UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Trojrozměrné kontingenční tabulky

Vedoucí diplomové práce:
Doc. RNDr. Eva Fišerová, Ph.D.
Rok odevzdání: 2012

Vypracovala:
Bc. Dana Cahová
MAP, II. ročník

Prohlášení

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně za vedení Doc. RNDr. Evy Fišerové, Ph.D. a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 26. března 2012

Poděkování

Ráda bych na tomto místě poděkovala vedoucí diplomové práce Doc. RNDr. Evě Fišerové, Ph.D. za obětavou spolupráci i za čas, který mi věnovala při konzultacích, a rodičům za podporu při studiu.

Obsah

Úvodní slovo

Při statistické analýze se setkáváme s daty povahy buď kvantitativní nebo kvalitativní. Kvalitativní data se často nazývají kategoriální nebo také diskrétní. Jde o data, která lze rozdělit do skupin podle určitých vlastností. Například rozlišení osob dle pohlaví, vzdělání, místa bydliště nebo podle barvy očí. Kategoriální data měříme většinou na škále nominální, kdy nelze data porovnat vůči sobě (např. pohlaví) nebo ordinální, kdy lze data odstupňovat, srovnat (např. vzdělání). V menší míře na kardinální škále, kterou však většinou dále rozdělujeme do intervalů (např. věk). Jednou z metod pro analyzování kategoriálních dat jsou kontingenční tabulky.[1] Vztahy mezi dvěma kategoriálními znaky se posuzují dvourozměrnými kontingenčními tabulkami. Chceme-li však zkoumat více kategoriálních proměnných musíme do vyšších dimenzí.

Čtenáře znalého statistických metod by mohlo napadnout, že analyzování trojrozměrných kontingenčních tabulek je poněkud zbytečné. Pro analyzování několika znaků existují známější metody, jako logistická regrese nebo zobecněné lineární modely, případně lze analyzovat proměnné po dvou. Každá trojrozměrná tabulka lze totiž rozložit na několik dvourozměrných tabulek například podle jednotlivých řádků a interpretovat přímo tyto dvourozměrné tabulky. Tímto rozkladem však přijdeme o vzájemné vazby mezi proměnnými (dvou proměnných natřetí) a následná interpretace by mohla vést k zavádějícím výsledkům. Rozklady na podtabulky se doporučují provádět až v případě, že víme, které proměnné jsou na kterých nezávislé. Pak je možné spočítat i příslušné šance a jejich poměry. Nejlepším nástrojem k určení vztahů mezi proměnnými se jeví loglineární modely, které si v této práci podrobně vysvětlíme. Loglineární modely mají oproti ostatním modelům bezespornou výhodu, kterou je aplikovatelnost na data bez znalosti jejich rozdělení a v případě, že nemáme představu, které proměnné jsou vysvětlované a které vysvětlující. Právě z důvodu nejednoznačnosti vysvět-

[1] Kontingence (z latinského contingit, stává se, přihodilo se) je nahodilost, vlastnost jevů, vztahů a věcí, které mohou, ale nemusí být, a v důsledku toho také vznikají, mění se a zanikají. Takovým skutečnostem říkáme kontingentní.[47]

lujících a vysvětlovaných proměnných nelze použít zmíněná logistická regrese ani zobecněné lineární modely.

V této práci se dočteme konkrétně o trojrozměrných kontingenčních tabulkách. Na dalších stránkách najdeme ucelenou teorii potřebnou k analyzování trojrozměrných kontingenčních tabulek.

V první kapitole je zavedeno značení a popis trojrozměrných kontingenčních tabulek. Ukážeme si také, jak z povahy dat poznáme, který model pro analýzu zvolit.

Ve druhé kapitole si ukážeme odvození loglineárního modelu, popíšeme si typy nezávislosti pro tři kategoriální znaky a dozvíme se o hierarchických modelech, které si blíže popíšeme.

V dalších dvou kapitolách se budeme věnovat odhadům očekávaných četností v kontingenční tabulce. Ukážeme si dva různé přístupy k hledání odhadů, metodu maximální věrohodnosti a metodu nejmenších čtverců. První přístup je velice známý a běžně využívaný. Druhý je víceméně alternativní a v praxi těžce aplikovatelný především kvůli značným omezením na znalost celkové populace. Určitě je ale vhodné tento přístup zmínit, neboť pak by čtenář mohl dojít ke klamné představě, že jedinou metodou, jak odhadnout parametry v loglineárním modelu, resp. očekávané četnosti, je metoda maximální věrohodnosti. Na konci každé z kapitol jsou uvedeny algoritmy pro výpočet očekávaných četností ve vybraném modelu. V příloze diplomové práce jsou pak algoritmy „ručně“ naprogramovány softwarem MATLAB, čímž je ucelena teorie loglineárních modelů.

Poslední kapitolou věnovanou loglineárním modelům je kapitola o posuzovaní těchto modelů a výběru „nejlepšího“. Najdeme zde několik kritérií a statistik k vhodnému porovnávání modelů mezi sebou.

Další samostatná kapitola je nazvaná Poměry a šance, kde se dozvíme, jak spočítat šance na určitou událost. Z dvourozměrného případu, tabulky 2×2, se dostaneme až k trojrozměrným tabulkám obecného počtu řádků, sloupců a hladin. Zjistíme také, že pomocí šancí lze názorně interpretovat parametry loglineárních modelů, což si ukážeme v závěru kapitoly.

Poslední a neméně významnou kapitolou následující bezprostředně po teoretické části je část čistě praktická. Zde jsou uvedeny výsledky z dotazníku, který byl v rámci této diplomové práce vytvořen. Výsledky jsou zpracovány především pomocí statistického softwaru SAS a jsou srovnány s výstupy dalších dvou programů (R a Statistica od StatSoft).

1 Trojrozměrné kontingenční tabulky

1.1 Popis trojrozměrné kontingenční tabulky

V této kapitole si popíšeme, jak vzniká trojrozměrná kontingenční tabulka, a ukážeme si značení, které budeme používat v dalších kapitolách.

