diagonálu

2) Pro distanční model (14) se nahradí proměnná D proměnnou $DD = (A - B)^{**2}$.

3) Mezi příkazy BEGIN DATA a END DATA je zařazen vstup tabulky 2.

B. Distanční model (15).

Zde je kromě dat nutno zajistit vstup s hodnotami pomocných proměnných W1, W2, W3, jejichž koeficienty jsou parametry η_1, η_2, η_3 v jednotlivých polích tabulky (viz tab. 10).

V tabulce $I \times I$ je takových proměnných $I - 1$.

```
DATA LIST LIST / A B N W1 W2 W3
VARIABLE LABELS A 'VZDEL. MUZE' B 'VZDEL. ZENY'
```

VALUE LABELS A B 1 'ZAKL'. 2 '2+3+4'
3 'STR. S MAT' 4 'VYSOK'

WEIGHT BY N
LOGLINEAR A B (1,4) WITH W1 W2 W3
/ PRINT = ALL
/ *DISTANCNI MODEL (15)
/ DESIGN = A B W1 W2 W3

BEGIN DATA

1	1	37.72	0	0	0
1	2	21.53	1	0	0
1	3	4.62	1	1	0
1	4	0.13	1	1	1
2	1	84.88	1	0	0
2	2	161.59	0	0	0
2	3	48.23	0	1	0
2	4	5.30	0	1	1
3	1	8.98	1	1	0
3	2	32.54	0	1	0
3	3	44.42	0	0	0
3	4	2.06	0	0	1
4	1	1.42	1	1	1
4	2	8.33	0	1	1
4	3	10.73	0	0	1
4	4	8.52	0	0	0

END DATA

C. Model symetrie (9) a kvazisymetrie (12)

Zadání ověření modelů symetrie a kvazisymetrie, pokud chceme znát všechny parametry, je velmi pracná záležitost, a to tím pracnější, čím větší je rozměr I tabulky. Při častějším používání, nebo při čtení dat přímo ze souboru, bude nutné připravit zadání pro vytvoření pomocných proměnných, či příslušnou makroinstrukci.

DATA LIST LIST / A B N L1 L2 L3 L11 L21 L22 L31 L32 L33
VARIABLE LABELS A 'VZDEL. MUZE' B 'VZDEL. ZENY'
VALUE LABELS A B 1 'ZAKL.' 2 '2+3+4'
3 'STR. S MAT.' 4 'VYSOK.'

WEIGHT BY N
LOGLINEAR A B (1,4) WITH L1 L2 L3
L11 L21 L22 L31 L32 L33
/ PRINT = ALL
/ *MODEL SYMETRIE
/ DESIGN = L1 L2 L3 L11 L21 L22 L31 L32 L33
/ *MODEL QUAZISYMETRIE
/ DESIGN = A B L11 L21 L22 L31 L32 L33

BEGIN DATA

1	1	37.72	2	0	0	2	0	0	0	0	0
1	2	21.53	1	1	0	0	1	0	0	0	0
1	3	4.62	1	0	1	0	0	0	1	0	0
1	4	0.13	0	-1	-1	-2	-1	0	-1	0	0
2	1	84.88	1	1	0	0	1	0	0	0	0
2	2	161.59	0	2	0	0	0	2	0	0	0
2	3	48.23	0	1	1	0	0	0	0	1	0
2	4	5.30	-1	0	-1	0	-1	-2	0	-1	0
3	1	8.98	1	0	1	0	0	0	1	0	0
3	2	32.54	0	1	1	0	0	0	0	1	0
3	3	44.42	0	0	2	0	0	0	0	0	2
3	4	2.06	-1	-1	0	0	0	0	-1	-1	-2
4	1	1.42	0	-1	-1	-2	-1	0	-1	0	0
4	2	8.33	-1	0	-1	0	-1	-2	0	-1	0
4	3	10.73	-1	-1	0	0	0	0	-1	-1	-2

