Apendix A
Log-lineární modely

Log-lineární modely jsou nástrojem pro analýzu kategorizovaných dat. Mezi tato data řadíme každou proměnnou (znak), jejíž varianty mají podobu diskrétních kategorií. Od poloviny 70. let minulého století, kdy se poznání o log-lineárních modelech začalo v sociálních vědách rozšiřovat, bylo napsáno několik učebnic o log-lineárních modelech (srov. například Bishop, Fienberg, Holland 1975; Knoke, Burke 1980; Hagenaars 1990; Agresti 1996, 2002; Powers, Xie 2000). Následující apendix vychází z těchto učebnic, v žádném ohledu je však nesupluje. Jedná se pouze o základní představení principů log-lineárního modelování. Jeho smyslem je doplnit analýzu dat popsanou v jednotlivých kapitolách, kde na podrobnější charakteristiku log-lineárních modelů nebyl prostor a bylo nutné předpokládat alespoň základní obeznámenost s touto technikou. V případě zájmu o detailnější studium log-lineárních modelů a všech jeho souvislostí s jinými pokročilými statistickými technikami je nutné využít některý z následujících textů: Bishop, Fienberg, Holland (1975); Everitt (1977); Goodman (1978); Haberman (1978, 1979); Andersen (1980); Fienberg (1980); Wickens (1989); Hagenaars (1990); Agresi (1984, 1996, 2002); Clogg, Shihadeh (1994); Christensen (1997); Long (1997); Vermunt (1997); Powers, Xie (2000).

Až do druhé poloviny 60. let 20. století byla kategorizovaná data a vztahy mezi nimi analyzovány na základě výpočtu hodnoty chí-kvadrátu, testem nezávislosti mezi proměnnými a nejrůznějšími variantami asociačních koeficientů. Když kontingenční tabulka obsahovala více než dvě proměnné, byla její analýza problematická. Na začátku 70. let 20. století Leo Goodman publikoval řadu článků o kategorizovaných datech, v nichž představil analýzu kontingenčních tabulek na základě log-lineárních modelů.[47] Přibližně ve stejné době byla vyvinuta binární logistická regrese jako způsob analýzy vztahů mezi dichotomickou závisle proměnnou a nezávisle proměnnými. Statistická analýza kategorizovaných dat se v této době dramaticky rozvíjela. V polovině

47 Většina těchto článků byla přetištěna v Goodmanových dvou knihách: *Analyzing Qualitative/Categorical Data* (1978) a *The Analysis of Crosss-Classified Data Having Ordered Categories* (1984).

70. let minulého století byly publikovány práce Bishopové, Fienberga a Hollanda (1975) a Habermana (1978, 1979), které tehdejší znalosti o log-lineárním modelování shrnovaly do přehledné a konzistentní podoby a na dlouhou dobu se staly standardními učebnicemi analýzy kategorizovaných dat.

Dnes již máme k dispozici celou řadu modelů pro kategorizovaná data. Nominální proměnné analyzujeme pomocí hierarchických (případně nehierarchických) modelů, proměnné, u nichž předpokládáme ordinalitu variant, analyzujeme pomocí log-lineárních a log-multiplikativních asociativních modelů; proměnné, které jsou ve vztahu závislosti k ostatním proměnným, analyzujeme pomocí logitových modelů. Každá tato obecná kategorie modelů obsahuje celou řadu sub-modelů pro řešení specifických případů dat.

V následujícím apendixu si nejdříve představíme kontingenční tabulky a uspořádání dat v nich pro log-lineární modely. Poté se budeme zabývat šancemi a poměry šancí v kontingenčních tabulkách, představíme si logiku log-lineárního modelování, zaměříme se na výpočet parametrů saturovaného log-lineárního modelu, ukážeme si souvislost mezi parametry log-lineárního modelu, šancemi a poměry šancí a budeme tyto parametry interpretovat. Dále se budeme zabývat principy statistického modelování, statistickými kritérii pro výběr nejadekvátnějšího log-lineárního modelu a zaměříme se na základní principy asociativních modelů pro ordinální proměnné v kontingenčních tabulkách. Také si ukážeme, s jakými typy dat při log-lineárním modelování pracujeme.

A/1 Kontingenční tabulky

Základním a nejjednodušším statistickým nástrojem pro analýzu kategorizovaných dat jsou kontingenční tabulky. Pomocí tohoto nástroje analyzujeme vztahy mezi proměnnými s omezeným počtem kategorií (variant). V případě, že máme dvě kategorizované proměnné, hovoříme o dvojrozměrné kontingenční tabulce, v případě, že analyzujeme tři kategorizované proměnné, hovoříme o trojrozměrné kontingenční tabulce. Každá další proměnná přidává do kontingenční tabulky nový rozměr, přičemž počet takto analyzovaných proměnných je teoreticky neomezený. Ve skutečnosti je ale tento počet omezen dostatečným počtem případů v polích vícerozměrné kontingenční tabulky.

Jako statistický nástroj pro analýzu kategorizovaných dat jsou kontingenční tabulky v sociálních vědách velmi populární. A to ze dvou důvodů: jednak proto, že je poměrně snadné je zkonstruovat a vztahy v nich interpretovat, a jednak proto, že se jedná o nástroj, který není omezen striktními parametrickými (distribučními) předpoklady.

I přes tyto výhody však kontingenční tabulky skrývají interpretační pasti. Jedná se o nástroj pro deskripci dat, nikoliv pro jejich analýzu a testování hypotéz. Z tohoto důvodu zjištění, která kontingenční tabulky poskytují, nemusejí být platná pro základní populaci, zvláště pokud analyzujeme vztahy mezi více proměnnými (vícerozměrné kontingenční tabulky). Jestliže v takovém případě nebudeme vztahy mezi proměnnými modelovat – to znamená, že nebudeme analyzovat vícerozměrné kontingenční tabulky v celku, ale rozložíme je na řadu dvojrozměrných subtabulek, mohou být naše závěry, vyčtené přímo z těchto dat, zavádějící. Naše intuice v takovém případě sehraje při interpretaci patrně větší roli než reálné měření.[48] Z tohoto důvodu je nezbytné kategorizovaná data ve vícerozměrných kontingenčních tabulkách analyzovat pomocí log-lineárních modelů.

A/2 Formální zápis frekvencí v kontingenčních tabulkách

Podle Leo Goodmana (1981) můžeme rozlišit tři typy vztahů mezi dvěma kategorizovanými proměnnými, jež jsou dány vzájemnými kombinacemi vysvětlujících a vysvětlovaných proměnných. Za prvé se jedná o vztah mezi dvěma vysvětlujícími proměnnými (například mezi sociální třídou a vzděláním). Za druhé se jedná o kauzální vztah mezi vysvětlovanou (závisle) proměnnou a vysvětlující (nezávisle) proměnnou (například kouření a rakovina plic). A za třetí se jedná o vztah mezi dvěma vysvětlovanými proměnnými (například postoje k interrupci a postoje k předmanželskému sexu).[49]

Rozdíly mezi těmito typy vztahů jsou konceptuální, nikoliv faktické. Všechny proměnné v jednotlivých vztazích jsou stejně zapsány a je pouze na výzkumníkovi, aby určil, která z nich je vysvětlující a která vysvětlovaná proměnná. V případě, že to lze určit, analyzujeme kategorizovaná data pomocí logistické regrese.[50] V případě, že to určit nelze, analyzujeme kategorizovaná data pomocí log-lineárních modelů.

Tabulka A.1 je čtyřrozměrná kontingenční tabulka, která ukazuje věkově homogamní a heterogamní sňatky (H) uzavřené podle sňatkového věku

48 Tento problém se označuje jako Simpsonův paradox: závěry, které učiníme na základě dvojrozměrné kontingenční tabulky, jsou v rozporu se závěry, pokud analyzujeme trojrozměrnou kontingenční tabulku. V prvním případě může být výsledkem pozitivní vztah mezi dvěma proměnnými, nicméně při zavedení třetí proměnné se tento vztah změní na negativní. Tento paradox způsobuje nerovné rozložení četností v kategoriích analyzovaných proměnných (více k tomu Christensen 1997; Agresti 2002).

49 Více k tomu také Powers a Xie (2000).

50 Má-li závisle proměnná dvě varianty, použijeme binární logistickou regresi, má-li závisle proměnná více uspořádaných variant, použijeme ordinární logistickou regresi a má-li závisle proměnná více variant, které nelze uspořádat, zvolíme multinomickou logistickou regresi (více k jednotlivým variantám logistické regrese viz Long 1997).

Tabulka A.1 Věkově homogamní a heterogamní sňatky podle sňatkového věku muže a typu věkového sňatku v letech 1994–2004 v ČR

Roky	Typ věkového sňatku	Sňatkový věk muže	Věková homogamie 0–2 roky	Věková heterogamie 3–5 let	Věková heterogamie 6+ let	Celkem
1994	tradiční	18–29	18 554	11 728	4 655	34 937
			20,23 %	*12,79 %*	*5,08 %*	*38,10 %*
		30+	1 109	1 580	4 469	7 158
			1,21 %	*1,72 %*	*4,87 %*	*7,81 %*
	netradiční	18–29	4 294	1 666	846	6 806
			4,68 %	*1,82 %*	*0,92 %*	*7,42 %*
		30+	361	276	115	752
			0,39 %	*0,30 %*	*0,13 %*	*0,82 %*
2004	tradiční	18–29	11 408	6 347	1 819	19 574
			12,44 %	*6,92 %*	*1,98 %*	*21,35 %*
		30+	3 191	4 574	6 079	13 844
			3,48 %	*4,99 %*	*6,63 %*	*15,10 %*
	netradiční	18–29	4 066	2 106	1 018	7 190
			4,43 %	*2,30 %*	*1,11 %*	*7,84 %*
		30+	771	516	147	1 434
			0,84 %	*0,56 %*	*0,16 %*	*1,56 %*
Celkem			43 754	28 793	19 148	91 695
			47,72 %	*31,40 %*	*20,88 %*	*100 %*

Poznámka: Procenta jsou sdružené (celkové) relativní četnosti.

muže (M) a typu věkového sňatku (T)[51] v letech 1994 a 2004 (R) v České republice. V této tabulce jsou zkombinovány čtyři proměnné. V případě, že si položíme otázku, jak věková homogamie (a heterogamie) souvisí se sňatkovým věkem muže a typem věkového sňatku a jak se tato souvislost mění v čase, je nezbytné tuto tabulku analyzovat pomocí log-lineárních modelů.

Pozorované (výběrové) četnosti se v log-lineárním modelování označují jako f a modelové (odhadnuté) četnosti jako F. Když variantu každé proměnné v kontingenční tabulce označíme dolním indexem – v našem případě jako i pro proměnnou H, j pro proměnnou M, k pro proměnnou T a l pro proměnnou R, kde $i = 1,\ldots, I$; $j = 1,\ldots, J$; $k = 1,\ldots, K$ a $l = 1,\ldots, L$; – můžeme každou pozorovanou četnost indikovat jako f_{ijkl} a modelovou (očekávanou) četnost jako F_{ijkl}. Dolní index označuje kategorie jednotlivých proměnných. Dohromady s horním indexem, který odkazuje k proměnným, například četnost 18554 v tabulce A.1 zapíšeme jako f_{1111}^{HMTR}, zatímco ve stejné tabulce četnost 147 zapíšeme jako f_{3222}^{HMTR}. Pozorovanou pravděpodobnost p přináležet do i-té kategorie proměnné H, j-té kategorie proměnné M, k-té kategorie proměnné T a l-té kategorie proměnné R označíme jako p_{ijkl}^{HMTR}.[52] V tabulce A.1 se $p_{1111}^{HMTR} = 18554/91695$, tedy 20,23 % (číslo 91 695 označuje všechny uzavřené sňatky). Platí tedy:

$$f_{ijkl}^{HMTR} = Np_{ijkl}^{HMTR} \tag{1}$$

Pravděpodobnost je pro populaci označována jako π. V našem případě π_{ijkl}^{HMTR} označuje pravděpodobnost, že v populaci náhodně vybraný sňatek přináleží do $H = i$, $M = j$, $T = k$ a $R = l$. Modelové četnosti F_{ijkl}^{HMTR}, které v tomto případě znamenají četnosti ve vzorku, který je přesnou kopií populace (nepředpokládáme existenci výběrové variace), pak vypočítáme podobně jako v rovnici 1:

$$F_{ijkl}^{HMTR} = N\pi_{ijkl}^{HMTR} \tag{2}$$

Symbol + ve formálním zápisu frekvencí označuje součet. V tabulce A.1 například f_{+111}^{HMTR} označuje řádkovou marginální četnost 34 937 sňatků uzavřených v roce 1994 jako věkově tradičních ve věku muže 18–29 let. Výpočet tohoto čísla zapíšeme jako $f_{+111}^{HMTR} = \sum_{i=1}^{I} f_{ijkl}^{HMTR}$, kde Σ odpovídá symbolu + a znamená součet četností napříč variantami dané proměnné. Podobně lze zapsat jakoukoliv sloupcovou marginální četnost a její výpočet. Například 19 148 věkově heterogamních sňatků 6+ let označíme jako f_{3+++}^{HMTR} a jejich výpočet zapíšeme jako $f_{3+++}^{HMTR} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} f_{ijkl}^{HMTR}$. Celkové N v tabulce A.1 pak můžeme označit jako f_{++++}^{HMTR} a jeho výpočet zapsat jako $f_{++++}^{HMTR} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} f_{ijkl}^{HMTR}$.

A/3 Šance, marginální a podmíněné šance a poměr šancí

Z tabulky A.1 můžeme vytvořit několik *marginálních tabulek*, které vzniknou součtem frekvencí napříč jednou nebo více proměnnými. Tabulka A.2 je *marginální tabulka*, která vznikla součtem četností napříč roky a napříč typy věkových sňatků (formálně tuto úpravu zapíšeme jako $f_{ij}^{HM} = f_{ij++}^{HMTR} = \sum_{k=1}^{K}\sum_{l=1}^{L} f_{ijkl}^{HMTR}$. Navíc jsou v této tabulce kategorie věkové homogamie sloučeny ze tří na dvě kategorie: věková homogamie (věková vzdálenost mezi manželi 0–2 roky) a věková heterogamie (věková vzdálenost mezi manželi 3+ let).