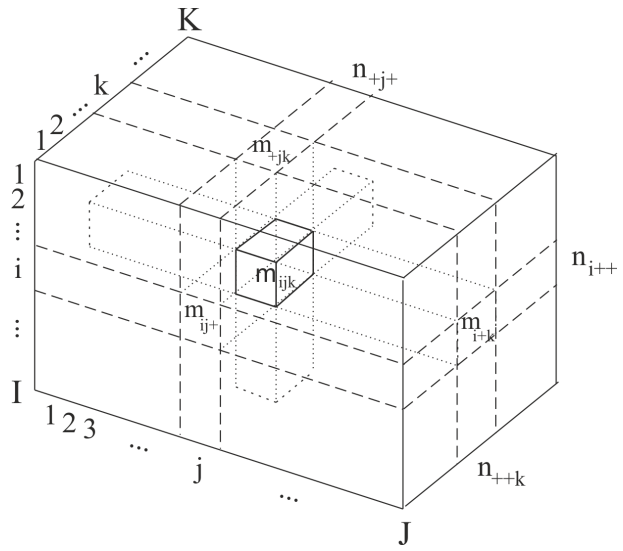
Kontingenční tabulka je nástrojem pro klasifikaci objektů podle kategoriálních znaků. V případě tří kategoriálních znaků přirozeně získáme trojrozměrnou kontingenční tabulku. Pro následný text uvažujme tři kategoriální znaky X, Y, Z.

Řádky v tabulce označme indexem i a bude je reprezentovat znak X. Znak X má I variant, tj. řádkový index i probíhá $1, 2, \ldots, I$. Sloupce označme indexem j. Sloupce bude reprezentovat znak Y s J variantami, tj. sloupcový index j tedy probíhá $1, 2, \ldots, J$. Třetí rozměr v tabulce nazveme hladiny, označme je indexem k a odpovídá jim znak Z. Hladinový index k probíhá varianty $1, 2, \ldots, K$. Hladin v tabulce bude K.

Každá trojice (X_i,Y_j,Z_k) určuje jednotku z populace rozsahu n a všechny tyto trojice lze seskupit do jedné trojrozměrné kontingenční tabulky, jejíž dimenze je $I \times J \times K$. Vektor (X_i,Y_j,Z_k) budeme pro jednoduchost značit Y_{ijk}. Označme n_{ijk} počet jednotek, které mají $i-$tou variantu znaku X, $j-$tou variantu znaku Y a $k-$tou variantu znaku Z. Počet n_{ijk} nazveme *pozorovanou četností* nebo pouze četností v buňce $\{i, j, k\}$. Na četnost n_{ijk} se lze dívat i jako na počet realizací znaku Y_{ijk}. Dále již budeme Y_{ijk} považovat za jednorozměrnou veličinu.

Výše uvedné si shrňme. Kontingenční tabulku tvoří tři kategoriální znaky X, Y, Z. Jednotka v tabulce na $ijk-$té pozici je tedy realizace trojrozměrného vektoru (X_i, Y_j, Z_k). Počet jednotek, které mají stejnou realizaci vektoru (X_i, Y_j, Z_k) bude již ale jednorozměrná náhodná veličina Y_{ijk}. Kontingenční tabulku lze zapsat jako náhodný vektor $(Y_{111}, Y_{112}, \ldots, Y_{IJK})$ jehož realizací je vektor četností $(n_{111}, n_{112}, \ldots, n_{IJK})$. [39]

Dále označme π_{ijk} pravděpodobnost, že se vybraná jednotka z populace ocitne



Obrázek 1: Intuitivní představa trojrozměrné kontingenční tabulky

v buňce $\{i, j, k\}$, a nazvěme ji *očekávaná (teoretická) pravděpodobnost*. Očekávané pravděpodobnosti tvoří vektor $(\pi_{111}, \pi_{112}, \ldots, \pi_{IJK})$. [41] *Očekávanou četnost* budeme značit m_{ijk}, což lze opět zapsat pomocí vektoru $(m_{111}, m_{112}, \ldots, m_{IJK})$.

Pro marginální četnosti budeme uvažovat klasicky zavedené sumační značení. Sčítání přes řádky $(i = 1, \ldots, I)$ budeme značit $n_{+jk} = \sum_i n_{ijk}$. Jedná se o tzv. řádkovou marginální četnost. Analogicky sčítání přes sloupce $(j = 1, \ldots, J)$ $n_{i+k} = \sum_j n_{ijk}$ a hladiny $(k = 1, \ldots, K)$ $n_{ij+} = \sum_k n_{ijk}$. Sčítání řádků a sloupců v $k-$té hladině pak označme $n_{++k} = \sum_i \sum_j n_{ijk}$. Analogicky sumace v $i-$tém řádku nebo $j-$tém sloupci. Sečtením všech pozorovaných četností v tabulce získáme celkový rozsah populace n, který lze značit opět sumačně $n_{+++} = \sum_i \sum_j \sum_k n_{ijk}$. Pro lepší představu nám může posloužit tabulka 1 nebo obrázek 1. Stejný styl značení aplikujeme jak na očekávané pravděpodobnosti π_{ijk}, tak i na očekávané četnosti m_{ijk}. Přirozenou podmínkou pro součet pravděpodobností bude $\sum_i \sum_j \sum_k \pi_{ijk}$=1. Označením π_{i++} budeme tedy rozumět součet všech pravděpodobností v $i-$tém řádku a bude to pravděpodobnost, že se objekt ocitne v $i-$tém řádku. Analogicky pro ostatní marginální pravděpodobnosti.