4 4 8.52 -2 -2 -2 2 2 2 2 2 2
END DATA

Určení pomocných proměnných pro zadání:

a) $I = 4$

Hodnoty pomocných proměnných pro $A = i$ a $B = j$ jsou

$$\begin{aligned} L1(i,j) &= q(i,1) + q(j,1), L2(i,j) = q(i,2) + q(j,2), L3(i,j) = q(i,3) + q(j,3), \\ L11(i,j) &= 2q(i,1) q(j,1), L21(i,j) = q(i,2) q(j,1) + q(i,1) q(j,2), \\ L22(i,j) &= 2q(i,2) q(j,2), L31(i,j) = q(i,3) q(j,1) + q(i,1) q(j,3), \\ L32(i,j) &= q(i,3) q(j,2) + q(i,2) q(j,3), L33(i,j) = 2q(i,3) q(j,3), \\ q(i,r) &= 1, i = r, \\ &= 0, i \neq r, i < 4, \\ &= -1, i = 4. \end{aligned}$$

b) Obecný případ tabulky $I \times I$

Je třeba vytvořit $(I-1)(I+2)/2$ pomocných proměnných s hodnotami pro $A = i$ a $B = j$

$$\begin{aligned} Lk(i,j) &= q(i,k) + q(j,k), k = 1, \dots, I-1 \\ Lrs(i,j) &= q(i,r) q(j,s) + q(i,s) q(j,r); r, s = 1, \dots, I-1; r \geq s \\ q(i,r) &= 1, i = r, \\ &= 0, i \neq r, i < I, \\ &= -1, i = I. \end{aligned}$$

Odhadý parametrů pro

a) model symetrie

Označme čísla, která dostaneme ve výstupech z SPSSX v části „odhadý parametrů“ u proměnných $L1, L2, \dots, L(I-1), L11, L21, \dots, L(I-1)$ jako $\mu(1), \mu(2), \dots, \mu(I-1), \mu(I,1), \mu(2,1), \dots, \mu((I-1), (I-1))$. Pak odhadý parametrů modelu (9) jsou

$$\begin{aligned} \lambda^A(i) &= \lambda^B(i) = \mu(i), 1 \leq i \leq I, \mu(i) = \sum_{i=1}^{I-1} \mu(i), \\ \lambda^{AB}(i,j) &= \lambda^{AB}(j,i) = \mu(i,j) = \mu(j,i), 1 \leq j \leq i \leq I, \\ \mu(I,j) &= \sum_{k=1, k \neq j}^{I-1} \mu(k,j) - 2\mu(j,j), \\ \lambda^{AB}(i,i) &= 2\mu(i,i), 1 \leq i \leq I, \\ \mu(I,I) &= \sum_{i=1}^{I-1} \sum_{\substack{j=1 \\ i \geq j}}^{I-1} \mu(i,j). \end{aligned}$$

b) model kvazisymetrie

Označme čísla, která dostaneme ve výstupech z SPSSX v části „odhadý parametrů“ u proměnných $A, B, L11, L21, \dots, L(I-1)$ jako $\mu^A(1), \dots, \mu^A(I-1), \mu^B(1), \dots, \mu^B(I-1), \mu(I,1), \mu(2,1) \dots \mu((I-1), (I-1))$. Odhadý parametrů modelu (12) z nich určíme jako

$$\begin{aligned} \hat{\lambda}^A(i) &= \mu^A(i), 1 \leq i \leq I, \mu^A(I) = - \sum_{i=1}^{I-1} \mu^A(i), \\ \hat{\lambda}^B(i) &= \mu^B(i), 1 \leq i \leq I, \mu^B(I) = - \sum_{i=1}^{I-1} \mu^B(i), \\ \hat{\lambda}^{AB}(i,j) &= \hat{\lambda}^{AB}(j,i) = \mu(i,j) = \mu(j,i), 1 \leq j < i \leq I, \\ \mu(I,j) &= - \sum_{k=1, k \neq j}^{I-1} \mu(k,j) - 2\mu(j,j), \end{aligned}$$

$$\hat{\lambda}^{AB}(i, i) = 2\mu(i, i), \quad 1 \leq i \leq I, \quad \mu(I, I) = \sum_{i=1}^{I-1} \sum_{j=1}^{I-1} \mu(i, j).$$