51 U této proměnné kategorie tradiční znamená, že muž je starší (nebo stejně starý) než jeho žena a kategorie netradiční znamená, že muž je mladší než jeho žena. Více k této proměnné viz kapitolu 2.

52 Vyjádření tabulkových četností pomocí procent či pravděpodobnosti je jedním ze způsobů standardizace dat.

Tabulka A.2 Věkově homogamní a heterogamní sňatky podle sňatkového věku muže v ČR

Sňatkový věk muže	Věková homogamie 0–2 roků	Věková heterogamie 3+ let	Celkem
18–29	38 322	30 185	68 507
	55,94 %	*44,06 %*	*100 %*
30+	5 432	17 756	23 188
	23,43 %	*76,57 %*	*100 %*
Celkem	43 754	47 941	91 695
	47,72 %	*52,28 %*	*100 %*

Poznámka: Procenta jsou řádkové relativní četnosti.

V tabulce A.2 vidíme, že 48 % uzavřených sňatků je věkově homogamních a 52 % věkově heterogamních. V případě, že dáme tyto dva podíly do poměru, dostaneme šance věkově homogamních sňatků oproti věkově heterogamním sňatkům: 47,72 / 52,28 = (43754 / 91695) / (47941 / 91695) = 43754 / 47941 = 0,913. Jedná se o *marginální šance* – jsou počítány z marginálních (celkových) tabulkových distribucí. Marginální šance věkově heterogamních sňatků oproti věkově homogamním sňatkům dostaneme prostou záměnou čísel: 52,28 / 47,72 = 47941 / 43754 = 1 / 0,913 = 1,096. Pravděpodobnost věkově homogamního sňatku je 0,913 pravděpodobnosti věkově heterogamního sňatku; pravděpodobnost věkově heterogamního sňatku je 1,096 pravděpodobnosti věkově homogamního sňatku. Šance na věkově homogamní sňatek jsou tedy 0,913krát menší ve srovnání s věkově heterogamním sňatkem a šance na věkově heterogamní sňatek jsou 1,096krát větší ve srovnání s věkově homogamním sňatkem. Rovnosti obou pravděpodobností (nebo frekvencí) odpovídá číslo 1. Šance se může pohybovat od 0 do ∞, přičemž čísla menší než 1 znamenají nižší šance a čísla větší než 1 vyšší šance. I když se to na první pohled nezdá, obě čísla (0,913 a 1,096) věcně říkají totéž. V termínech násobků a podílu jsou stejně vzdálená od čísla 1 (0,913 = 1/1,096). Vzhledem k číslu 0 jejich ekvidistanci ukazuje převod na jejich přirozené logaritmy.[53] Rovnost přirozených logaritmů šancí tedy ukazuje číslo 0 a přirozené logaritmy šancí se mohou pohybovat od -∞ do ∞.

Šance není totéž co pravděpodobnost, i když mezi šancemi a pravděpodobností existuje vztah. V případě, že pravděpodobnost na věkově homogamní sňatek u mužů označíme jako p a pravděpodobnost na věkově heterogamní sňatek jako opak, tedy jako $1-p$, šanci O (z anglického *odd*) vypočítáme jako:

$$O = \frac{p}{1-p} \tag{3}$$

Jednoduchou úpravou této rovnice vypočítáme pravděpodobnost ze šance jako:

$$p = \frac{O}{O+1} \tag{4}$$

Obecně platí, že čím vyšší je šance na událost, tím vyšší je také pravděpodobnost této události. Například šance 2 : 1 odpovídá 66,7 %,[54] šance 3 : 1 odpovídá 75 % a šance 50 : 1 odpovídá 98 %. Čím více se naopak šance blíží číslu 0, tím více se také pravděpodobnost blíží číslu 0. Tabulka A.3 ukazuje vztah mezi vybranými šancemi a procenty.

Tabulka A.3 Vztah mezi šancemi a procenty

Šance	Procenta
1000 : 1	99,90 %
500 : 1	99,80 %
100 : 1	99,01 %
50 : 1	98,04 %
10 : 1	90,91 %
5 : 1	83,33 %
3 : 1	75,00 %
2 : 1	66,67 %
1 : 1	50,00 %
1 : 2	33,33 %
1 : 3	25,00 %
1 : 5	16,67 %
1 : 10	9,09 %
1 : 50	1,96 %
1 : 100	0,99 %
1 : 500	0,20 %
1 : 1000	0,10 %

Při analýze vztahu mezi dvěma proměnnými pracujeme s *podmíněnými šancemi*. Výpočet podmíněných šancí je totožný s výpočtem marginálních šancí. Oproti marginálním šancím se podmíněné šance liší tím, že jsou počítány pro jednotlivé podskupiny – přesněji řečeno v rámci variant jiné proměnné. V tabulce A.2 jsou šance věkově homogamních sňatků oproti věkově heterogamním sňatkům pro muže, kteří uzavřeli sňatek mezi 18–29 lety,

53 ln (0,913) = – 0,091 a ln (1,096) = 0,091.

54 Vypočítané jako (2/(2+1) × 100).

1,270;[55] pro muže, kteří uzavřeli sňatek později (30+ let), jsou tyto šance 0,306. Pokud se muž ožení do 29 let, má větší šanci uzavřít věkově homogamní než věkově heterogamní sňatek. V případě, že se ožení ve 30 letech nebo později, šance na věkově homogamní sňatek jsou ve srovnání s věkově heterogamním sňatkem mnohem nižší.

Čím více se podmíněné šance na jednu a tutéž věc v rámci kategorií jiné proměnné od sebe odlišují, tím silnější vztah mezi dvěma zkoumanými proměnnými existuje. Když srovnáme dvě podmíněné šance, dostaneme *poměr šancí*. V našem případě je poměr šancí 4,149.[56] Šance na věkově homogamní sňatek (ve srovnání s věkově heterogamním sňatkem) u mužů ženících se do 29. roku života je 4,149krát vyšší oproti šancím mužů ženících se později. Vypočítáme-li kontrastní podmíněné šance – na věkově heterogamní sňatek mužů do 29 let –, dostaneme číslo 0,241.[57] Šance na věkově heterogamní sňatek (oproti věkově homogamnímu sňatku) mužů ženících se do 29 let jsou 0,241krát menší než šance na tentýž sňatek mužů ženících se ve 30 a více letech. Pokud dáme do poměru podmíněné šance mužů ženících se ve 30 a více letech na věkově homogamní sňatek a podmíněné šance mužů ženících se do 29 let na věkově homogamní sňatek (ve srovnání s věkově heterogamním sňatkem), dostaneme také číslo 0,241. V případě, že vypočítáme podmíněné šance mužů ženících se ve 30 a více letech na věkově heterogamní sňatek (ve srovnání s věkově homogamním sňatkem) a pak podmíněné šance mužů ženících se do 29 let na věkově heterogamní sňatek (ve srovnání s věkově homogamním sňatkem), jejich poměrem dostaneme opět číslo 4,149. Tyto čtyři možné reciproční poměry šancí pro kontingenční tabulku A.2 shrnuje tabulka A.4.

Tabulka A.4 Čtyři možné poměry šancí pro dvojrozměrnou tabulku A.2

Sňatkový věk muže	Věková homogamie 0–2 roky	Věková heterogamie 3+ let
18–29	4,15	0,24
30+	0,24	4,15

Čísla 4,149 a 0,241 opět věcně sdělují totéž. Při interpretaci je však nutné dávat pozor, ke kterým kategoriím dvou proměnných se vztahují (jejich ekvidistanci vzhledem k číslu 0 opět ukazuje převod na přirozené logaritmy těchto čísel). To znamená, že poměr šancí je pro dvojrozměrnou tabulku symetrickým indikátorem asociace. I když můžeme poměr šancí identifikovat pro každé pole dvojrozměrné tabulky, k jejímu popisu stačí znát pouze jeden poměr šancí (zbylé poměry šancí v případě nutnosti vyvodíme z tohoto poměru). Obecně poměr šancí *OR* (z anglického *odds ratio*) z pozorovaných četností vypočítáme:

$$OR_{11} = \frac{f_{11}}{f_{12}} \Big/ \frac{f_{21}}{f_{22}} = \frac{f_{11}}{f_{21}} \Big/ \frac{f_{12}}{f_{22}} = \frac{f_{11} f_{22}}{f_{12} f_{21}} \qquad (5)$$

V případě, že se poměr šancí rovná číslu 1 (podmíněné šance se neliší), najdeme stejné rozložení věkově homogamních a věkově heterogamních sňatků u mužů, kteří se žení do 29 let, a u mužů, kteří se žení ve 30 a více letech. Věková homogamie a sňatkový věk muže jsou statisticky nezávislé.

Při interpretaci můžeme poměr šancí (nebo také podmíněné či marginální šance) vyjádřit v procentech. Například je-li poměr šancí číslo 2, znamená to dvakrát větší šance na událost, tedy 200 případů na každých 100 případů, čili o 100 % větší šance. Naopak je-li poměr šancí číslo 0,4, znamená to o 60 % menší šance na událost, neboli výskyt 40 případů na každých 100 případů. Pro tyto převody obecně platí, že je-li poměr šancí větší než číslo 1, číslo 1 od tohoto poměru šancí odečteme a výsledek vynásobíme číslem 100. Je-li poměr šancí menší než číslo 1, toto číslo od čísla 1 odečteme a výsledek opět vynásobíme číslem 100. V obou případech poměr šancí interpretujeme jako procentuálně větší či menší než referenční kategorie.[58]

A/4 Invariance poměru šancí

Poměr šancí je invariantní ke změnám v datech. Jeho velikost nepoznamenává ani změna v celkovém počtu případů, ani změny v marginálních řádkových nebo sloupcových distribucích kontingenční tabulky.

Představme si situaci, že by se celkový počet případů v souboru změnil *n*-krát – tedy o faktor *c* (například třikrát). Všechny frekvence tím změníme o stejný faktor *c*, nicméně poměr šancí zůstane nezměněn, protože:

$$OR_{11} = \frac{cf_{11}}{cf_{12}} \Big/ \frac{cf_{21}}{cf_{22}} = \frac{cf_{11}cf_{22}}{cf_{12}cf_{21}} = \frac{f_{11} f_{22}}{f_{12} f_{21}} \qquad (6)$$

55 Vypočítáno jako 55,94 / 44,06 = 38322 / 30185.
56 Vypočítáno jako 1,270 / 0,306 = (38322 / 30185) / (5432 / 17756).
57 Vypočítáno jako: (30185 / 38322) / (17756 / 5432)) = 0,306/1,270 = 1 / 4,149.

58 Tato procentuální interpretace šancí nesmí být ale zaměňována s převodem šancí na procenta podle rovnice 3 a 4 (viz tabulku A.3). V procentuální interpretaci šancí jde o vyjádření velikosti jednoho čísla vzhledem k číslu jinému v procentech (v rozmezí 0 % až ∞ %), převádíme-li však šance na procenta, říkáme, jaké procento odpovídá dané šanci (v rozmezí 0 % až 100 %).

Pokud změníme marginální řádkové četnosti v kontingenční tabulce tak, že první řádek tabulky vynásobíme faktorem c a druhý řádek tabulky faktorem d nebo změníme marginální sloupcové četnosti tak, že první sloupec vynásobíme faktorem k a druhý sloupec faktorem l, celkový poměr šancí zůstane opět nezměněn, protože:

$$OR_{11} = \frac{cf_{11}}{cf_{12}} \Big/ \frac{df_{21}}{df_{22}} = \frac{cf_{11}df_{22}}{cf_{12}df_{21}} = \frac{f_{11}f_{22}}{f_{12}f_{21}} \tag{7}$$

$$OR_{11} = \frac{kf_{11}}{lf_{12}} \Big/ \frac{kf_{21}}{lf_{22}} = \frac{kf_{11}lf_{22}}{lf_{12}kf_{21}} = \frac{f_{11}f_{22}}{f_{12}f_{21}} \tag{8}$$

Poměry šancí jsou invariantní ke změnám v marginálních distribucích, jelikož tyto změny se odrážejí v proporčním nárůstu nebo poklesu napříč řádky i sloupci. Díky této charakteristice je poměr šancí využíván v analýzách, které potřebují odhlédnout od změn v marginálních distribucích (například změny zaměstnanecké struktury rodičů a jejich potomků v sociálněstratifikačním výzkumu). Pokud bychom měli dva náhodné výběry ze stejné populace provedené ve stejném časovém okamžiku, jeden o velikosti 1 000 respondentů a druhý o velikosti 10 000 respondentů, a měli bychom dvě stejné kontingenční tabulky z těchto dat, poměry šancí v obou tabulkách by se neměly lišit, pokud by neexistovala výběrová variace.

A/5 Nonredundantní počet poměrů šancí v kontingenční tabulce

K popsání vztahů mezi proměnnými v kontingenční tabulce potřebujeme méně poměrů šancí, než je polí v kontingenční tabulce. U dvojrozměrné kontingenční tabulky je nonredundantní (nezbytný, někdy také lokální) počet poměrů šancí dán vzorcem $(I\text{-}1)(J\text{-}1)$, kde I označuje počet variant pro proměnnou I a J označuje počet variant pro proměnnou J (tabulka o rozměrech $I \times J$). Zbylé poměry šancí jsou odvoditelné z těchto nonredundantních poměrů šancí.

Pro jakoukoliv dvojrozměrnou $I \times J$ tabulku vypočítáme poměry šancí podle následující rovnice:

$$OR_{ij} = \frac{f_{ij}f_{(i+1)(j+1)}}{f_{i(j+1)}f_{(i+1)j}}, \qquad i = 1,\ldots,I-1;\ j = 1,\ldots,J-1 \tag{9}$$

Jelikož každý poměr šancí ve dvojrozměrné kontingenční tabulce zahrnuje kombinace dvou kategorií jedné a dvou kategorií jiné proměnné, můžeme pro tabulku $I \times J$ spočítat mnoho poměrů šancí. Například máme-li tabulku o rozměrech 2 × 3, spočítáme dva nonredundantní poměry šancí (v případě, že budeme počítat i reciproční poměry šancí, tak čtyři a v případě, že budeme počítat všechny poměry šancí, tak dvanáct poměrů šancí). Ke smysluplnému popsání asociace mezi proměnnými v této tabulce potřebujeme ale pouze dva poměry šancí. Zbylé, nereciproční poměry šancí lze z těchto dvou poměrů šancí odvodit jejich vynásobením. Podle rovnice (9) vypočítáme nejdříve poměr šancí pro řádek 1 a 2 a sloupec 1 a 2. Poté vypočítáme poměr šancí pro řádek 1 a 2 a sloupec 2 a 3. Chceme-li spočítat poměr šancí pro řádek 1 a 2 a sloupec 1 a 3, můžeme to udělat buď podle rovnice (9), anebo stačí vynásobit poměr šancí řádku 1 a 2 a sloupce 1 a 2 a poměr šancí řádku 1 a 2 a sloupce 2 a 3.