		Y_1	Y_2	Y_3	Y_4	n_{i+k}
Z_1	X_1	n_{111}	n_{121}	n_{131}	n_{141}	n_{1+1}
	X_2	n_{211}	n_{221}	n_{231}	n_{241}	n_{2+1}
Z_2	X_1	n_{112}	n_{122}	n_{132}	n_{142}	n_{1+2}
	X_2	n_{212}	n_{222}	n_{232}	n_{242}	n_{2+2}
Z_3	X_1	n_{113}	n_{123}	n_{133}	n_{143}	n_{1+3}
	X_2	n_{213}	n_{223}	n_{233}	n_{243}	n_{2+3}
	n_{++k}	n_{+1+}	n_{+2+}	n_{+3+}	n_{+4+}	n

Tabulka 1: Ukázka rozmístění znaků a četností

1.2 Volba vhodného modelu dle typu dat

Nejdříve si připomeneme některá diskrétní rozdělení pravděpodobnosti náhodných veličin, která budeme dále potřebovat. Poté se budeme věnovat volbě vhodného modelu podle typu dat. Na závěr této kapitoly si ukážeme ekvivalenci multinomického a Poissonova rozdělení.

Binomické rozdělení

[44] Uvažujeme n statisticky nezávislých pokusů. V každém pokusu může sledovaný jev buď nastat (= „úspěch“) nebo nenastat (= „neúspěch“). Odpovídající pravděpodobnosti označíme π a $1-\pi$ a jsou v každém pokusu stejné. Celkový počet úspěchů X v n nezávislých pokusech je binomická veličina. Tato náhodná veličina může nabývat pouze celočíselných hodnot od 0 do n. Je-li v každém pokusu pravděpodobnost úspěchu π, potom pravděpodobnost, že v n nezávislých pokusech nastane přesně k úspěchů, je

$$P(X=k) = \binom{n}{k} \pi^k (1-\pi)^{n-k} \qquad k = 0, 1, 2, \ldots, n.$$

Situaci, kdy má náhodná velična binomické rozdělení s parametry n a p označíme $X \sim Bi(n, \pi)$.

Střední hodnota binomického rozdělení je $E(X) = n\pi$ a rozptyl je $D(X) = n\pi(1-\pi)$.

Poissonovo rozdělení

[44] Uvažujme náhodnou veličinu X, která představuje počet výskytů nějaké výjimečné události v daném intervalu. Veličina X tedy může nabývat celočíselných hodnot od 0 do nekonečna. Nechť λ je kladná konstanta označující průměrný počet událostí v intervalu (času nebo prostoru). Potom pravděpodobnost řídkého jevu, tj. výskytu méně časté události je dána vztahem

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

a X je náhodná veličina s Poissonovým rozdělením s parametrem λ, což budeme značit $X \sim Po(\lambda)$. Toto diskrétní rozdělení vznikne buď jako limitní případ binomického rozdělení nebo tehdy, když události nějakého druhu nastávají náhodně s malou pravděpodobností v čase či prostoru. Je-li pravděpodobnost π nějaké výjimečné události relativně malá a rozsah výběru poměrně velký, pak Poissonovo rozdělení v podstatě splývá s binomickým, ale je mnohem výhodnější pro počítání. Střední hodnota Poissonova rozdělení je $E(X) = \lambda$ a rozptyl je $D(X) = \lambda$.

Multinomické rozdělení

[44] Uvažujme n nezávislých pokusů, z nichž každý musí skončit právě jedním z k možných výsledků. Vedle nezávislosti jednotlivých pokusů dále předpokládáme, že pravděpodobnost výsledku A_j je rovna číslu π_j bez ohledu na pořadí pokusu, což musí platit pro všechna j, $j = 1, ..., k$. Pravděpodobnosti π_j jsou nezáporné a jejich součet je roven jedničce. Zajímá nás počet pokusů n_1, v nichž nastal výsledek A_1, počet pokusů n_2, v nichž nastal výsledek A_2,..., počet pokusů n_k, v nichž nastal výsledek A_k. Jsou-li $n_1, \ldots, n_k$ nezáporná celá čísla splňující $\sum_j n_j = n$, pak budeme četnosti $n_1, \ldots, n_k$ v multinomickém rozdělení očekávat s pravděpodobností

$$P(n_1, \ldots, n_k) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \ldots \pi_k^{n_k}.$$

Označíme-li náhodný počet pokusů s výsledkem A_j symbolem X_j, pak náhodný vektor $\mathbb{X} = (X_1, \ldots, X_k)$ má multinomické rozdělení s parametry $n, \pi_1, \ldots, \pi_k$,

což značíme $\mathbb{X} \sim M(n, \pi_1, \ldots, \pi_k)$.

Náhodné veličiny X_j mají binomické rozdělení se střední hodnotou $E(X_j) = n\pi_j \quad , j = 1, \ldots, k$ a rozptylem $D(X_j) = n\pi_j(1 - \pi_j) \quad j = 1, \ldots, k$. Kovariance mezi $i-$tou a $j-$tou veličinou je $cov(X_i, X_j) = -\pi_i\pi_j \quad i, j = 1, \ldots, k$, přičemž $i \neq j$. Střední hodnota vektoru $\mathbb{X}$ je vektor středních hodnot veličin X_j a variační matice vektoru $\mathbb{X}$ je sestavená z kovariancí náhodných veličin X_i a X_j, přičemž na diagonále se nachází rozptyly veličin X_j.

Multinomické rozdělení je zobecněním rozdělení binomického. Používáme je tehdy, když máme určit pravděpodobnost výskytu daného počtu ne jednoho jevu a jevu k němu opačného, ale obecně více (konečně mnoha) různých jevů.