Pokud bychom nechtěli ověřovat model symetrie, vynechali bychom $L1, L2, L3$ (obecně $L(I-1)$) v řádku DATA LIST a LOGLINEAR, celý DESIGN pro model symetrie a čtvrtý, pátý a šestý (obecně $I + \text{druhý}$) sloupec mezi BEGIN DATA a END DATA. (9)

D. Model symetrie a kvazisymetrie s vynechanou diagonálou:

a) Symetrie

Pro tabulku $I \times I$ musíme vytvořit $I-1$ proměnných Lk , kde $Lk(i, j) = Lk(j, i) = q(i, k) + q(j, k)$; $k=1, \dots, I-1$; $i, j=1, \dots, I$ (též můžeme $Lk(i, i)$ položit rovno nule pro $i=1, \dots, I$, $k=1, \dots, I-1$) a dále $I(I-3)/2$ proměnných Lrs , kde $r > s$, $r = 2, \dots, I-1$, $s=1, \dots, I-3$ a

$$\begin{aligned} Lrs(i, i) &= 0, \\ Lrs(I, I-2) &= Lrs(I-2, I) = q(I, r) q(I-2, s) + q(I, s) q(I-2, r) + 1, \\ Lrs(I, I-1) &= Lrs(I-1, I) = q(I, r) q(I-1, s) + \alpha I, s) q(I-1, r) + 1, \\ Lrs(I-1, I-2) &= Lrs(I-2, I-1) = -1, \\ Lrs(i, j) &= Lrs(j, i) = q(i, r) q(j, s) + q(i, s) q(j, r) \\ \text{pro ostatní dvojice } (i, j), &\text{ přičemž} \\ q(i, r) &= 1, i = r, \\ &= 0, i \neq r, i < I, \\ &= -1, i = I. \end{aligned}$$

Prakticky to znamená, že máme-li už připravené hodnoty $Lk(i, j)$ a $Lrs(i, j)$ pro kompletní tabulku, dostaneme z nich hodnoty pro tabulku s blokovanou diagonálou tak, že $Lk(i, j)$ převezmeme beze změny, $Lrs(i, i)$ pro $i=1, \dots, I$ nahradíme nulami, $Lrs(I-1, I-2)$ a $Lrs(I-2, I-1)$ nahradíme minus jedničkami, k hodnotám $Lrs(I, I-1)$, $Lrs(I-1, I)$, $Lrs(I, I-2)$ a $Lrs(I-2, I)$ přičteme jedničky a zbytek převezmeme beze změny.

Odhadu parametrů ve výstupech není nutné přepočítávat, ale je třeba dopočítat $\hat{\lambda}^A(I)$, $\hat{\lambda}^B(I)$, $\hat{\lambda}^{AB}(I-1, I-2)$, $\hat{\lambda}^{AB}(I, j)$ pro $j=1, \dots, I-1$ podle vzorců

$$\begin{aligned} \hat{\lambda}^A(I) &= \sum_{i=1}^{I-1} \hat{\lambda}^A(i) = \hat{\lambda}^B(I) = - \sum_{i=1}^{I-1} \hat{\lambda}^B(i), \\ \hat{\lambda}^{AB}(I-1, I-2) &= \sum_{i=2}^{I-1} \hat{\lambda}^{AB}(i, 1) - \sum_{i=3}^{I-1} \hat{\lambda}^{AB}(i, 2) - \dots - \sum_{i=I-2}^{I-1} \hat{\lambda}^{AB}(i, I-3) = \\ &= - \sum_{i=1}^{I-1} \sum_{j=1}^{I-3} \hat{\lambda}^{AB}(i, j), \\ \hat{\lambda}^{AB}(I, j) &= - \sum_{k=1}^{I-1} \hat{\lambda}^{AB}(j, k), \quad j < I. \end{aligned}$$

b) Kvazisymetrie

Pro tabulku $I \times I$ musíme vytvořit jen $I(I-3)/2$ proměnných Lrs podle stejných vztahů jako