A/6 Poměr šancí ve vícerozměrné kontingenční tabulce

Poměr šancí lze také počítat mezi třemi a více kategorizovanými proměnnými. Kdybychom do tabulky A.2 zavedli další proměnnou, jíž by byl rok, v němž byl sňatek uzavřen (dvě kategorie: 1994 a 2004), mohli bychom se ptát, jak se liší vztah mezi věkovou homogamií a sňatkovým věkem muže podle roků, v nichž byl sňatek uzavřen. Při tomto výpočtu nejdříve spočítáme podmíněné poměry šancí – pro každý rok zvláště – a potom spočítáme poměr mezi dvěma poměry šancí. Rovnice pro tento výpočet je následující.

$$OR_{11} = \frac{f_{111}f_{122}}{f_{112}f_{121}} \Big/ \frac{f_{211}f_{222}}{f_{212}f_{221}} = \frac{f_{111}f_{122}f_{212}f_{221}}{f_{112}f_{121}f_{211}f_{222}} \tag{10}$$

V čitateli rovnice jsou všechny frekvence, u nichž součet dolních indexů dává liché číslo; ve jmenovateli jsou všechny frekvence, u nichž suma dolních indexů dává sudé číslo. Pro trojrozměrnou interakci se někdy také používá označení *interakce druhého řádu* (*second-order interaction*) (Bishop, Fienberg, Holland 1975; Rudas 1998). Tento poměr šancí musíme interpretovat s ohledem na třetí proměnnou. Jedná se o vyjádření toho, do jaké míry (kolikrát) se podmíněný poměr šancí (dvojrozměrná interakce) liší v jednotlivých variantách (kategoriích) třetí proměnné. Čím vyšší nebo nižší je toto číslo než číslo 1, tím větší význam třetí proměnná hraje v trojrozměrné tabulce. V případě, že toto číslo odpovídá číslu 1, podmíněné poměry šancí jsou totožné. Hovoříme o homogennosti podmíněných poměrů šancí. Třetí proměnná v trojrozměrné kontingenční tabulce nehraje roli a při analýze na ni nemusíme brát zřetel (tj. lze analyzovat jen dvourozměrnou tabulku).

Princip výpočtu poměru šancí v trojrozměrné tabulce lze použít i pro tabulky o více rozměrech. Čtyřrozměrnou interakci označíme jako *interakci třetího řádu*, pětirozměrnou interakci jako *interakci čtvrtého řádu* atd. (Rudas 1998). Rovnice pro výpočet jednotlivých poměrů šancí je totožná s rovnicí (10), obsahuje pouze všechny nezbytné frekvence dané počty tabulkových

rozměrů. Máme-li lichý počet rozměrů, všechny frekvence, jejichž součet (dolních) indexů je lichý, umísťujeme do čitatele rovnice, a všechny frekvence, jejichž součet (dolních) indexů je sudý, dáváme do jmenovatele rovnice. Máme-li naopak sudý počet tabulkových rozměrů, všechny frekvence, které mají sudou sumu indexů, dáváme do čitatele rovnice, a všechny frekvence, jejichž suma je lichá, umisťujeme do jmenovatele rovnice. Výpočet poměru šancí pro čtyřrozměrnou tabulku (čtyři proměnné o dvou variantách) ukazuje rovnice 11.

$$OR_{11} = \frac{f_{1111}f_{1122}}{f_{1121}f_{1112}} \Big/ \frac{f_{1211}f_{1222}}{f_{1221}f_{1112}} \Big/ \frac{f_{2111}f_{2122}}{f_{2121}f_{2112}} \Big/ \frac{f_{2211}f_{2222}}{f_{2221}f_{2212}} = \frac{f_{1111}f_{1122}f_{1212}f_{1221}f_{2112}f_{2121}f_{2211}f_{2222}}{f_{2111}f_{1211}f_{1121}f_{1112}f_{1222}f_{2122}f_{2212}f_{2221}} \quad (11)$$

A/7 Parciální šance, aritmetický a geometrický průměr

Parciální šance jsou definovány jako průměrné podmíněné šance. Parciální šance na věkově homogamní sňatek v tabulce A.2 odpovídá na otázku, jaká je šance na věkově homogamní sňatek oproti věkově heterogamnímu sňatku *v průměru* pro kategorie sňatkového věku muže. Podobně parciální šance na uzavření sňatku muže ve věku 18–29 let odpovídá na otázku, jaká je jeho šance oženit se v tomto věku oproti pozdějšímu věku (30+ let) *v průměru* pro věkově homogamní a heterogamní sňatky.

Parciální šance počítáme jako geometrický průměr z podmíněných šancí. Geometrický průměr, stejně jako aritmetický průměr, je mírou centrální tendence (Hendl 2004). Abychom lépe pochopili princip výpočtu geometrického průměru, a tedy parciálních šancí, začneme definicí a logikou aritmetického průměru.

Aritmetický průměr je definován jako součet všech hodnot dělený počtem pozorování (rovnice 12). Suma odchylek hodnot od aritmetického průměru se rovná vždy číslu 0 (rovnice 13). Jedná se o vlastnost aritmetického průměru. Charakterizujeme-li tedy v souboru každého člověka průměrnou hodnotou – například průměrným věkem při uzavření sňatku, podhodnocujeme jeho sňatkový věk naprosto stejně, jako jej nadhodnocujeme (v termínech rozdílů a součtů). V tomto smyslu leží aritmetický průměr ve středu distribuce hodnot, z nichž je spočítán, neboť součet odchylek všech hodnot od něj je nulový.

$$\bar{X} = (\sum_{i=1}^{N} X_i)/N \quad (12)$$

$$\sum_{i=1}^{N} (X_i - \bar{X}) = 0 \quad (13)$$

S geometrickým průměrem pracujeme v případech, kdy lze uvažovat o poměrech mezi čísly.[59] K vysvětlení logiky geometrického průměru J. A. Hagenaars (1990) uvádí následující příklad: Cena koně je $100. Dva muži mají za úkol odhadnout jeho cenu. Kůň připadne tomu z nich, jehož odhad bude blíže skutečné ceně koně. První muž tipuje cenu $10, druhý muž tipuje $1000. Komu připadne kůň? Pokud bychom odhadnuté ceny odečítali od skutečné ceny (v logice aritmetického průměru), první muž by byl vítězem. Kůň ale nepřipadne žádnému z mužů, protože (v logice geometrického průměru) oba muži tipovali stejně. První muž podcenil cenu koně 10krát, druhý muž přecenil jeho cenu rovněž 10krát.

Geometrický průměr vypočítáme jako součin všech hodnot odmocněný počtem pozorování (rovnice 14). V našem případě by cena koně ze dvou odhadů ($10 a $1000) byla $100. Součin podílů hodnot a hodnoty geometrického průměru se rovná vždy číslu 1 (rovnice 15). Jedná se o vlastnost geometrického průměru. Charakterizujeme-li tedy v souboru každého člověka geometrickým průměrem – opět například věkem při uzavření sňatku – podhodnocujeme jeho sňatkový věk v násobcích, stejně jako jeho sňatkový věk (opět v násobcích) nadhodnocujeme. V tomto smyslu leží geometrický průměr přesně ve středu distribuce hodnot, z nichž je počítán, neboť součin jednotlivých podílů hodnot a geometrického průměru je číslo 1.

$$\bar{X}_{geom} = \sqrt[N]{X_1 X_2 \ldots X_N} = \left(\prod_{n=1}^{N} X_i\right)^{1/N} \quad (14)$$

$$\prod_{n=1}^{N} (X_i / \bar{X}_{geom}) = 1 \quad (15)$$

Aritmetický průměr je míra vhodná pro případy, kdy pracujeme se součty a rozdíly – s aditivními modely. Geometrický průměr používáme v těch případech, kdy pracujeme s násobky a podíly, tedy se šancemi a poměry šancí – s multiplikativními modely.

Mezi aritmetickým a geometrickým průměrem existuje vztah. Pokud hodnoty, z nichž je geometrický průměr počítán, převedeme na přirozené logaritmy a spočítáme z nich aritmetický průměr, exponent tohoto aritmetického průměru se rovná původnímu geometrickému průměru. Například geometrický průměr z hodnot 2, 3 a 4 je 2,885. Aritmetický průměr z hodnot přirozených logaritmů čísel 2, 3 a 4 je 1,059. Platí, že exp(1,059) = 2,885 a ln(2,885) = 1,059. Přirozený logaritmus geometrického průměru se rovná

59 Většinou se jedná o proměnné, v jejichž distribucích má 0 přirozený počátek (vyjadřuje neexistenci jevu) a jejichž rozpětí nabývá hodnot 0 až ∞. Četnost u takové proměnné ukazuje, kolikrát daný jev nastal.

aritmetickému průměru vypočítanému z přirozených logaritmů hodnot geometrického průměru. A naopak: exponent aritmetického průměru se rovná geometrickému průměru, který je vypočítán z exponentů hodnot aritmetického průměru.

V tabulce A.2 platilo, že podmíněné šance na věkově homogamní sňatek oproti věkově heterogamnímu sňatku pro muže, kteří se oženili mezi 18.–29. rokem, byly 1,270; pro muže, kteří se oženili později (30+ let) tyto šance byly 0,306. Parciální šance na věkově homogamní sňatek je počítána jako geometrický průměr z těchto dvou podmíněných šancí: $\sqrt{(1{,}270)(0{,}306)} = 0{,}623$. V průměru věkových kategorií jsou šance na věkově homogamní sňatek mužů menší než na věkově heterogamní sňatek. To koresponduje s marginálními šancemi mužů na věkově homogamní sňatek oproti věkově heterogamnímu sňatku v téže tabulce.

A/8 Saturovaný log-lineární model

Rovnice saturovaného log-lineárního modelu je podobná rovnici lineární regrese. Na levé straně rovnice je přirozený logaritmus frekvencí (přesněji řečeno měřené četnosti jsou konvertovány na svůj přirozený logaritmus), pravá strana rovnice je lineární kombinací vysvětlujících parametrů. Z tohoto důvodu hovoříme o *log-lineárních* či *logaritmicko-lineárních* modelech – o přirozených logaritmech četností předpokládáme, že jsou lineární funkcí sady parametrů.[60]

Saturovaný model znamená, že rovnice obsahuje všechny nezbytné parametry k objasnění velikostí (přesněji řečeno velikostí přirozených logaritmů) frekvencí. Žádné omezení pro proměnné v modelu nepředpokládáme, stejně jako nepředpokládáme žádné omezení pro vztahy mezi proměnnými. Všechny parametry a kombinace vztahů mezi nimi jsou v modelu přítomny.

Data v tabulce A.5 ukazují věkově homogamní a heterogamní sňatky podle sňatkového věku muže a typu věkového sňatku v roce 2004 v České republice. Jedná se o trojrozměrnou kontingenční tabulku, kterou (v multiplikativní podobě) popisuje následující saturovaný model (parametry jsou označeny jako τ).

$$F_{ijk}^{HMT} = \eta \tau_i^H \tau_j^M \tau_k^T \tau_{ij}^{HM} \tau_{ik}^{HT} \tau_{jk}^{MT} \tau_{ijk}^{HMT} \qquad (16)$$

60 V anglosaských zemích se pro přirozený logaritmus používá zkratka *log*, zatímco u nás zkratka *ln* (zkratka *log* označuje dekadický logaritmus). Jelikož se jedná o log-lineární modely, bude v dalším textu pro přirozený logaritmus používána zkratka *log*.

Tabulka A.5 Věkově homogamní a heterogamní sňatky podle sňatkového věku muže a typu věkového sňatku v roce 2004 v ČR

Typ věkového sňatku	Sňatkový věk muže	Věková homogamie 0–2 roky	Věková heterogamie 3+ let	Celkem
tradiční	18–29	11 408	8 166	19 574
	30+	3 191	10 653	13 844
netradiční	18–29	4 066	3 143	7 209
	30+	771	663	1 434

Modelové frekvence *F* v jednotlivých polích kontingenční tabulky jsou vyjádřeny jako součin jednotlivých parametrů a jejich kombinací. Z tohoto důvodu nazýváme model multiplikativní (součinový). Každou četnost ovlivňuje jednak parametr η (obdoba konstanty v regresní analýze), dále jednotlivé kategorie proměnné *H* (věková homogamie), *M* (sňatkový věk muže) a *T* (typ věkového sňatku) – parametry τ_i^H, τ_j^M, τ_k^T, dvojrozměrné interakce mezi těmito kategoriemi proměnných *HM*, *HT* a *MT* – parametry τ_{ij}^{HM}, τ_{ik}^{HT}, τ_{jk}^{MT} a trojrozměrná interakce *HMT* – parametr τ_{ijk}^{HMT}.

Levá strana rovnice však není „klasická" závisle proměnná. Jedná se o počet případů v jednotlivých polích kontingenční tabulky – o výskyt událostí. Z tohoto důvodu se někdy log-lineárním modelům říká frekvenční modely. Frekvence neboli četnosti jsou poměrným kardinálním znakem – číslo 0 má přirozený počátek a záporný počet událostí nemůže nastat (například -1 dítě nebo -5 sňatků je nesmyslný údaj). Neobvyklé je také jiné vyjádření událostí (četností) v kontingenční tabulce než v celých číslech (například 1,8 sebevražd nebo 2,3 sňatků je nelogický údaj). V tomto ohledu se rovnice pro log-lineární modely liší od rovnice lineární regrese, která taková omezení nemá (číslo 0 obvykle není přirozeným počátkem a rozpětí hodnot se může pohybovat od $-\infty$ do $+\infty$, hodnoty případů nemají pouze podobu celých kladných čísel).