Nyní se již věnujme otázce volby modelu. Podle povahy získaných dat volíme model, který nám bude trojrozměrnou kontingenční tabulku vhodně reprezentovat. Vybíráme-li jednotky z populace tak, že dopředu nevíme kolik jich budeme mít, získáme tzv. *Poissonův model*. Objekty pozorování jsou nezávislé náhodné veličiny (znaky) Y_{ijk} s Poissonovým rozdělením pravděpodobnosti se střední hodnotou $E(Y_{ijk}) = m_{ijk}$, přičemž četnosti n_{ijk} jsou realizace Y_{ijk}. Rozsah výběru je rovněž náhodná veličina s Poissonovým rozdělením, jehož parametrem je $\sum_i \sum_j \sum_k m_{ijk}$. Sdružené rozdělení pravděpodobností $I \times J \times K$ četností je tedy

$$P(Y_{111} = n_{111}, \ldots, Y_{IJK} = n_{IJK}) = \prod_i \prod_j \prod_k \frac{e^{-m_{ijk}} m_{ijk}^{n_{ijk}}}{n_{ijk}!}. \qquad (1)$$

Pokud rozsah výběru jednotek z populace známe dopředu (předem fixujeme n), avšak počet řádků, sloupců a hladin ne, má každá četnost Y_{ijk} binomické rozdělení s parametry n a π_{ijk} a střední hodnotou $m_{ijk} = n\pi_{ijk}$. Četnosti n_{ijk} již nejsou vzájemně nezávislé a rozdělení $I \times J \times K$ četností je multinomické s parametry n a vektorem $(\pi_{111}, \pi_{112}, \ldots, \pi_{IJK})$. Aplikujeme tedy *model multinomický.* Sdružená pravděpodobnost pak vypadá následovně [1, 25]

$$P(Y_{111} = n_{111}, \ldots, Y_{IJK} = n_{IJK}) = \frac{n!}{n_{111}! \ldots n_{IJK}!} \prod_i \prod_j \prod_k \pi_{ijk}^{n_{ijk}}. \qquad (2)$$

Ukažme si nyní, že modely (1) a (2) jsou ekvivalentní, tj. podmíníme-li sdruženou pravděpodobnostní funkci Poissonova rozdělení součtem $\sum_i \sum_j \sum_k n_{ijk} = n$, pak se neliší od multinomického rozdělení. Označme $m = \sum_i \sum_j \sum_k m_{ijk}$ a nechť platí $m = n$, tj. předpokládáme, že celkové očekávané i pozorované četnosti jsou stejné. Potom

$$P(Y_{111} = n_{111}, \ldots, Y_{IJK} = n_{IJK} | \sum_i \sum_j \sum_k n_{ijk} = n) =$$

$$= \frac{\prod_i \prod_j \prod_k \frac{e^{-m_{ijk}} m_{ijk}^{n_{ijk}}}{n_{ijk}!}}{\frac{e^{-m} m^n}{n!}} =$$

$$= \frac{n!}{\prod_i \prod_j \prod_k n_{ijk}!} \prod_i \prod_j \prod_k \left(\frac{m_{ijk}}{\sum_i \sum_j \sum_k m_{ijk}} \right)^{n_{ijk}} =$$

$$= \frac{n!}{\prod_i \prod_j \prod_k n_{ijk}!} \prod_i \prod_j \prod_k \left(\frac{n\pi_{ijk}}{n} \right)^{n_{ijk}} =$$

$$= \frac{n!}{\prod_i \prod_j \prod_k n_{ijk}!} \prod_i \prod_j \prod_k \pi_{ijk}^{n_{ijk}}.$$

Prohození multinomického za Poissonovo a obráceně nepředstavuje žádný významný rozdíl, což se dočteme například v [1, 25, 41]. Vše je při analýze dat stejné jak počítání očekávaných četností, tak maximálně věrohodné odhady a testování podmodelů.

Dalším typem modelu může být některý ze *součinově multinomických* modelů. Určíme-li předem kromě rozsahu výběru i složení výběru vzhledem k jednomu sledovanému znaku např. X, tj. fixujeme-li některé marginální četnosti (např. n_{i++}), pak sdružené četnosti lze považovat za výsledky třídění I nezávislých náhodných výběrů. Četnosti $n_{i11}, \ldots, n_{iJK}$ v řádcích tabulky pak mají nezávislá multinomická rozdělení s parametry $(n_{i++}, \frac{\pi_{i11}}{\pi_{i++}}, ..., \frac{\pi_{iJK}}{\pi_{i++}})$ a s pravděpodobnostmi

$$P(Y_{i11} = n_{i11}, \ldots, Y_{iJK} = n_{iJK}) = \frac{n_{i++}!}{\prod_j \prod_k n_{ijk}} \prod_j \prod_k \left(\frac{\pi_{ijk}}{\pi_{i++}} \right)^{n_{ijk}} \quad i = 1, 2, \ldots, I.$$

Tedy sdružená pravděpodobnost vzniku tabulky rozsahu $I \times J \times K$ je

$$P(Y_{111} = n_{111}, \ldots, Y_{IJK} = n_{IJK}) = \frac{\prod_i n_{i++}!}{\prod_i \prod_j \prod_k n_{ijk}} \prod_j \prod_k \left(\frac{\pi_{ijk}}{\pi_{i++}} \right)^{n_{ijk}}.$$

Analogická situace nastává i pro předem známé marginální četnosti dvou znaků, např. řádkové a sloupcové marginální četnosti. Podrobněji se o součinově mutltinomických modelech můžeme dočíst například v knihách autorů Anděl, Pecáková a Prášková [4, 25, 26].

2 Loglineární modely

2.1 Odvození loglineárního modelu

Praktickou metodou k analyzování trojrozměrných kontingenčních tabulek je sestavení takzvaného loglineárního modelu, který vychází z teorie klasických lineárních modelů. Odvození loglineárního modelu si nejdříve popíšeme pro výběr pocházející z Poissonova rozdělení, kde není dopředu známý rozsah výběru n. Dále si ukážeme, jak situace vypadá pro pevné n (dopředu známé), přičemž za této podmínky přecházíme na multinomické rozdělení. Uvedené odvození popisuje i literatura [1, 41].