(9) Čtenář se může na příkladě tohoto dodatku přesvědčit, že ani světové renomované programové systémy nejsou příliš uživatelsky vlídné. Zde je to vidět hned na dvou aspektech: nutnost dopočítávat a dokonce přepočítávat parametry, abychom dostali interpretačně důležitou informaci a nutnost složité vytvářet pomocné proměnné pro složité zadání jednoduchých modelů. A což teprve kdybychom chtěli znát i odhady rozptýlů odhadů parametrů, které nejsou ve výstupech! Pro zájemce připomeňme tři užitečné vzorce:

$$\begin{aligned} \text{var} \left(\sum_{i=1}^N X_i \right) &= \sum_{i=1}^N \text{var} X_i + 2 \sum_{i < j} \text{cov} (X_i, X_j), \\ \text{var} aX &= a^2 \text{var} X, \text{cov} (aX, bY) = a b \text{cov} (X, Y). \end{aligned}$$

u symetrie. Rovněž dopočty odhadů parametrů provedeme stejným způsobem, zde ovšem $\lambda^A(I) \neq \lambda^B(I)$.

Následuje ukázka zadání na datech z tab. 2.

```
DATA LIST LIST / A B N L1 L2 L3 L21 L31
VARIABLE LABELS A 'VZDEL. MUZE' B 'VZDEL. ZENY'
VALUE LABELS A B 1 'ZAKL.' 2 '2 + 3 + 4'
                3 'STR. S MAT.' 4 'VYSOK.'
```

```
COMPUTE ND = 1
IF (A EQ B) ND = 0
WEIGHT BY N
LOGLINEAR A B (1,4) WITH L1 L2 L3 L21 L31
/ PRINT = ALL
/ CWEIGHT = ND
/ *MODEL SYMETRIE
/ DESIGN = L1 L2 L3 L21 L31
/ *MODEL QUAZISYMETRIE
/ DESIGN = A B L21 L31
```

BEGIN DATA

1	1	37.72	2	0	0	0	0
1	2	21.53	1	1	0	1	0
1	3	4.62	1	0	1	0	1
1	4	0.13	0	-1	-1	-1	-1
2	1	84.88	1	1	0	1	0
2	2	161.59	0	2	0	0	0
2	3	48.23	0	1	1	-1	-1
2	4	5.30	-1	0	-1	0	1
3	1	8.98	1	0	1	0	1
3	2	32.54	0	1	1	-1	-1
3	3	44.42	0	0	2	0	0
3	4	2.06	-1	-1	0	1	0
4	1	1.42	0	-1	-1	-1	-1
4	2	8.33	-1	0	-1	0	1
4	3	10.73	-1	-1	0	1	0
4	4	8.52	-2	-2	-2	0	0

END DATA

Literatura

- Bishop, Y. M. M. - Fienberg, S. E. - Holland, P. W.: *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts, and London, England, The MIT Press, 1977.
- Haberman, S. J.: *The Analysis of Frequency Data*. Chicago and London, The University of Chicago Press, 1974.
- Haberman, S. J.: *Analysis of Qualitative Data*. Vol. 1 Introductory Topics. New York, Academic Press, 1978.
- Haberman, S. J.: *Analysis of Qualitative Data*. Vol. 2 New Developments. New York, Academic Press, 1979.
- Havránek, T.: *Třídění třetího a čtvrtého stupně — úlohy a metody*. In Kapitoly ze sociologických metod a technik III. „Statistická analýza dat a kauzální závěry“ (red. Jan Řehák). Československá sociologická společnost při ČSAV, Sokola sociologických metod a technik, 1980, s. 87—121.
- Matějů, P.: *K využití logaritmicko-lineárních modelů v analýze sociální mobility*. Sociologický časopis XXI, 1985, č. 1, s. 53—70.
- Řehák, J. - Řeháková, B.: *Analýza kontingenčních tabulek: rozlišení dvou základních typů úloh a znaménkové schéma*. Sociologický časopis XIV, 1978, č. 6, s. 619—631.
- Řehák, J. - Řeháková, B.: *Porovnání distribucí u párově závislých dat ve čtvercových kontingenčních tabulkách*. Sociologický časopis XX, 1984, č. 5, s. 516—541.
- Řehák, J. - Řeháková, B.: *Analýza kategorizovaných dat v sociologii*. Praha, Academia, 1986.
- SPSSX *User's Guide*, 2nd ed. Chicago, Illinois, SPSS Inc., 1986.
- Upton, G. J. G.: *The Analysis of Cross-tabulated Data*. New York, John Wiley, 1978.