Další podstatný rozdíl ve srovnání s rovnicí lineární regrese spočívá v tom, že u log-lineárních modelů nás zajímá především to, co je umístěno na pravé straně rovnice, nicméně v regresní analýze se zajímáme o to, co je umístěno jak na pravé, tak na levé straně rovnice. Stručně řečeno: klasické rozdělení na závisle (vysvětlovanou) proměnnou a nezávisle (vysvětlující) proměnné (levá a pravá strana rovnice v lineární regresi) v případě log-lineárních modelů neplatí. Závisle proměnná neexistuje – suplují ji frekvence v jednotlivých polích kontingenční tabulky.

S tím souvisí další vlastnost log-lineární analýzy. Tato analýza je dimenzována a lze ji použít pouze na agregovaná, tabulková data. V případě, že

máme individuální data, musíme z nich buď vytvořit kontingenční tabulku (kolik rozměrů tabulka bude mít, záleží na tom, kolik proměnných do ní z dat vložíme),[61] nebo použijeme některou z variant logistické regrese (binární, ordinální nebo multinomickou logistickou regresi), které však již předpokládají rozdělení na závisle a nezávisle proměnnou.

V případě, že obě strany rovnice 16 převedeme na přirozené logaritmy, dostaneme následující rovnici:

$$G_{ijk}^{HMT} = \theta + \lambda_i^H + \lambda_j^M + \lambda_k^T + \lambda_{ij}^{HM} + \lambda_{ik}^{HT} + \lambda_{jk}^{MT} + \lambda_{ijk}^{HMT} \quad (17)$$

kde

$$G_{ijk}^{HMT} = \ln(F_{ijk}^{HMT}),\ \theta = \ln(\eta),\ \lambda_i^H = \ln(\tau_i^H),\ \lambda_i^H = \ln(\tau_j^M) \ldots \lambda_{ijk}^{HMT} = \ln(\tau_{ijk}^{HMT})$$

Jedná se o aditivní (součtové) vyjádření saturovaného modelu pro tabulku A.5 (v této podobě je rovnice podobná rovnici lineární regrese). Přirozený logaritmus každé četnosti v tabulce je lineární kombinací přirozeného logaritmu celkového průměru a přirozených logaritmů efektů jednotlivých kategorií proměnných a vztahů mezi nimi. Úprava multiplikativní rovnice do podoby přirozených logaritmů se provádí z důvodů numerické identifikace modelu.[62] Jedná se o logaritmickou transformaci, po níž je již model lineární (v parametrech). Exponenciováním této rovnice dostaneme původní multiplikativní rovnici. Exponenciální podoba rovnice 17 je následující:

$$e^{G_{ijk}^{HMT}} = e^{(\theta + \lambda_i^H + \lambda_j^M + \lambda_k^T + \lambda_{ij}^{HM} + \lambda_{ik}^{HT} + \lambda_{jk}^{HT} + \lambda_{ijk}^{HMT})} \quad (18)$$

$$e^{G_{ijk}^{HMT}} = e^{\theta} e^{\lambda_i^H} e^{\lambda_j^M} e^{\lambda_k^T} e^{\lambda_{ij}^{HM}} e^{\lambda_{ik}^{HT}} e^{\lambda_{jk}^{HT}} e^{\lambda_{ijk}^{HMT}} \quad (19)$$

A/9 Restrikce parametrů pro identifikaci log-lineárního modelu

Rovnice 17, 18 a 19 pro saturovaný log-lineární model mají z hlediska identifikace parametrů více řešení. Například pro trojrozměrnou interakci bychom identifikovali tolik parametrů, kolik je polí v kontingenční tabulce. Nicméně samotné efekty kategorií proměnných nás ve statistické analýze dat nezajímají. Samy o sobě, bez referenčního rámce (například efektu jiné kategorie) nemají význam a nejsou interpretovatelné. Otázkou, která nás tedy zajímá, je, zda a jak se efekt jedné varianty proměnné liší od jiné varianty stejné proměnné. Nakolik například v tabulce A.5 sňatkový věk mužů 18–29 let ve srovnání s věkem 30 a více let ovlivňuje šance na věkově homogamní sňatek.

Tato relační perspektiva řeší problém identifikace parametrů v log-lineárních (ale i všech ostatních regresních) modelech. Buď můžeme parametry vypočítat tak, že jsou vztaženy ke svému průměru, nebo můžeme parametry identifikovat k sobě navzájem. Obě řešení dávají věcně stejné výsledky. Představme si, že máme muže, který získá v matematickém testu 100 bodů, a ženu, jejíž skóre v tomtéž testu je 170 bodů. Průměrné skóre z těchto dvou případů je 135 bodů. Ve srovnání s tímto průměrem pohlaví v případě ženy zvyšuje skóre o 35 bodů, v případě muže snižuje skóre také o 35 bodů. Celkový rozdíl mezi oběma skóry je 70 bodů ve prospěch ženy nebo v neprospěch muže – záleží na tom, z jaké perspektivy data interpretujeme. Ke stejnému závěru bychom dospěli, kdybychom vztáhli obě kategorie k sobě navzájem – přesněji řečeno, pokud bychom se ptali, o kolik je skóre v jedné kategorii vyšší než skóre ve druhé (referenční) kategorii (70 bodů ve prospěch ženy ve srovnání s mužem nebo 70 bodů v neprospěch muže ve srovnání s ženou).

První řešení se v log-lineárním modelování nazývá *effect coding* (někdy také ANOVA *coding*), druhé řešení se nazývá *dummy coding*. Effect coding znamená, že efekty log-lineárních parametrů jsou identifikovány ve vztahu k průměrnému efektu – jedná se o odchylky od průměrného efektu. Dummy coding znamená, že efekty log-lineárních modelů jsou identifikovány k sobě navzájem. Jedná se o odchylky od jednoho, arbitrárně zvoleného parametru, jehož hodnota je nahrazena konstantou, obvykle číslem 0 (v log-lineárním režimu) nebo číslem 1 (v multiplikativním režimu), což znamená, že efekt neexistuje.

Effect coding a dummy coding jsou dvě rozdílné parametrizace, které lze použít pro identifikaci parametrů stejného modelu. Ať použijeme první nebo druhé řešení, parametry jsou vzájemně převoditelné (Rudas 1998). S ohledem na zvolenou parametrizaci musíme však odhadnuté parametry adekvátně interpretovat (Alba 1987; Kaufman, Schervish 1986, 1987; Long 1984). V log-lineárních modelech je rozšířenější používat effect coding, v regresních a logistických modelech dummy coding.[63]

Effect coding znamená, že součet log-lineárních parametrů λ vymezených dolním indexem se rovná číslu 0 (charakteristika odchylek od aritmetického

61 Některé statistické programy transformují individuální data do podoby kontingenčních tabulek, aniž by to jejich uživateli explicitně sdělily.

62 Pracovat s přirozenými logaritmy čísel při maximálně věrohodnostním odhadu parametrů je numericky snazší než pracovat s celými čísly. Na podobu výsledku přitom tato úprava nemá vliv.

63 Také rozdílné statistické programy pro odhad log-lineárních modelů mají implementovány rozdílné typy parametrizace efektů. Například GLIM, Stata, S-Plus nebo SAS používají dummy coding. SPSS nebo LEM mají přednastavený effect coding, který lze ale velmi pohodlně změnit na dummy coding.

průměru) a součin multiplikativních parametrů τ se rovná číslu 1 (charakteristika odchylek od geometrického průměru). Rovnice 20 a 21 ukazují tuto restrikci pro parametry saturovaného log-lineárního modelu tabulky A.5.

$$\sum_{i=1}^{I}\lambda_i^H = \sum_{j=1}^{J}\lambda_j^M = \sum_{k=1}^{K}\lambda_k^T = \sum_{i=1}^{I}\lambda_{ij}^{HM} = \sum_{j=1}^{J}\lambda_{ij}^{HM} = \ldots = \sum_{i=1}^{I}\lambda_{ijk}^{HMT} = \sum_{j=1}^{J}\lambda_{ijk}^{HMT} = \sum_{k=1}^{K}\lambda_{ijk}^{HMT} = 0 \quad (20)$$

$$\prod_{i=1}^{I}\tau_i^H = \prod_{j=1}^{J}\tau_j^M = \prod_{k=1}^{K}\tau_k^T = \prod_{i=1}^{I}\tau_{ij}^{HM} = \prod_{j=1}^{J}\tau_{ij}^{HM} = \ldots = \prod_{i=1}^{I}\tau_{ijk}^{HMT} = \prod_{i=1}^{J}\tau_{ijk}^{HMT} = \prod_{k=1}^{K}\tau_{ijk}^{HMT} = 1 \quad (21)$$

V případě použití parametrizace *dummy coding* je nezbytné si vždy zvolit jednu z kategorií analyzovaných proměnných, která bude kategorií referenční. Pokud si zvolíme u každé proměnné první kategorii, znamená to, že log-lineární parametry se pro tuto kategorii rovnají číslu 0 – multiplikativní parametry číslu 1. Pro tabulku A.5 saturovaného log-lineárního modelu toto omezení ukazují rovnice 22 a 23.

$$\lambda_1^H = \lambda_1^M = \lambda_1^T = \lambda_{1j}^{HM} = \lambda_{i1}^{HM} = \ldots = \lambda_{1jk}^{HMT} = \lambda_{i1k}^{HMT} = \lambda_{ij1}^{HMT} = 0 \quad (22)$$

$$\tau_1^H = \tau_1^M = \tau_1^T = \tau_{1j}^{HM} = \tau_{i1}^{HM} = \ldots = \tau_{1jk}^{HMT} = \tau_{i1k}^{HMT} = \tau_{ij1}^{HMT} = 1 \quad (23)$$

Tato omezení umožňují parametry log-lineárních modelů identifikovat. Počet nonredundantních (nezbytných) parametrů pro saturovaný log-lineární model v trojrozměrné kontingenční tabulce je dán vzorcem $(I\text{-}1)(J\text{-}1)(K\text{-}1)$, kde I, J a K označují dimenze (počty kategorií) analyzovaných proměnných. Dohromady s celkovým efektem počet nonredundantních parametrů saturovaného modelu odpovídá rozměrům kontingenční tabulky. Pro saturovaný model dvojrozměrné tabulky o rozměrech 3 × 3 (dvě proměnné, každá obsahuje tři kategorie) je například nezbytné (při restrikci dummy coding) odhadnout devět parametrů: hlavní průměr (jeden parametr), $(I\text{-}1)$ a $(J\text{-}1)$ pro každou proměnnou (čtyři parametry) a $(I\text{-}1)(J\text{-}1)$ parametrů pro interakce mezi variantami obou proměnných (čtyři parametry). Pro trojrozměrnou tabulku 3 × 3 × 3 by to bylo (opět při restrikci dummy coding) 27 nonredundantních parametrů saturovaného modelu.

A/10 Interpretace parametrů saturovaného log-lineárního modelu

Výpočet vybraných parametrů, identifikovaných jako *effect coding*, saturovaného log-lineárního modelu pro data tabulky A.5 ukazují rovnice 24 až 27.[64]

64 U saturovaného modelu není nutné modelové četnosti odhadovat. Modelové četnosti F odpovídají pozorovaným četnostem F a všechny parametry jsou počítány z těchto pozorovaných četností. U jiných než saturovaných modelů je nezbytné modelové četnosti již pomocí speciálních algoritmů odhadnout (viz následující podkapitoly A/13 a A/14).

Zbylé parametry vypočítáme podle stejných vzorců, ovšem s jinými (jim odpovídajícími) hodnotami a restrikcemi. Vzorce pro výpočet parametrů, identifikovaných jako *dummy coding*, najde čtenář v příslušné literatuře (srov. Bishop, Fienberg, Holland 1975; Haberman 1978, 1979).

$$\eta = \left(\prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K}F_{ijk}\right)^{1/IJK} \qquad \theta = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}G_{ijk} \quad (24)$$

$$\tau_i^H = \frac{\left(\prod_{i=1}^{I}\prod_{j=1}^{J}F_{ijk}\right)^{1/JK}}{\eta} \qquad \lambda_i^H = \frac{1}{JK}\sum_{i=1}^{I}\sum_{j=1}^{J}G_{ijk} - \theta \quad (25)$$

$$\tau_{ij}^{HM} = \frac{\left(\prod_{k=1}^{I}F_{ijk}\right)^{1/K}}{\eta\tau_i^H\tau_j^M} \qquad \lambda_{ij}^{HM} = \frac{1}{K}\sum_{k=1}^{K}G_{ijk} - \theta - \lambda_i^H - \lambda_j^M \quad (26)$$

$$\tau_{ijk}^{HMT} = \frac{F_{ijk}}{\eta\tau_i^H\tau_j^M\tau_k^T\tau_{ij}^{HM}\tau_{ik}^{HT}\tau_{jk}^{MT}} \qquad \lambda_{ijk}^{HMT} = G_{ijk} - \theta - \lambda_i^H - \lambda_j^M - \lambda_k^T - \lambda_{ij}^{HM} - \lambda_{ik}^{HT} - \lambda_{jk}^{MT} \quad (27)$$

Tabulka A.6 ukazuje hodnoty parametrů saturovaného modelu identifikované jako *effect coding* a *dummy coding* pro data tabulky A.5. Podle rovnice 24 celkový efekt (celkový parametr) η a θ ukazuje průměrnou hodnotu frekvencí v kontingenční tabulce (z tohoto důvodu se také někdy nazývá jako hlavní průměr). Geometrický průměr frekvencí v tabulce A.5 je 3 463 a aritmetický průměr přirozených logaritmů frekvencí je 8,150. Celkový efekt je poznamenán celkovým počtem případů N v tabulce. Pokud změníme velikost N, změní se také velikost tohoto parametru.

Efekty jednotlivých proměnných (marginální či jednorozměrné efekty) odrážejí distribuční zešikmení napříč kategoriemi. Znamená to, že ukazují, v jakých kategoriích je více nebo méně případů. Pokud je například uzavřeno více sňatků ve věku 18–29 let než ve věku 30 a více let, můžeme říci, že první věková kategorie působí silněji na velikost četností v kontingenční tabulce. Numericky se jedná o geometrický průměr z frekvencí dané kategorie proměnné, který je poté vztažený k celkovému průměru (rovnice 25). V tabulce A.5 nás například zajímá efekt sňatkového věku muže 18–29 let (τ_1^M). Z polí

$M = 1$ spočítáme geometrický průměr a poté zjistíme, nakolik – přesněji kolikrát – se liší od celkového průměru.[65] Číslo 1,696 říká, že sňatky jsou v této kategorii 1,696krát častější než v (celkovém) průměru frekvencí tabulky. Pro kategorii $M = 2$ je to 0,590,[66] což znamená, že ve věkové kategorii mužů 30 a více let je 0,590krát méně sňatků než v průměru frekvencí celé tabulky. Log-lineární marginální parametry λ mají stejnou interpretaci jako multiplikativní marginální parametry τ, ale s tím rozdílem, že odchylka od celkového průměru není vyjádřena jako poměr, ale jako rozdíl.