Uvažujme trojrozměrnou kontingenční tabulku rozsahu $I \times J \times K$ znaků X, Y, Z. Pozorované četnosti n_{ijk} pochází z Poissonova rozdělení se střední hodnotou $E(Y_{ijk}) = m_{ijk}$. Pro výběr z Poissonova rozdělení je obecně $m \neq n$, kde m značí celkovou očekávanou četnost a n rozsah výběru. Za předpokladu nezávislosti znaků X, Y a Z lze psát

$$m_{ijk} = m\pi_{ijk} = mP(X = i, Y = j, Z = k) = \\ = mP(X = i)P(Y = j)P(Z = k) = m\alpha_i\beta_j\gamma_k, \quad \forall i, j, k \quad (3)$$

kde α_i, β_j a γ_k jsou kladné konstanty splňující podmínky $\sum_i \alpha_i = 1$, $\sum_j \beta_j = 1$ a $\sum_k \gamma_k = 1$. Jedná se o tzv. *multiplikativní model*. Takový model není příliš vhodný k počítání odhadů neznámých parametrů m_{ijk}, a proto se logaritmováním převádí na tzv. *aditivní model*

$$\ln(m_{ijk}) = \lambda + \alpha_i^* + \beta_j^* + \gamma_k^*,$$

kde $\lambda = \ln(m), \alpha_i^* = \ln(\alpha_i), \beta_j^* = \ln(\beta_j)$ a $\gamma_k^* = \ln(\gamma_k)$.

Součet $\sum_i \sum_j \sum_k n_{ijk}$ v kontingenční tabulce má také Poissonovo rozdělení se střední hodnotou $\sum_i \sum_j \sum_k m_{ijk} = m$.

Na konci předchozí kapitoly jsme si uvedli, že Poissonovo rozdělní podmíněné celkovým součtem n je multinomické rozdělení. Z toho plyne, že za podmínky

$\sum_i \sum_j \sum_k n_{ijk} = n$ mají dle (3) četnosti n_{ijk} multinomické rozdělení s pravděpodobnostmi

$$\pi_{ijk} = \alpha_i \beta_j \gamma_k.$$

Za předpokladu nezávislosti znaků X, Y a Z platí pro marginální pravděpodobnosti

$$\pi_{i++} = \frac{m_{i++}}{m} = \frac{m\alpha_i}{m} = \alpha_i,$$

$$\pi_{+j+} = \frac{m_{+j+}}{m} = \frac{m\beta_j}{m} = \beta_j,$$

$$\pi_{++k} = \frac{m_{++k}}{m} = \frac{m\gamma_k}{m} = \gamma_k.$$

Řádkové, sloupcové a hladinové součty tvoří náhodné vektory, které mají také multinomické rozdělení, tj.

$$(n_{1++}, n_{2++}, \ldots, n_{I++}) \sim M(n, \alpha_1, \ldots, \alpha_I),$$

$$(n_{+1+}, n_{+2+}, \ldots, n_{+J+}) \sim M(n, \beta_1, \ldots, \beta_J),$$

$$(n_{++1}, n_{++2}, \ldots, n_{++K}) \sim M(n, \gamma_1, \ldots, \gamma_K).$$

Tyto vektory jsou vzájemně nezávislé. Nyní si uvědomme, že při pevném n získáváme z Poissonova modelu multinomický model. Opět za předpokladu nezávislosti znaků X, Y a Z zapišme pravděpodobnosti v loglineárním kontextu $\pi_{ijk} = \alpha_i \beta_j \gamma_k = \pi_{i++}\pi_{+j+}\pi_{++k}$. V multinomickém modelu, kde $m = n$, mají očekávané četnosti m_{ijk} tvar

$$m_{ijk} = n\pi_{i++}\pi_{+j+}\pi_{++k}.$$

Logaritmováním dostaneme

$$\ln(m_{ijk}) = \ln(n\pi_{i++}\pi_{+j+}\pi_{++k}),$$

což lze dle pravidel pro počítání s logaritmy přepsat do tvaru

$$\ln(m_{ijk}) = \ln(n) + \ln(\pi_{i++}) + \ln(\pi_{+j+}) + \ln(\pi_{++k}).$$

Při označení $\mu = \ln(n), \lambda_i^X = \ln(\pi_{i++}), \lambda_j^Y = \ln(\pi_{+j+})$ a $\lambda_k^Z = \ln(\pi_{++k})$ získáme model nazývaný *aditivní loglineární model nezávislosti*

$$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z. \tag{4}$$

V loglineárním modelu je přirozený logaritmus očekávaných četností modelován jako součet tzv. „efektů“, čímž rozumíme vlivy, které do dat, resp. modelu vnáší jednotlivé znaky X, Y a Z. V modelu (4) je to konkrétně součet vlivu celkového průměru μ a přímých vlivů znaků X, Y a Z. Parametry $\lambda_i^X, \lambda_j^Y, \lambda_k^Z$ v modelu (4) pak reprezentují přirozené logaritmy očekávaných četností jako dopad těchto efektů. Parametry lze chápat jako velikost efektu a konkrétně ukazují relativní počet případů v jednotlivých variantách znaků X, Y, Z.[16]

2.2 Typy nezávislosti mezi třemi znaky X, Y a Z

Zde se budeme podrobněji zabývat nezávislostí tří kategoriálních znaků. V trojrozměrných kontingenčních tabulkách lze rozlišit několik druhů nezávislosti, podle níž pak sestavujeme loglineární modely. Různé jsou totiž pravděpodobnosti vzniku kontingenční tabulky pokud jsou vzájemně nezávislé všechny tři znaky nebo pokud jsou některé mezi sebou závislé a jiné nezávislé. Předpokládáme multinomický model a pravděpodobnostmi $\{\pi_{ijk}\}$, kde $\sum_i \sum_j \sum_k \pi_{ijk} = 1$. Totéž lze aplikovat i pro model Poissonův s parametrem m_{ijk}.

Znaky X, Y, Z nazveme *úplně nezávislé*, jestliže pro pravděpodobnosti platí

$$P(X = i, Y = j, Z = k) = P(X = i)P(Y = j)P(Z = k)$$

neboli

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}, \quad i = 1, 2, ..., I, j = 1, 2, ..., J, k = 1, 2, ...K.$$

Znaky X a Z nazveme *sdruženě nezávislé* na znaku Y, jestliže pro pravděpodobnosti platí

$$P(X = i, Y = j, Z = k) = P(X = i, Z = k)P(Y = j)$$

neboli

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+}, \quad i = 1, 2, ..., I, j = 1, 2, ..., J, k = 1, 2, ...K.$$

Totéž platí analogicky pro sdruženou nezávislost znaků Y a Z na znaku X a sdruženou nezávislost znaků X a Y na znaku Z. Znaky X, Y nazveme *okrajově (marginálně) nezávislé*, jestliže platí

$$P(X = i, Y = j) = P(X = i)P(Y = j)$$

neboli

$$\pi_{ij+} = \pi_{i++}\pi_{+j+}, \quad i = 1, 2, ..., I, j = 1, 2, ..., J.$$

Jestliže jsou znaky X a Y nezávislé v parciální tabulce, vzniklé pro $k-$tou hladinu znaku Z, pak znaky X a Y nazveme *podmíněně nezávislé vzhledem ke $k-$té variantě znaku Z*. Podmíněná pravděpodobnost $X = i$, $Y = j$ při dané $k-$té variantě znaku Z je

$$P(X = i, Y = j|Z = k) = P(X = i, Y = j, Z = k)/P(Z = k)$$

neboli

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{++k}} \quad i = 1, 2, ..., I, j = 1, 2, ..., J.$$

Znaky X a Y jsou podmíněně nezávislé vzhledem ke $k-$té variantě znaku Z, jestliže platí

$$P(X = i, Y = j|Z = k) = P(X = i|Z = k)P(Y = j|Z = k)$$

neboli

$$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k} \quad i = 1, 2, ..., I, j = 1, 2, ..., J. \tag{5}$$

Jsou-li znaky X a Y podmíněně nezávislé vzhledem ke vešem variantám znaku Z říkáme, že znaky X a Y jsou *podmíněně nezávislé vzhledem k Z*. Pro pravděpodobnosti přitom platí

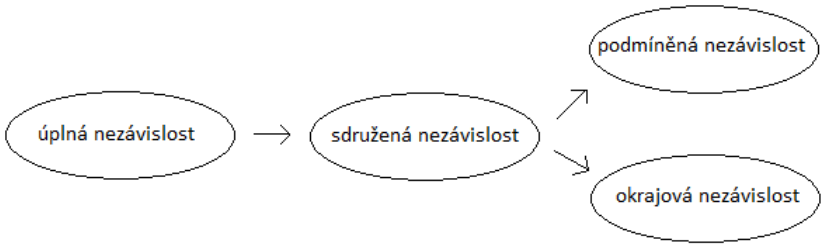
$$P(X = i, Y = j, Z = k) = P(X = i, Z = k)P(Y = j, Z = k)/P(Z = k)$$

neboli

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}, \quad i = 1, 2, ..., I, j = 1, 2, ..., J, k = 1, 2, ...K.$$

[1, 30]

Podmíněná nezávislost je slabší než úplná nezávislost i než sdružená nezávislost. Obecně platí, že úplná nezávislost implikuje sdruženou nezávislost jedné proměnné na ostatních a ta implikuje podmíněnou nezávislost. Pokud jsou tedy znaky X, Y a Z úplně nezávislé, pak jsou X a Y oba sdruženě nezávislé na Z a znaky X a Y jsou podmíněně nezávislé na Z. Sdružená nezávislost rovněž implikuje okrajovou nezávislost. Lépe vztahy mezi nezávislostmi ukazuje obrázek 2. Zde si ještě uveďme, že v loglineární notaci je častější používání symbolů π namísto klasickému značení pravděpodobnosti symbolem $P()$. Budeme tedy nadále používat symboliky π. [1, 30]



Obrázek 2: Znázornění vztahů mezi nezávislostmi

2.3 Saturovaný model

Saturovaný model je model úplný. Jedná se o model, který přesně vysvětluje vstupní data a obsahuje veškeré informace, které loglineárním modelováním lze získat. Tento model se používá především pro testování podmodelů.

Předpokládejme, že četnosti n_{ijk} v IJK buňkách jsou nezávislá pozorování s Poissonovým rozdělením s parametrem m_{ijk}. Struktura saturovaného modelu je následující

$$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \tag{6}$$

kde

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = 0, \tag{7}$$

$$\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = \sum_i \lambda_{ik}^{XZ} = \sum_k \lambda_{ik}^{XZ} = \sum_j \lambda_{jk}^{YZ} = \sum_k \lambda_{jk}^{YZ} = 0, \tag{8}$$

$$\sum_i \lambda_{ijk}^{XYZ} = \sum_j \lambda_{ijk}^{XYZ} = \sum_k \lambda_{ijk}^{XYZ} = 0. \tag{9}$$

Parametr μ reprezentuje celkový průměr logaritmů očekávaných četností a zajišťuje, že $\sum_i \sum_j \sum_k m_{ijk} = n$. [40] Parametry $\lambda_i^X, \lambda_j^Y, \lambda_k^Z$ nazýváme *hlavní efekty* jednotlivých znaků X, Y, Z a reprezentují odchylky od celkového průměru. Parametry λ_{ij}^{XY}, λ_{ik}^{XZ}, λ_{jk}^{YZ} jsou *interakce 1. řádu* neboli *dvou faktorové efekty* a parametry λ_{ijk}^{XYZ} jsou *interakce 2. řádu* neboli *tří faktorové efekty*. Model označme symbolickým zápisem nejvyšší interakce (XYZ).

Interakce 1. řádu ukazují vazby mezi dvěma znaky. Velikost těchto interakcí určuje míru závislosti příslušných dvou znaků. Interakce 2. řádu určují rozdíly mezi interakcemi 1. řádu odpovídající třetí proměnné [14]. To znamená, že parametry λ_{ijk}^{XYZ} určují rozdíly například mezi parametry λ_{ij}^{XY} a λ_k^Z.