Parciální šance jsme definovali jako geometrický průměr z podmíněných šancí. V tabulce A.5 jsou parciální šance na sňatky mužů ve věku 18–29 let oproti věku 30 a více let 2,875.[67] Když dáme do poměru parametry τ_1^M / τ_2^M z tabulky A.6, odpovíme na otázku, kolikrát je průměrná četnost sňatků u mužů ve sňatkovém věku 18–29 let větší či menší než u mužů ve sňatkovém věku 30 a více let ($\tau_1^M / \tau_2^M = 2.875$). Jedná se o tytéž parciální šance. Sňatky mužů ve věku 18–29 let jsou v průměru 2,875krát častější než sňatky mužů ve 30 nebo více letech. Sňatky mužů ve 30 a více letech (τ_2^M / τ_1^M) jsou naopak v průměru 0,348krát méně časté ve srovnání se sňatky mužů v 18–29 letech.[68]

U hierarchických log-lineárních modelů nejsou marginální efekty proměnných interpretačně příliš užitečné. Později uvidíme, že v marginálních distribucích hierarchické log-lineární modely odpovídají marginálním distribucím pozorovaných dat. Navíc rozšíření kontingenční tabulky o každou další dimenzi mění efekt proměnné na tabulkové četnosti. To znamená, že velikost marginálních parametrů není nezávislá na počtu dimenzí kontingenční tabulky.

Z hlediska interpretace nás mnohem více než marginální efekty zajímají interakce (asociace) mezi proměnnými (dvojrozměrné nebo vícerozměrné efekty a jim odpovídající parametry). Začněme dvojrozměrnými interakcemi saturovaného modelu. Tyto parametry ukazují průměrný vztah mezi dvěma proměnnými kontrolovaný pro kategorie ostatních proměnných v kontingenční tabulce. Například v tabulce A.6 parametry τ_{ij}^{MH} indikují parciální interakce mezi proměnnými *MH* pro konstantní úroveň *T*. V případě restrikce *effect coding* tyto parametry vypočítáme jako geometrické průměry z polí *MH*, které vztáhneme k součinu parametrů nižšího řádu (η, τ_i^H, τ_j^M) (rovnice 26). Například parametr $\tau_{11}^{MH} = 1.229$,[69] což znamená, že průměrné frekvence

65 Tedy: $(11408 \times 8166 \times 4066 \times 3143)^{0.25} / 3463 = 1,696$.
66 Vypočítáno jako $1/1,696 = (3191 \times 10653 \times 771 \times 663)^{0.25} / 3463$.
67 Vypočítáno jako $[(11408 / 3191) \times (4066 / 771) \times (8166 / 10653) \times (3143 / 663)]^{0.125}$.
68 $\tau_2^M / \tau_1^M = (\tau_2^M)^2$ a podobně pak $\tau_1^M / \tau_2^M = (\tau_1^M)^2$, přičemž $(\tau_2^M)^2 = 1/(\tau_1^M)^2$ a $(\tau_1^M)^2 = 1/(\tau_2^M)^2$.
69 Vypočítáno jako $(11408 \times 4066)^{0.5} / (1,696 \times 0,944 \times 3463)$.

Tabulka A.6 Parametry saturovaného modelu pro proměnné v kontingenční tabulce A.5

			Effect coding		Dummy coding	
Efekt	**Popis kategorií**	**Kategorie**	λ	τ	λ	τ
Celkový			8,150	3463	9,342	11408
T	tradiční	1	0,773	2,166	0	1
	netradiční	2	-0,773	0,462	-1,032	0,356
M	18–29 let	1	0,528	1,696	0	1
	30+ let	2	-0,528	0,590	-1,274	0,280
H	věková homogamie	1	-0,058	0,944	0	1
	věková heterogamie	2	0,058	1,060	-0,334	0,716
TM		11	-0,276	0,759	0	1
		12	0,276	1,318	0	1
		21	0,276	1,318	0	1
		22	-0,276	0,759	-0,389	0,678
TH		11	-0,160	0,852	0	1
		12	0,160	1,173	0	1
		21	0,160	1,173	0	1
		22	-0,160	0,852	0,077	1,080
MH		11	0,206	1,229	0	1
		12	-0,206	0,814	0	1
		21	-0,206	0,814	0	1
		22	0,206	1,229	1,540	4,664
TMH		111	0,179	1,196	0	1
		112	-0,179	0,836	0	1
		121	-0,179	0,836	0	1
		122	0,179	1,196	0	1
		211	-0,179	0,836	0	1
		212	0,179	1,196	0	1
		221	0,179	1,196	0	1
		222	-0,179	0,836	-1,433	0,239

mezi variantami M1 a H1 jsou 1,229krát vyšší než frekvence, které generuje součin parametrů nižšího řádu.

V případě vícerozměrné kontingenční tabulky je dvojrozměrná interakce parciální interakcí. U trojrozměrné (a vícerozměrné) tabulky se jedná o geometrický průměr ze dvou (nebo více) podmíněných dvojrozměrných interakcí – tedy o *průměrnou podmíněnou interakci*. V tabulce A.5 je například (parciální) interakce mezi sňatkovým věkem muže a věkovou homogamií (parametr τ_{ij}^{MH}) geometrickým průměrem z podmíněných interakcí mezi

sňatkovým věkem muže a věkovou homogamií pro tradiční ($\tau_{ij}^{MH|T}{}_{1}$) a netradiční ($\tau_{ij}^{MH|T}{}_{2}$) sňatky.[70]

Jak podmíněné, tak parciální dvojrozměrné interakce souvisejí s poměry šancí. V tabulce A.5 pro pole f_{111} vypočítáme podmíněný poměr šancí (pro věkově tradiční sňatek) na věkově homogamní sňatek podle sňatkového věku jako poměr dvou podmíněných interakcí ($\tau_{11}^{MH|T}{}_{1}/\tau_{12}^{MH|T}{}_{1}$) / ($\tau_{21}^{MH|T}{}_{1}/\tau_{22}^{MH|T}{}_{1}$), čemuž odpovídá ($\tau_{11}^{MH|T}{}_{1}$)^4.[71] Podobně vypočítáme pro tabulkové pole f_{112} podmíněný poměr šancí (nyní pro věkově netradiční sňatek). Geometrický průměr z těchto podmíněných poměrů šancí odpovídá parametru (τ_{11}^{MH})^4 v tabulce A.6 neboli interakci mezi sňatkovým věkem muže a věkovou homogamií (*MH*) vyjádřenou jako poměr šancí ($\tau_{11}^{MH}/\tau_{12}^{MH}$) / ($\tau_{21}^{MH}/\tau_{22}^{MH}$) pro konstantní úroveň proměnné manželství.[72]

Podobně jako u efektu jednotlivých proměnných na tabulkové četnosti, také u dvojrozměrných interakcí jsou velikosti parametrů odlišné podle přítomnosti nebo nepřítomnosti další proměnné v kontingenční tabulce.

Poslední parametr, který v tabulce A.6 zbývá objasnit, je trojrozměrná interakce τ_{ijk}^{MHT}. Existuje souvislost mezi věkovou homogamií, sňatkovým věkem mužů a typem věkového sňatku? Tuto otázku můžeme přeformulovat do tří následujících otázek: 1. Liší se souvislost mezi věkovou homogamií a sňatkovým věkem mužů pro věkově tradiční a netradiční sňatek? 2. Liší se souvislost mezi věkovou homogamiií a typem věkového sňatku pro brzký (18–29 let) a pozdější (30+ let) sňatkový věk? 3. Liší se souvislost mezi typem věkového sňatku a sňatkovým věkem pro věkově homogamní a věkově heterogamní sňatky? Na všechny tyto otázky odpovídá trojrozměrná interakce, neboť parametry této interakce jsou v hierarchických log-lineárních modelech symetrické.

Podle rovnice 27 parametr τ_{ijk}^{MHT} vypočítáme jako podíl příslušné frekvence a součinu efektů nižšího řádu. Jedná se o odchylku tabulkové četnosti od četnosti generované hlavním průměrem a jednorozměrnými (marginálními) a dvojrozměrnými parametry. Zatímco tedy například parametr τ_{ij}^{MH} je *průměrnou* podmíněnou dvojrozměrnou interakcí *MH*, parametr τ_{ijk}^{MHT} říká, nakolik – kolikrát – se podmíněné dvojrozměrné interakce ($\tau_{ij}^{MH|T}{}_{1}$ a $\tau_{ij}^{MH|T}{}_{2}$) od sebe odlišují. Jinými slovy řečeno, do jaké míry se podmíněné dvojrozměrné interakce odlišují od parciální (průměrné) interakce (parametr τ_{ij}^{MH}). Totéž platí i pro zbylé dvojrozměrné interakce (parametry τ_{ik}^{MT} a τ_{jk}^{HT}), přičemž trojrozměrná interakce (parametr τ_{ijk}^{MHT}) má stejnou hodnotu.[73] Pokud se všechny trojrozměrné parametry τ_{ijk}^{MHT}= 1 (v multiplikativním režimu) nebo λ_{ijk}^{MHT}= 0 (log-lineárním režimu), trojrozměrná interakce neexistuje a všechny podmíněné dvojrozměrné interakce (vztahy) mezi proměnnými jsou stejné.

V tabulce A.6 se τ_{111}^{MHT}= 1,196 a τ_{112}^{MHT}= 0,836. To znamená, že podmíněná interakce mezi věkovou homogamií a brzkým sňatkovým věkem mužů (18–29 let) ve věkově tradičních sňatcích ($\tau_{11}^{MH|T}{}_{1}$) je 1,196krát vyšší než průměrná interakce (τ_{11}^{MH}= 1,229). Podmíněný vztah mezi věkovou homogamií a sňatkovým věkem 18–29 let je 1,470 ($\tau_{11}^{MH|T}{}_{1}$).[74] Interakce mezi věkovou homogamií a brzkým sňatkovým věkem muže (18–29 let) ve věkově netradičních sňatcích $\tau_{11}^{MH|T}{}_{2}$ je 0,836krát ($1/\tau_{11}^{MH|T}{}_{1}$) menší než průměrná interakce (τ_{11}^{MH}= 1,229). Podmíněný vztah mezi věkovou homogamií a sňatkovým věkem 18–29 let je 1,027 ($\tau_{11}^{MH|T}{}_{2}$).[75]

Na základě těchto údajů můžeme konstatovat, že věková homogamie souvisí se sňatkovým věkem mužů. Pokud nebereme zřetel na typ věkového sňatku, muži, kteří se ožení dříve (18–29 let), mají 1,229krát (o 22,9 %) vyšší šance uzavřít věkově homogamní sňatek a naopak 0,814krát (o 18,6 %) menší šanci uzavřít věkově heterogamní sňatek než průměrný muž (generovaný součinem efektů nižších řádů). Pokud bereme zřetel na typ věkového sňatku, vztah mezi sňatkovým věkem mužů a věkovou homogamiií má stejný (pozitivní) směr – pro věkově tradiční sňatky je však mnohem silnější než pro sňatky netradiční.

Vztah mezi věkovou homogamií a typem věkového sňatku bez ohledu na sňatkový věk muže je 0,852 (ve věkově tradičních sňatcích je šance na věkově homogamní sňatek nižší, ve věkově netradičních naopak vyšší). Zahrneme-li sňatkový věk mužů, je interakce mezi věkově homogamním sňatkem a věkově tradičním sňatkem pro sňatkový věk mužů 18–29 let 0,712[76] a pro sňatkový věk mužů 30 a více let 1,018.[77] Šance, že věkově tradiční sňatek bude věkově homogamní, jsou při brzkém uzavření sňatku mužů nižší a při jejich pozdějším sňatkovém věku mírně vyšší než průměr.

Podobně interpretujeme vztah mezi sňatkovým věkem muže a typem věkového sňatku podle věkové homogamie. Bez ohledu na to, zda je sňatek vě-

---

70 $\tau_{ij}^{MH} = \left[\left(\tau_{ij}^{MH|T}{}_{1}\right)\left(\tau_{ij}^{MH|T}{}_{2}\right)\right]^{0,5}$. Přitom podmíněnou interakci pro T = 1 vypočítáme jako $\tau_{ij}^{MH|T}{}_{1} = \frac{F_{ij1}}{\eta_1^T \tau_{i\ 1}^{H|T} \tau_{j\ 1}^{M|T}}$, a pro T = 2 jako $\tau_{ij}^{MH|T}{}_{2} = \frac{F_{ij2}}{\eta_2^T \tau_{i\ 2}^{H|T} \tau_{j\ 2}^{M|T}}$. Například parciální interakce $\tau_{11}^{MH} = \left[\left(\tau_{11}^{MH|T}{}_{1}\right)\left(\tau_{11}^{MH|T}{}_{2}\right)\right]^{0,5}$ = (1,470 × 1,027)^{0,5} = 1,229 a parciální interakce $\tau_{12}^{MH} = \left[\left(\tau_{12}^{MH|T}{}_{1}\right)\left(\tau_{12}^{MH|T}{}_{2}\right)\right]^{0,5}$ = ((1/1,470) × (1/1,027))^{0,5} = 0,814.

71 OR_1^T = ((1,470 / 0,681) / (0,681 / 1,470) =(1,470)^2 = 4,663. Toto číslo odpovídá dvojrozměrnému parametru při restrikci *dummy coding*.

72 $(OR_1^T)\times(OR_2^T)$ = (4,663 × 1,112)^{0,5} = (1,229)^4 = (1,229 / 0,814) / (0,814 / 1,229) = 2,278.

73 Symetričnost trojrozměrné interakce můžeme vyjádřit: $\tau_{ijk}^{MHT} = \tau_{ij\ k}^{MH|T}/\tau_{ij}^{MH} = \tau_{ik\ j}^{MT|H}/\tau_{ik}^{MT} = \tau_{jk\ i}^{HT|M}/\tau_{jk}^{HT}$.