V každém modelu jsou vzájemně provázané všechny parametry. Interakce 2. řádu ovlivňují interakce 1. řádu a hlavní efekty. Interakce 1. řádu ovlivňují hlavní efekty. Proto interpretujeme v modelu vždy nejvyšší významnou interakci. Na interakce a parametry nižšího řádu se poté již nedíváme, neboť interpretace nižších významných interakcí nebo efektů může být kvůli závislosti zavádějící. Pokud tedy budou v saturovaném modelu významné interakce 2. řádu, interpretujeme ji jako vztah tří znaků. Ve stejném modelu mohou být významné i interakce 1. řádu nebo hlavní efekty, avšak z výše uvedených důvodů je neinterpretujeme.

Rovnicím (7),(8) a (9) říkáme *podmínky identifikovatelnosti* [4]. Podmínky identifikovatelnosti tvoří restrikce na parametry a zavádíme je kvůli jednoznačné identifikaci loglineárního modelu. Model (6) má více řešení z hlediska identifikace parametrů. Například pro interakci 2. řádu bychom museli identifikovat stejný

počet parametrů jako je počet buněk v tabulce. Avšak při zavedení podmínek (7), (8) a (9) bude identifikace jednoznačná. Efekty jsou v tomto případě identifikovány ve vztahu k průměrnému efektu μ, tj. odchylky od průměrného efektu. Jedná se o tzv. kódování typu efekt. [16, 40]

parametry	počet
μ	1
λ_i^X	$I-1$
λ_j^Y	$J-1$
λ_k^Z	$K-1$
λ_{ij}^{XY}	$(I-1)(J-1)$
λ_{ik}^{XZ}	$(I-1)(K-1)$
λ_{jk}^{YZ}	$(J-1)(K-1)$
λ_{ijk}^{XYZ}	$(I-1)(J-1)(K-1)$

Tabulka 2: Počet nezávislých parametrů v saturovaném modelu

Počet nezávislých parametrů v loglineárním modelu určíme z tabulky 2. Hlavním efektům odpovídá $I+J+K$ parametrů a interakcím $IJ+JK+IK+IJK$ parametrů. K tomu musíme ještě přičíst jeden parametr μ. Tyto parametry jsou však na sobě určitým způsobem závislé. Závislost se odstraňuje například již zmíněným kódováním typu efekt. Saturovaný model má maximální možný počet nezávislých parametrů, tj.

$$1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+ \\ +(J-1)(K-1)+(I-1)(J-1)(K-1)=IJK.$$

Při testování podmodelů je potřeba znát počet stupňů volnosti, které přísluší statistikám, jež se pro testování podmodelů používají. Jedná se vlastně o ukazatele složitosti modelu. Počet stupňů volnosti v saturovaném modelu je 0, což se lehce odvodí ze vztahu (10), který najdeme například v literatuře [26]

$$\textbf{stupně volnosti} = \textbf{počet buněk} - \textbf{nezávislé parametry}, \tag{10}$$

neboť platí $0 = IJK - IJK$.

V saturovaném modelu nelze explicitně určit očekávané četnosti m_{ijk}. K výpočtu m_{ijk} používáme iteračního algoritmu, který je popsán v samostatné kapitole. Grafické znázornění saturovaného modelu najdeme na obrázku 3. Jak jsme již uvedli, saturovaný model je příliš složitý a používá se především pro srovnání s jednodušími modely.[1, 4, 16, 40]

2.4 Model párové závislosti neboli homogenní asociace

Předpokládejme, že četnosti v buňkách jsou nezávislá pozorování s Poissonovým rozdělením s parametrem m_{ijk}. V modelu párové závislosti neexistuje žádná společná interakce mezi znaky X, Y a Z. Jedná se tedy o model bez interakcí druhého řádu, kdy $\lambda_{ijk}^{XYZ} = 0$ pro všechny i,j,k. Znaky X, Y a Z jsou po dvou závislé. Vztah mezi znaky X a Y označme jako XY, vztah mezi znaky X a Z označme XZ a analogicky pro znaky Y a Z označme YZ. Pak symboly nejvyšších interakcí (XY, YZ, XZ) budeme značit model párové závislosti znaků X, Y a Z. Žádný znak není tedy nezávislý na ostatních znacích. Model párové závislosti je tvaru

$$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{11}$$

Počet nezávislých parametrů v modelu je dle tabulky 2 roven

$$1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+(J-1)(K-1) = \\ = IJ + IK + JK - (I + J + K) + 1.$$

Podle vztahu (10) má model párové závislosti $(I-1)(J-1)(K-1)$ stupňů volnosti. V modelu párové závislosti je vztah mezi dvěma znaky vůči danému třetímu znaku stejný pro každou variantu třetího znaku. Například mezi dvojicí znaků X a Y je stejný vztah vůči znaku Z v každé hladině k. Analogicky vztah dvojice znaků X a Z je stejný v každém sloupci j a vztah dvojice Y a Z je stejný v každém řádku i.

V modelu párové závislosti opět nelze explicitně určit očekávané četnosti m_{ijk} a pro jejich výpočet používáme iteračního algoritmu jako v saturovaném modelu. Grafické znázornění modelu párové závislosti najdeme na obrázku 3. [1, 4, 16, 40]

2.5 Model podmíněné nezávislosti

Předpokládejme, že četnosti v buňkách jsou nezávislá pozorování s Poissonovým rozdělením s parametrem m_{ijk}. Jsou-li nyní dva znaky podmíněně nezávislé při každé pevně zvolené variantě třetího znaku, jedná se o model *podmíněné nezávislosti*. Situaci, kdy při každé pevně zvolené variantě (hladině) znaku Z jsou znaky X a Y podmíněně nezávislé budeme označovat symbolem nejvyšších interakcí, tj.(XZ, YZ). Analogicky lze sestavit další dva modely podmíněné nezávislosti. Podmíněnou nezávislost X, Z na Y označíme (XY, YZ) a podmíněnou nezávislost Y, Z na X označíme (XY, XZ).