74 Vypočteno jako 1,229 × 1,196.

75 Vypočteno jako 1,229 × 0,836.

76 Vypočteno jako 0,852 × 0,836.

77 Vypočteno jako 0,852 × 1,196.

kově homogamní nebo heterogamní, šance mužů na věkově tradiční sňatek ve sňatkovém věku 18–29 let jsou nižší (0,759krát) než průměr, ve sňatkovém věku 30 a více let jsou naopak vyšší než průměr (1,318krát). Pozdější sňatkový věk muže znamená vyšší šance na věkově tradiční podobu sňatku, naopak brzký sňatkový věk znamená vyšší šance na věkově netradiční podobu sňatku. Pro věkově homogamní sňatky je interakce mezi sňatkovým věkem mužů a typem věkového sňatku 0,907, pro věkově heterogamní sňatky je tato interakce 0,759.[78] Šance na netradiční podobu sňatku v brzkém sňatkovém věku jsou vyšší pro věkově homogamní sňatky než pro věkově heterogamní sňatky.

Shrneme-li to, můžeme říci, že jednotlivé parametry log-lineárního modelu ukazují, jak celková velikost vzorku, marginální distribuce proměnných, dvojrozměrné a vícerozměrné interakce mezi proměnnými „přispívají" k vysvětlení variace četností v kontingenční tabulce. Při omezení *effect coding* je každý parciální efekt τ počítán jako geometrický průměr z odpovídajících podmíněných efektů a každý další efekt vyššího řádu ukazuje odchylku podmíněných efektů od parciálního efektu. Hodnoty τ se mohou pohybovat od 0 do ∞, hodnoty λ od $-\infty$ do ∞.[79] Efekt nepozorujeme, pokud $\tau = 1$ a $\lambda = 0$. Nevýhodou hodnot τ parametrů je, že nejsou symetricky rozložené okolo čísla 1. Negativní a pozitivní efekty nemůžeme přímo srovnávat.[80] Oproti tomu hodnoty λ parametrů jsou symetricky rozložené okolo čísla 0, což znamená, že pozitivní a negativní efekty jsou přímo srovnatelné. Nevýhodou λ parametrů ovšem je, že musejí být interpretovány v termínech logaritmů frekvencí, za nimiž si je obtížné představit konkrétní četnosti případů. Oproti tomu τ parametry interpretujeme velmi snadno – jako poměry mezi frekvencemi nebo pravděpodobnostmi.

Dosadíme-li vypočítané (nezaokrouhlené) parametry (z tabulky A.6) do jednotlivých log-lineárních rovnic, dostaneme modelové (v případě saturovaného modelu pozorované) četnosti z tabulky A.5. Například pro frekvence F_{111} nebo F_{112} a jejich přirozené logaritmy platí:

78 Vypočteno jako 0,759 × 1,196 a jako 0,759 × 0,836.

79 Krajní meze těchto intervalů jsou dosažitelné pouze teoreticky. Prakticky by to znamenalo, že by tabulková frekvence byla nulová. V takovém případě je však parametr log-lineárního modelu neidentifikovatelný, protože jev nenastal. Vyskytne-li se takový případ, je nutné buď číslo 0 nahradit velmi nízkým číslem (pracujeme-li s výběry z populace, předpokládáme, že případ se vyskytuje, ale není obsažen v našem vzorku), nebo jej považovat za „strukturální" nulu (pracujeme-li s vyčerpávajícím šetřením, musíme konstatovat, že případ se nevyskytuje) a při odhadu parametrů vzít tuto skutečnost v úvahu (více k tomu viz Knoke, Burke 1980).

80 Pokud například chceme odpovědět, zda $\tau = 1,2$ je silnější interakce než $\tau = 0,8$, musíme negativní efekt nejdříve převést na pozitivní efekt (1/0,8 = 1,25) a pak oba efekty z hlediska velikosti srovnat.

$$F_{111} = 11408 = 3463 \times 2,166 \times 1,696 \times 0,944 \times 0,759 \times 0,852 \times 1,229 \times 1,196$$

$$\ln(F_{111}) = 9,342 = 8,150 + 0,773 + 0,528 + (-0,058) + (-0,276) + (-0,160) + 0,206 + 0,179$$

$$F_{112} = 8166 = 3463 \times 2,166 \times 1,696 \times 1,060 \times 0,759 \times 1,173 \times 0,814 \times 0,836$$

$$\ln(F_{112}) = 9,008 = 8,150 + 0,773 + 0,528 + 0,058 + (-0,276) + 0,160 + (-0,206) + (-0,179)$$

A/11 Nesaturovaný log-lineární model

Saturovaný model není příliš interpretačně užitečný. Jedná se o parametrizaci pozorovaných četností – pozorované případy převedeme na odpovídající počet parametrů (Powers, Xie 2000). Interpretujeme stejný počet parametrů jako počet četností, což je jedno a totéž. Takový model je sice přesný (to znamená, že vypočítané modelové frekvence v jednotlivých polích kontingenční tabulky se neliší od pozorovaných – měřených – frekvencí), nicméně není úsporný (neobsahuje méně parametrů než pozorování), a proto není ani interpretačně užitečný.

Smyslem statistického modelování hromadných dat je najít úspornější model (popis struktury dat), než je model saturovaný (princip parsimonie). Úspornější znamená jednodušší (některé z parametrů jsou vynechány nebo jinak omezeny). Jednodušší ale obvykle znamená i méně přesný (modelová data se liší od pozorovaných dat). Ideálem statistického modelování je proto najít vždy takový model, který je ještě dostatečně přesný (modelová data se významně neliší od pozorovaných dat), který je ale také maximálně možně úsporný (obsahuje co nejméně vazeb mezi proměnnými ve srovnání se saturovaným modelem). Přesnost a úspornost jsou v protikladu. Zvyšováním přesnosti snižujeme úspornost a naopak. Jedná se o soukolí, v němž je každý výzkumník při explanaci proměnných a vazeb mezi nimi. Zvýšením počtu vazeb ve struktuře modelu zvyšujeme jeho přesnost, nicméně na úkor úspornosti a jeho interpretovatelnosti. Opomíjením vazeb ve struktuře modelu snižujeme přesnost modelu, tím však snižujeme také pravděpodobnost, že budeme moci na jeho základě pozorovaná data ještě interpretovat.

Většina vědců preferuje úspornost před přesností. Jednodušší model je pro interpretaci vhodnější než model složitější. Tento princip je obsažen v zákonu Occamovy břitvy. Podle něho by výzkumník měl vždy hledat takové řešení, které je nejjednodušší, přitom ovšem data (generovaná v rámci) modelu, který interpretuje, by se statisticky významně neměla lišit od pozoro-

vaných (měřených) dat. V případě, že můžeme volit ze dvou stejných řešení, nicméně jedno je složitější a druhé je jednodušší, měli bychom volit vždy to jednodušší nebo méně komplikované řešení.

Modely, které neobsahují všechny nezbytné parametry k popsání kontingenční tabulky, se nazývají nesaturované. V log-lineárním modelování existuje mnoho způsobů, jak parametry omezit. V případě, že předpokládáme, že efekt parametru odpovídá číslu 0 (v aditivní rovnici modelu) nebo číslu 1 (v multiplikativní rovnici modelu) a přitom zachováváme pravidlo, že všechny vyšší interakce, v nichž se tento parametr také vyskytuje, se rovnají číslu 0 (nebo číslu 1), hovoříme o hierarchických log-lineárních modelech. Například pokud předpokládáme, že asociace *HM* (vztah mezi věkovou homogamií a sňatkovým věkem muže) pro data v tabulce A.5 neexistuje (odpovídá číslu 0 v aditivní rovnici modelu), musíme předpokládat, že všechny interakce vyššího řádu, které interakci *HM* obsahují, rovněž neexistují (rovnají se také číslu 0 v aditivní rovnici). Rovnice pro takový nesaturovaný model pak vypadá následovně:

$$G_{ijk}^{HMT} = \theta + \lambda_i^H + \lambda_j^M + \lambda_k^T + \lambda_{ik}^{HT} + \lambda_{jk}^{MT} \qquad (28)$$

Jiným příkladem nesaturovaného hierarchického log-lineárního modelu může být model nezávislosti, kdy předpokládáme, že interakce mezi věkovou homogamií a sňatkovým věkem mužů nebo mezi věkovou homogamií a typem sňatku nebo mezi sňatkovým věkem mužů a typem sňatku, neexistuje. Rovnice pro takový model je následující:

$$G_{ijk}^{HMT} = \theta + \lambda_i^H + \lambda_j^M + \lambda_k^T \qquad (29)$$

Při hledání modelu, který adekvátně reprodukuje pozorovaná data (je přesný) a přitom obsahuje pouze tolik vazeb, kolik je nezbytně nutné (je úsporný), se obvykle postupuje dvojím způsobem. Buď začneme odhadem saturovaného modelu a postupně odstraňujeme z modelu interakce vyššího a pak nižšího řádu (postupujeme tedy od nejsložitějších po nejednodušší vazby v datech), až najdeme model, jehož reprodukce dat je stále ještě přesná, a přitom je tento model dostatečně úsporný. Nebo začneme nejjednodušším modelem (obvykle modelem nezávislosti mezi proměnnými) a postupně přidáváme složitější interakce, až nalezneme model, který adekvátně reprodukuje pozorovaná data, přitom ovšem je stále ještě dostatečně úsporný. Prvnímu postupu se říká sestupný výběr modelu (*backward selection*), druhý postup se nazývá vzestupný výběr modelu (*forward selection*). V log-lineárním modelování je rozšířenější druhý postup.

Víme již, že princip hierarchie znamená, že jsou-li v log-lineárním modelu přítomny interakce vyššího řádu, jsou zároveň také přítomny všechny efekty nižších řádů proměnných, které interakci vyššího řádu tvoří. Je-li v modelu například přítomna trojrozměrná interakce mezi proměnnými, jsou implicitně přítomny všechny dvojrozměrné a jednorozměrné interakce stejných proměnných, včetně hlavního průměru. Oproti hierarchické struktuře modelů existuje také nehierarchická struktura. Jedná se o log-lineární modely, které obsahují interakce vyšších řádů mezi proměnnými, aniž by byly v modelu přítomné interakce nižších řádů nebo efekty jednotlivých proměnných, včetně hlavního průměru. Tyto modely nejsou ale příliš rozšířené. Jednak proto, že není vždy snadné odhadnout jejich modelové četnosti, a jednak proto, že jsou obtížně interpretovatelné.

V log-lineárním modelování je obvyklé model specifikovat pomocí jednotlivých proměnných – přesněji řečeno pomocí arbitrárně zvolených písmen pro tyto proměnné ve složených závorkách. Saturovaný model pro data v tabulce A.5 (rovnice 17, 18 a 19) můžeme buď specifikovat jako {T M H TM TH MH THM}, nebo jako {TMH}. V prvním případě písmena odpovídají jednotlivým parametrům v rovnicích 17, 18 nebo 19 (hlavní průměr se nespecifikuje), přičemž písmena vedle sebe znamenají interakce proměnných. Ve druhém případě je uvedena pouze trojrozměrná interakce, protože v hierarchické struktuře modelu jsou interakce nižších řádů a efekty jednotlivých proměnných automaticky přítomny. V případě, že chceme vyjádřit nezávislost mezi proměnnými, ponecháme mezi písmeny jednoduše mezeru (například model {T M H} odpovídá rovnici 29).

Modelové proměnné a vazby mezi nimi indikované písmeny v závorkách nemají pouze symbolický, ale také praktický význam. Označují marginální kontingenční tabulky generované (pod jednotlivými modely) z celkové kontingenční tabulky. To znamená, že máme-li hypotézu, která určuje vztahy mezi proměnnými, marginální distribuce pro tyto proměnné v kontingenčních tabulkách odpovídají marginálním distribucím pro tytéž proměnné v pozorovaných datech. Modelové frekvence F a pozorované frekvence f se sice liší (s výjimkou saturovaného modelu), jejich součet napříč řádky nebo sloupci se však neliší od stejného součtu pozorovaných četností napříč řádky nebo sloupci. Procedury k odhadnutí modelových (očekávaných) četností tedy vycházejí z totožnosti modelových a pozorovaných marginálních distribucí kontingenčních tabulek. Toto východisko je součástí tradičního testu chí-kvadrátu, kdy očekávané četnosti odpovídají modelu nezávislosti mezi dvěma proměnnými (poměr šancí $OR = 1$), přitom však v marginálních

Tabulka A.7 Četnosti saturovaného modelu {TMH} a četnosti modelu {TM MH} pro data tabulky A.5

Typ věkového sňatku	Sňatkový věk	Model {TMH} Věková homogamie 0–2 roky	Model {TMH} Věková heterogamie 3+ let	Model {TM MH} Věková homogamie 0–2 roky	Model {TM MH} Věková heterogamie 3+ let
tradiční	18–29 let	11 408	8 166	11 308,967	8 265,033
	30+ let	3 191	10 653	3 590,25	10 253,88
netradiční	18–29 let	4 066	3 143	4 165,033	3 043,967
	30+ let	771	663	371,875	1 062,125

distribucích mezi modelem nezávislosti a pozorovanými daty nenajdeme rozdíl.

Tabulka A.7 ukazuje četnosti dvou modelů pro data tabulky A.5 – saturovaného modelu a modelu, který předpokládá existenci pouze dvou dvojrozměrných interakcí – jednak mezi typem sňatku (*T*) a sňatkovým věkem muže (*M*) a jednak mezi sňatkovým věkem muže (*M*) a věkovou homogamií (*H*). Z hlediska marginálií lze oba modely zapsat následovně: 1) {TMH}; 2) {TM MH}. Součet četností u proměnných TM druhého modelu odpovídá součtu těchto četností u saturovaného modelu a součet četností u proměnných MH u druhého modelu odpovídá součtu stejných četností u saturovaného modelu. U druhého modelu přitom nepředpokládáme existenci interakce TH (poměr šancí *OR* pro tuto interakci vypočítaný z modelových četností je číslo 1), stejně jako nepředpokládáme existenci trojrozměrné interakce (poměr kombinací poměrů šancí podle variant třetí proměnné odpovídá také číslu 1).

A/13 Výpočet modelových četností

Principy log-lineárního modelování jsou totožné s principy jakéhokoliv jiného statistického modelování hromadných dat. Když v realitě pozorujeme (měříme) data, součástí těchto dat jsou (obvykle) struktury – vazby mezi proměnnými, jež odhalujeme, abychom data mohli interpretovat. Ve statistickém modelování jsou však pouze ve výjimečných případech struktury hledány přímo v pozorovaných datech. Pokud bychom takto postupovali, vystavovali bychom se riziku, že vazeb mezi proměnnými (interakcí) najdeme nekonečně mnoho. Nebyli bychom pak schopni rozlišit, která vazba je pro interpretaci dat ještě zásadní a která už nikoliv.

Obvykle se proto postupuje naopak. Navrhne se model, který obsahuje strukturu vazeb mezi proměnnými (model je obvykle reprezentací testované hypotézy). Na základě tohoto modelu vypočítáme modelové četnosti (frekvence) a srovnáváme je s reálnými (pozorovanými) četnostmi. V případě, že odlišnost mezi nimi není statisticky významná, můžeme konstatovat, že navržené (modelové) vazby existují v datech. Pomocí nich pak data interpretujeme. V případě, že odlišnost mezi modelovými a pozorovanými četnostmi je statisticky významná, musíme navrhnout model s jinou strukturou vazeb mezi proměnnými. A opět testujeme, zda se vypočítané četnosti na základě tohoto modelu statisticky významně odlišují od pozorovaných četností.[81]

Výpočet modelových četností byl dlouhou dobu jedním z velkých problémů log-lineárního modelování, a dokud nebyly nalezeny adekvátní algoritmy, brzdil pokrok v tomto typu analýzy.

A/14 Generování modelových četností

Začněme příkladem jednoduchého statistického modelu, jímž je model nezávislosti mezi dvěma proměnnými. Na příkladě dat tabulky A.2 bychom testovali hypotézu, že věková homogamie (*H*) a sňatkový věk muže (*M*) nesouvisejí. Modelovou četnost F_{ij} v jednotlivých polích kontingenční tabulky vypočítáme jako součin modelové pravděpodobnosti p_{ij} a celkového počtu respondentů *N*:

$$F_{ij} = Np_{ij} \tag{30}$$

Modelovou pravděpodobnost p_{ij} neznáme, ale víme, že je výsledkem součinu dvou marginálních modelových pravděpodobností p_{i+} a p_{+j}

$$p_{ij} = p_{i+}p_{+j} \tag{31}$$

Marginální modelové pravděpodobnosti vypočítáme jako podíl marginálních pozorovaných četností a celkového počtu případů v tabulce (*N* či f_{++}):

$$p_{i+} = f_{i+} / f_{++} \tag{32}$$

$$p_{+j} = f_{+j} / f_{++} \tag{33}$$

81 Komparace modelových a pozorovaných četností platí především pro modelování kategorizovaných dat. V technikách vyvinutých pro spojité (kardinální) proměnné se v modelech komparují pozorované a modelové korelace.

Tabulka A.8 Pozorované četnosti a generované četnosti pro model nezávislosti (kurzívou) pro věkově homogamní a heterogamní sňatky podle sňatkového věku muže v ČR

Sňatkový věk muže	Věková homogamie 0–2 roky	Věková heterogamie 3+ let	Celkem
18–29	38 322	30 185	68 507
	32 689,41	*35 817,59*	*68 507*
30+	5 432	17 756	23 188
	11 064,59	*12 123,41*	*23 188*
Celkem	4 3754	47 941	91 695

Prostou kombinací rovnic 30, 31, 32 a 33 dostaneme rovnici pro výpočet četností modelu nezávislosti, který je také znám jako vzorec pro výpočet očekávaných četností v kontingenční tabulce pro identifikaci velikosti statistiky chí-kvadrát:

$$F_{ij} = f_{i+} f_{+j} / f_{++} \tag{34}$$

Podle tohoto vzorce modelové četnosti v jednotlivých polích kontingenční tabulky určují pouze marginální tabulkové distribuce. Existence vztahu mezi dvěma proměnnými je z rovnice eliminována. Jinými slovy řečeno, pomocí tohoto vzorce vypočítáme takové rozložení četností v tabulce, při němž mezi dvěma proměnnými neexistuje vztah. Tabulka A.8 ukazuje pozorované četnosti a četnosti pro model nezávislosti.

Pro generování četností složitějších modelů pro vícerozměrné kontingenční tabulky musíme použít speciální algoritmy. Používá se buď *algoritmus iterativního proporčního sednutí* (*iterative proportional fitting altgorithm*), někdy označovaný také jako Demingův a Stefanův algoritmus pro hierarchické modely nebo Newtonův-Raphsonův algoritmus.[82] Oba tyto algoritmy generují odhady maximální věrohodnosti (*maximum likelihood estimates – MLE*) modelových četností. Přitom podobně jako u přímého výpočtu četností modelu nezávislosti, i u těchto odhadů četností zůstávají marginální tabulkové distribuce totožné s pozorovanými marginálními distribucemi.

82 K popisu obou algoritmů viz Bishop, Fienberg, Holland (1975), Haberman (1978, 1979), Fienberg (1980).

Parametry log-lineárního modelu jsou interpretačně nosné pouze do té míry, do jaké odhadnutý model reprodukuje pozorovaná data. K poznání, který z odhadnutých modelů nejlépe reprodukuje pozorovaná data, se používá několik statistických kritérií. Těmi základními jsou Pearsonův test chí-kvadrát (χ^2) (rovnice 35 pro trojrozměrnou tabulku) a test poměru maximální věrohodnosti (L^2) (rovnice 36 pro trojrozměrnou tabulku). V obou těchto testech jsou srovnávány (i když odlišným způsobem) modelové (F) a pozorované (f) četnosti a v obou těchto testech nám jde o to, aby rozdíl mezi těmito četnostmi byl co nejmenší.[83] Výsledky obou testů jsou podobné (zvláště při malém počtu případů v datech). Test poměru maximální věrohodnosti (L^2) je však před testem Pearsonova chí-kvadrátu (χ^2) mnoha výzkumníky upřednostňován.

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(f_{ijk} - F_{ijk})^2}{F_{ijk}} \tag{35}$$

$$L^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} f_{ijk} \log\left(\frac{f_{ijk}}{F_{ijk}}\right) \tag{36}$$

Hodnoty L^2 mají přibližné chí-kvadrát rozdělení s příslušnými stupni volnosti (*df*). Pro log-lineární modely platí, že stupně volnosti označují počet vynechaných parametrů, které jsou nezbytné k identifikaci saturovaného modelu. Čím více parametrů z modelu vynecháme, tím více stupňů volnosti model má a tím je také úspornější. Když srovnáme hodnotu L^2 a počet stupňů volnosti, platí, že model data reprodukuje odpovídajícím způsobem, pokud jsou tyto hodnoty rovny nebo se liší jen nepatrně. Čím více jsou tyto hodnoty odlišné, tím více modelové četnosti nekorespondují s pozorovanými četnostmi. Na tuto skutečnost poukazuje statistická významnost u L^2, která odpovídá na otázku, zda se modelová data statisticky významně odlišují od dat pozorovaných.[84] V případě, že tomu tak není, model (navržené vazby mezi proměnnými) můžeme přijmout a data na jeho základě interpretovat. V případě, že se model statisticky významně odlišuje, musíme jej zamítnout a hledat jiný model (jiné vztahy mezi proměnnými).

83 U poměru maximální věrohodnosti (L^2) se jedná o nepodmíněný test, protože srovnáváme statistiku L^2 se saturovaným modelem ($L^2 = 0$).

84 V jakých polích kontingenční tabulky model neadekvátně reprodukuje data, ukazují rezidua (rozdíly mezi pozorovanými a modelovými četnostmi). Jejich standardizovaná podoba (pro účely srovnání) je počítána podle vzorce: $R_s = (f - F)/\sqrt{F}$. V případě, že model data reprodukuje adekvátně, jsou rezidua v podobné míře kladná i záporná, mají přibližně stejnou velikost, a to napříč všemi poli kontingenční tabulky.

V analýze dat jsou rozšířené testy statistické významnosti koeficientů a statistické významnosti modelů. V případě koeficientů pomocí testu zkoumáme, zda se daný koeficient liší od nuly natolik, že to nemůže být náhoda, a proto jej očekáváme i v základní populaci (ovšem s určitou pravděpodobností). V případě statistických modelů provádíme obvykle dva testy statistické významnosti. Buď zkoumáme, nakolik struktura (námi navržených) modelových vztahů odpovídá (zase s určitou pravděpodobností) struktuře pozorovaných dat (test chí-kvadrát a test poměru maximální věrohodnosti), nebo zkoumáme, jestli se struktura složitějšího modelu liší od struktury jednodušího („vsazeného do složitějšího") modelu (s ohledem na stupně volnosti) natolik, že proměnná, která v (jednodušším) modelu chybí, je pro interpretaci dat nezbytná.[85] Přitom musíme mít na paměti, že se jedná o testy statistické, nikoliv reálné odlišnosti. Jedná se o nalezení hranice, kdy jsou dva statistické modely natolik různé, že můžeme z hlediska interpretace jeden považovat za významnější než jiný. Testy se používají obvykle tehdy, kdy více modelů uspokojivě reprodukuje data.

Rovnice 37 a 38 ukazují princip tohoto testu. Poměr maximální věrohodnosti u jednodušího modelu (o – omezený model) je srovnáván s poměrem maximální věrohodnosti u složitějšího modelu (n – neomezený model). Výsledkem je podmíněný poměr maximální věrohodnosti $L^2_{o|n}$, který s ohledem na rozdíl v počtech stupňů volnosti (počet parametrů, jimiž se modely liší) odpovídá na otázku, zda nepřítomnost parametrů v omezeném modelu je statisticky významná – tedy zda se omezený model statisticky významně odlišuje od modelu neomezeného.[86] V případě, že nikoliv, přijmeme jednoduší model a konstatujeme, že přítomnost parametrů, které v omezeném modelu chybějí, není pro interpretaci dat nezbytná (úspornější model není statisticky horší než model složitější). V případě, že statistickou významnost odlišnosti mezi modely nalezneme, chybějící parametry v jednodušším modelu musíme považovat za statisticky významné a pro interpretaci dat nezbytné (úspornější model je statisticky horší než model složitější).

85 Máme například tři modely, z nichž model 3 je saturovaný model pro dvojrozměrnou tabulku, model 2 je modelem nezávislosti proměnných pro stejnou tabulku (dvojrozměrný parametr v něm chybí) a model 1 obsahuje pouze parametr pro proměnnou v řádku tabulky. Model 2 je ve srovnání s modelem 3 omezený a říkáme, že je v něm z hlediska hierarchie parametrů „vsazen", model 1 je zase omezený ve srovnání s modelem 2 a je v něm také z hlediska hierarchie parametrů „vsazen".

86 Jedná se o podmíněný test, jelikož srovnáme statistiky L^2 u dvou nesaturovaných modelů.

$$L^2_{o|n} = L^2_o - L^2_n \tag{37}$$

$$df_{o|n} = df_o - df_n \tag{38}$$

Jiný a v současnosti velmi rozšířený přístup k výběru modelu je založen na informačních kritériích. Tato kritéria (BIC, AIC) odkazují ke zkoumané realitě. V případě koeficientu identifikují míru informace, kterou o realitě daný koeficient přináší. V případě statistického modelu odkazují k velikosti informace, kterou daný model o zkoumané realitě poskytuje (Raftery 1986, 1995). Čím „bohatší" informaci model poskytuje, tím je také pro interpretaci výsledků vhodnější. Rovnice pro výpočet statisticky BIC a AIC pro log-lineární modely jsou následující:

$$\text{BIC} = L^2 - \log N df \tag{39}$$

$$\text{AIC} = L^2 - 2df \tag{40}$$

V reálných aplikacích se obvykle bere zřetel jak na testy významnosti, tak na informační kritéria. V případě statistických modelů se hledá model, který se statisticky významně neliší od dat a má zápornou hodnotu informačních kritérií. V případě velkých vzorků však testy statistické významnosti selhávají, protože nelze najít model, který není svou strukturou statisticky nevýznamně odlišný od struktury pozorovaných dat. V takových případech se při výběru modelu spoléháme na informační kritéria. Data interpretujeme na základě modelu, který má nejnižší hodnotu informačních kritérií a o takovém modelu hovoříme jako o modelu, který nám o zkoumané realitě přináší nejvíce informací. V případě, že je statistika BIC pro všechny modely kladné číslo, nezbývá nám než pro data přijmout saturovaný model a konstatovat, že úspornější model se nepodařilo nalézt (srov. Powers, Xie 2000).

A/17 Asociativní modely

Parametry log-lineárních modelů mohou být omezené více způsoby, než že jsou pouze vynechány (jejich hodnota je 0 v aditivní rovnici nebo 1 v multiplikativní rovnici). Mohou být specifikovány jednak tak, že se jejich hodnoty rovnají, nebo tak, že jeden parametr odpovídá násobku jiného parametru. Pokud jsou varianty proměnné ordinální (lze je seřadit), máme o proměnné navíc informaci, kterou postrádáme, pokud se jedná o nominální proměnnou (její varianty lze pouze pojmenovat). V takovém případě můžeme předpokládat, že vzdálenosti mezi variantami ordinální proměnné jsou ekvidistantní.

Přiřadíme-li těmto variantám číselné hodnoty, aby vzdálenost mezi nimi byla stejná,[87] vztah mezi nimi můžeme modelovat pomocí jednoho parametru.

S lineární specifikací parametrů pracují asociativní log-lineární modely (Goodman 1978; Clogg, Shihadeh 1994). V případě, že nás zajímá vztah mezi dvěma ordinálními proměnnými, lze tento vztah modelovat lineárně pomocí jednoho parametru (jedná se o model lineární interakce). Tento model se také někdy nazývá jako model uniformní asociace (unidiff model či U-model), protože asociace mezi jednotlivými variantami proměnných je modelována na základě jednoho parametru (β), a nikoliv pomocí sady (*I*-1) a (*J*-1) nezávislých parametrů (τ nebo λ). Parametr (β) je lineární pro skóry variant řádkové proměnné v jednotlivých variantách sloupcové proměnné a lineární pro skóry variant sloupcové proměnné v jednotlivých variantách řádkové proměnné. Specifikace takového modelu pro tabulku 5 × 5 mohou vypadat následovně:

	-2	-1	0	1	2
-2	4β	2β	0β	-2β	-4β
-1	2β	1β	0β	-1β	-2β
0	0β	0β	0β	0β	0β
1	-2β	-1β	0β	1β	2β
2	-4β	-2β	0β	2β	4β

	1	2	3	4	5
1	1β	2β	3β	4β	5β
2	2β	4β	6β	8β	10β
3	3β	6β	9β	12β	15β
4	4β	8β	12β	16β	20β
5	5β	10β	15β	20β	25β

Předpokládejme, že máme dvě ordinální proměnné: *H* a *M*. Saturovaný log-lineární model pro četnosti těchto dvou proměnných má následující podobu:

$$F_{ij}^{HM} = \eta\tau_i^H\tau_j^M\tau_{ij}^{HM} \qquad G_{ij}^{HM} = \theta + \lambda_i^H + \lambda_j^M + \lambda_{ij}^{HM} \tag{41}$$

Nahradíme-li dvojrozměrnou interakci parametrem ijβ, kde *i* a *j* označují číselné hodnoty variant řádkové a sloupcové proměnné, dostaneme rovnici modelu uniformní asociace:

$$F_{ij}^{HM} = \eta\tau_i^H\tau_j^M e^{ij\beta} \qquad G_{ij}^{HM} = \theta + \lambda_i^H + \lambda_j^M + ij\beta \tag{42}$$

Asociaci v kontingenční tabulce charakterizuje pouze jeden parametr β, jehož velikost je pro jednotlivé kombinace řádků a sloupců uniformní (stejná) a hodnoty všech nezbytných poměrů šancí jsou totožné (OR = expβ, nebo OR vypočítáme z modelových četností).

U proměnných v řádcích a sloupcích kontingenční tabulky můžeme také linearitu předpokládat jednotlivě. V případě, že takto specifikujeme pouze sloupcovou proměnnou, pro řádkovou proměnnou předpokládáme nominální kategorie, dostaneme řádkovou strukturu asociace. Hovoříme pak o modelu řádkové asociace – *R* model (z anglického *Row model*). To znamená, že pro každý řádek máme sadu parametrů (μ_i – tzv. řádkové skóry), které ukazují lineární vztahy mezi jednotlivými variantami řádkové proměnné a skóry sloupcové proměnné. V případě, že specifikujeme lineárně řádky tabulky a pro sloupce předpokládáme nominální kategorie, platí totéž, ale pro řádky a sloupce převráceně. Jedná se o model sloupcové asociace – *C* model (z anglického *Column model*). V takovém případě interpretujeme parametry mezi jednotlivými variantami sloupcové proměnné a skóry řádkové proměnné. V případě, že předpokládáme linearitu u řádkové i sloupcové proměnné dohromady, dostaneme model řádkové a sloupcové asociace (*R* + *C* model, někdy také jako model *RC I*). Předpokladem tohoto modelu je ordinalita variant proměnných a jejich uspořádání před odhadem modelu (změníme-li uspořádání kategorií, změníme také hodnoty odhadnutých parametrů). Pro varianty obou proměnných dostaneme sadu rozdílných parametrů (μ_i a μ_j). Odhadované parametry lze následně omezit tak, aby byly odhadnuty jako totožné pro obě proměnné, což je úspornější řešení. Model řádkové asociace je zapsán v rovnici 43, model sloupcové asociace v rovnici 44 a model řádkové a sloupcové asociace ukazuje rovnice 45.

$$F_{ij}^{HM} = \eta\tau_i^H\tau_j^M e^{j\mu_i} \qquad G_{ij}^{HM} = \theta + \lambda_i^H + \lambda_j^M + j\mu_i \tag{43}$$

$$F_{ij}^{HM} = \eta\tau_i^H\tau_j^M e^{i\mu_j} \qquad G_{ij}^{HM} = \theta + \lambda_i^H + \lambda_j^M + i\mu_j \tag{44}$$

$$F_{ij}^{HM} = \eta\tau_i^H\tau_j^M e^{j\mu_i, i\mu_j} \qquad G_{ij}^{HM} = \theta + \lambda_i^H + \lambda_j^M + j\mu_i + i\mu_j \tag{45}$$

Log-multiplikativní asociativní model, navržený Leo Goodmanem (1978) nebo Cliffordem Cloggem (1982), se od předchozích asociativních modelů (*U*, *R*, *C* a *R*+*C*) liší v tom, že skóry pro řádky nebo sloupce tabulky či řádky a sloupce tabulky dohromady nejsou číselně specifikovány před odhadem modelu, ale jejich hodnoty jsou odhadnuty. To znamená, že vzdálenosti mezi uspořádanými kategoriemi nejsou předpokladem, ale výsledkem modelu. Tento model používáme tehdy, nejsme-li si jisti, že uspořádání kategorií proměnných je správné, nebo tehdy, když je naším cílem identifikace vzdáleností mezi kategoriemi proměnných. Jediným předpokladem tohoto modelu je

87 Jednotlivé statistické programy přiřazují variantám proměnných jiné hodnoty, což je nezbytné brát v potaz při výpočtu modelových četností na základě odhadnutých parametrů. Například v LEMu jsou hodnoty pro lichý počet pěti variant specifikovány jako -2 -1 0 1 2, pro sudý počet šesti variant jako -2,5 -1,5 -0,5 0,5 1,5 2,5. Ať použijeme tuto nebo odlišnou specifikaci (pro lichý počet například 1 2 3 4 5, pro sudý počet 1 2 3 4 5 6), velikost odhadnutých parametrů se nemění.

ordinalita kategorií proměnných. Skóry pro řádky a sloupce tabulky jsou neznámé parametry μ_i a μ_j a jsou odhadovány dohromady s parametrem β, který indikuje tabulkovou asociaci. Z tohoto důvodu se tento model nazývá log-multiplikativní (neboli *RC* model, někdy také model *RC II*). Rovnice pro tento model je následující:

$$F_{ij}^{HM} = \eta\tau_i^H\tau_j^M e^{\mu_i\mu_j\beta} \qquad G_{ij}^{HM} = \theta + \lambda_i^H + \lambda_j^M + \mu_i\mu_j\beta \tag{46}$$

A/18 Model log-multiplikativního mezitabulkového efektu

O více než desetiletí později Y. Xie (1992) nebo R. Erikson a J. H. Goldthorpe (1992) rozšířili log-multiplikativní princip na mezitabulkovou asociaci (trojrozměrná a vyšší interakce). Nezávisle na sobě navrhují model, v němž jsou odhadnuty parametry pro dvojrozměrnou (tabulkovou) asociaci, přitom je ale pro každou variantu třetí proměnné odhadnuta také multiplikativní odchylka od této dvojrozměrné asociace. Z hlediska interpretace tato odchylka ukazuje, jak se mění dvojrozměrná asociace podle variant třetí proměnné.

Předpokládejme, že modelujeme vztah mezi věkovou homogamií (*H*) a sňatkovým věkem muže (*M*) v jednotlivých letech (*R*). Log-lineární (aditivní) rovnice pro saturovaný model vypadá následovně:

$$G_{ijk}^{HMR} = \theta + \lambda_i^H + \lambda_j^M + \lambda_k^R + \lambda_{ki}^{RH} + \lambda_{kj}^{RM} + \lambda_{ij}^{HM} + \lambda_{ijk}^{HMR} \tag{47}$$

Chceme-li odhadnout model log-multiplikativního mezitabulkového efektu pro tato data, musíme součet parametrů $\lambda_{ij}^{HM} + \lambda_{ijk}^{HMR}$ v této rovnici nahradit součinem parametrů $\psi_{ij}\phi_c$. Parametr ψ_{ij} ukazuje asociaci mezi jednotlivými variantami věkové homogamie a sňatkového věku muže (bez ohledu na roky), parametr ϕ_c ukazuje násobek této asociace neboli její velikost pro jednotlivé roky. Rovnice modelu pak vypadá následovně:

$$G_{ijk}^{HMR} = \theta + \lambda_i^H + \lambda_j^M + \lambda_k^R + \lambda_{ki}^{RH} + \lambda_{kj}^{RM} + \psi_{ij}\phi_c \tag{48}$$

Model se nazývá log-multiplikativní, protože log-lineární rovnice obsahuje multiplikaci dvou parametrů. Jeho předpokladem je, že všechny tabulkové poměry šancí se mění stejným směrem (podle variant třetí proměnné). Z tohoto důvodu je změna v asociaci modelována pouze pomocí jednoho parametru. Díky této charakteristice je tento model v sociálněstratifikačním výzkumu nazýván jako model uniformní diference neboli *unidiff model* (Erikson, Goldthorpe 1992). Pro identifikaci vývoje či změny asociace podle variant třetí proměnné se jedná o velmi vhodný model. Problém spočívá v tom, že na jeho základě nejsme schopni popsat změnu, k níž v poměrech šancí (tabulkové asociaci) podle variant třetí proměnné dochází. Řešení tohoto problému nabízí až model navržený o šest let později Leo Goodmanem a Mikem Houtem (1998, 2001).

Oba badatelé vyšli z předpokladu, že model uniformní diference je příliš restriktivní. Z hlediska úspornosti je to nesporně výhoda, z hlediska popsání změny v tabulkové asociaci se však jedná o značnou nevýhodu. Navrhují proto model, který je dnes znám jako Goodman-Hout model nebo jako model regresního mezitabulkového efektu. V jeho rámci můžeme modelovat jak proměnu poměrů šancí (změnu vzorce tabulkové asociace), tak vývoj velikosti této asociace (trend v asociaci). Vyjdeme-li ze saturovaného modelu v rovnici 47, model regresního mezitabulkového efektu dostaneme tak, že součet parametrů $\lambda_{ij}^{HM} + \lambda_{ijk}^{HMR}$ nahradíme součtem a součinem parametrů $\lambda_{ij}^{HM} + \psi_{ij}\phi_c$. Parametr λ_{ij}^{HM} ukazuje základní vzorec tabulkové asociace, ψ_{ij} ukazuje části asociace, které se mění podle třetí proměnné – v letech – a parametr ϕ_c ukazuje velikost změny asociace pro jednotlivé roky. Log-lineární rovnice takového modelu je pak následující:

$$G_{ijk}^{HMR} = \theta + \lambda_i^H + \lambda_j^M + \lambda_k^R + \lambda_{ki}^{RH} + \lambda_{kj}^{RM} + \lambda_{ij}^{HM} + \psi_{ij}\phi_c \tag{49}$$

Pomocí tohoto modelu dokážeme identifikovat jak změny ve struktuře asociace (poměrech šancí), tak velikost změny asociace v jednotlivých variantách třetí proměnné. Jedná se zatím o poslední a velmi významný posun na poli log-lineárních modelů. Za jistou nevýhodu tohoto modelu lze považovat to, že zatím nebyl uspokojivě aplikován na data, která obsahují více než tři rozměry (na čtyřrozměrné a vícerozměrné tabulky).

A/19 Podoba dat pro log-lineární analýzu

Data pro statistickou analýzu mají buď individuální, nebo agregovanou podobu. V případě, že pracujeme s individuálními daty, analyzujeme matici, v níž je (v jednotlivých polích) zapsaná pozorovaná (měřená) varianta proměnné (bývá obvykle ve sloupcích matice) pro jednotlivé případy (obvykle bývají v řádcích matice). V log-lineární analýze s tímto typem dat nepracujeme. Pokud bychom měli individuální data a chtěli bychom je analyzovat pomocí log-lineárních modelů, bylo by nezbytné je převést na data agregovaná.[88]

88 Jiným řešením je použít logistickou regresi, kterou lze aplikovat jak na individuální, tak agregovaná data, přičemž hodnoty koeficientů logitových a log-lineárních modelů, které se neliší svou strukturou a jsou aplikovány na stejná data, jsou totožné.

Tabulka A.9 Data z tabulky A.1 ve formě četnostních záznamů pro kombinace variant analyzovaných proměnných

Roky	Typ věkového sňatku	Sňatkový věk muže	Věková homogamie a heterogamie	Četnost
1	1	1	1	18 554
1	1	1	2	11 728
1	1	1	3	4 655
1	1	2	1	1 109
1	1	2	2	1 580
1	1	2	3	4 469
1	2	1	1	4 294
1	2	1	2	1 666
1	2	1	3	846
1	2	2	1	361
1	2	2	2	276
1	2	2	3	115
2	1	1	1	11 408
2	1	1	2	6 347
2	1	1	3	1 819
2	1	2	1	3 191
2	1	2	2	4 574
2	1	2	3	6 079
2	2	1	1	4 066
2	2	1	2	2 106
2	2	1	3	1 018
2	2	2	1	771
2	2	2	2	516
2	2	2	3	147

Agregovaná data, prezentovaná obvykle ve formě kontingenčních tabulek, ukazují počet opakujících se pozorování pro jednotlivé kombinace variant proměnných. V tomto případě se nejedná o nic jiného než o přepis (jakkoliv mnohorozměrné) kontingenční tabulky podle variant jednotlivých proměnných do řádků a sloupců matice.

V tabulce A.1 máme agregovaná data, která ukazují počet věkově homogamních a heterogamních sňatků podle sňatkového věku muže a typu věkového sňatku v letech 1994 a 2004 v České republice. Tato data můžeme zapsat také v podobě četností pro jednotlivé kombinace tabulkových proměnných. Tabulka A.9 ukazuje tento zápis (názvy variant jednotlivých proměnných jsou nahrazeny čísly). Jedná se o vymezení všech možných případů z hlediska variant jednotlivých proměnných. Každá četnost ukazuje, kolikrát se daná kombinace variant v datech vyskytuje. Tato data analyzujeme naprosto stejným způsobem jako data individuální, pouze kombinacím jednotlivých proměnných přiřadíme (jim odpovídající) četnosti jako váhy. V log-lineárním modelování pracujeme buď s tímto zápisem dat, nebo s daty v podobě kontingenční tabulky (věcně se jedná o jedno a totéž).

Agregovaná data lze jednoduše převést na individuální tak, že do každého řádku matice (v němž předpokládáme případy) vepíšeme odpovídající počty kombinací jednotlivých variant proměnných. V našem případě víme, že kombinace variant 1 1 1 1 se vyskytuje 18 554 (tabulka A.9). Je nezbytné tedy vepsat 18 554 řádků s hodnotou 1 u každé proměnné. Podobně pak zapíšeme počet řádků daných četnostmi pro všechny zbylé kombinace variant proměnných. Celkový počet řádků v matici pak odpovídá celkovému počtu případů v kontingenční tabulce. V případě tabulky A.9 by to bylo 91 695.