Uvažujme konkrétně model (XZ, YZ), kde může být vztah mezi X a Z i mezi Y a Z, avšak mezi X a Y může existovat vztah pouze přes společný vztah k Z. Tedy vzájemný vztah mezi znaky X a Y může být vysvětlen jedině prostřednictvím znaku Z, a proto $\lambda_{ij}^{XY} = 0 \quad \forall i, j$. Rovněž neexistuje interakce mezi všemi třemi znaky $\lambda_{ijk}^{XYZ} = 0 \quad \forall i, j, k$. Model podmíněné nezávislosti (XZ, YZ) vypadá následovně

$$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{12}$$

Počet nezávislých parametrů v modelu je dle tabulky 2

$$1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (J-1)(K-1) = IK + JK - K.$$

Počet stupňů volnosti užitím vztahu (10) bude $(I-1)(K-1)J$.

Pravděpodobnost π_{ijk} v tomto modelu pro každou hladinu znaku Z vypočítáme následovně. Pro podmíněnou pravděpodobnost platí

$$P(X = i, Y = j | Z = k) = P(X = i, Y = j, Z = k)/P(Z = k),$$

odtud

$$\pi_{ijk} = \pi_{ij|k}\pi_{++k}.$$

Pak užitím vztahu (5) pro podmíněné nezávislosti znaků X a Y vzhledem ke k-té variantě Z pro všechna k dostaneme

$$\pi_{ijk} = \pi_{i+|k}\pi_{+j|k}\pi_{++k} = \frac{\pi_{i+k}}{\pi_{++k}}\frac{\pi_{+jk}}{\pi_{++k}}\pi_{++k}.$$

Pokrácením získáme vztah pro výpočet očekávaných pravděpodobností v modelu podmíněné nezávislosti

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}. \tag{13}$$

Grafické znázornění modelu podmíněné nezávislosti najdeme na obrázku 3. [1, 4, 16, 40, 39]

2.6 Model sdružené nezávislosti

Předpokládejme, že četnosti v buňkách jsou nezávislá pozorování s Poissonovým rozdělením. Pokud jsou dva znaky současně nezávislé na třetím, jedná se o model *sdružené nezávislosti*. Opět existují tři modely, které lze označit (XZ, Y), (XY, Z), (YZ, X). Symbolem (XY, Z) budeme značit model, kde jsou znaky X a Y sdruženě nezávislé na znaku Z, proto je v zápise znak Z samostatně. Neexistují tedy žádné interakce mezi znakem Z a znaky X a Y, tj. $\lambda_{ik}^{XZ} = 0$ a $\lambda_{jk}^{YZ} = 0$. V modelu sdružené nezávislosti je vždy $\lambda_{ijk}^{XYZ} = 0 \quad \forall i, j, k$. Analogicky označujeme další dva modely (XZ, Y) a (YZ, X) vždy podle toho, které dva znaky jsou sdruženě nezávislé na třetím.

Uvažujme konkrétně model (XY, Z), kdy jsou znaky X a Y nezávislé na znaku Z, takže interakce XZ a YZ jsou nulové, což lze zapsat ve tvaru

$$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}. \tag{14}$$

Počet parametrů v modelu (XY, Z) je dle tabulky 2 roven

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

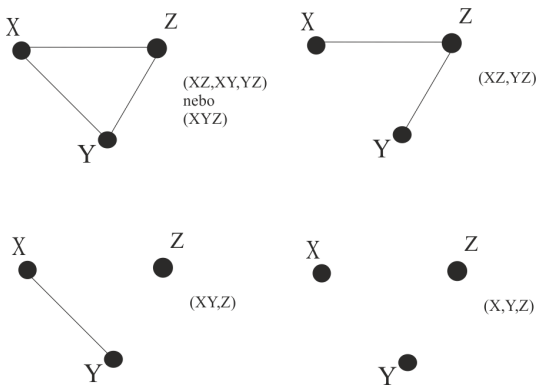
a počet stupňů volnosti je $(K-1)(IJ-1)$, což plyne ze vztahu (10). Pravděpodobnosti π_{ijk} v modelu sdružené nezávislosti (XY, Z) splňují

$$\pi_{ijk} = P(X = i, Y = j)P(Z = k) = \pi_{ij+}\pi_{++k}, \tag{15}$$

neboť předpokládáme, že X a Y jsou závislé a současně X a Z, resp. Y a Z jsou nezávislé. Grafické znázornění modelu sdružené nezávislosti najdeme na obrázku 3. [1, 4, 16, 40]

2.7 Model úplné nezávislosti

Předpokládejme, že četnosti v buňkách jsou nezávislá pozorování s Poissonovým rozdělením. Jsou-li znaky X, Y a Z nezávislé, získáme model *úplné nezávislosti*, jež je nejjednodušším modelem. Model úplné nezávislosti označme (X, Y, Z).



Obrázek 3: Grafické znázornění párové závislosti, podmíněné nezávislosti, sdružené nezávislosti a úplné nezávislosti(po řádcích zleva).
[14]

Zřejmě mezi znaky X, Y, Z nejsou žádné interakce, tj. $\lambda_{ijk}^{XYZ} = \lambda_{ij}^{XY} = \lambda_{ik}^{XZ} = \\ = \lambda_{jk}^{YZ} = 0$ pro každé i, j, k a dostáváme model odvozený v kapitole 2.1

$$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z. \tag{16}$$

Počet parametrů v tomto modelu je dle tabulky 2 roven

$$1 + (I-1) + (J-1) + (K-1) = I + J + K - 2.$$

Dosazením do vztahu (10) získáme počet stupňů volnosti roven

$$IJK - I - J - K + 2.$$

Pravděpodobnosti π_{ijk} v modelu úplné nezávislosti se vzhledem k nezávislosti všech tří znaků X, Y a Z modelují jako součiny marginálních pravděpodobností

$$P(X = i, Y = j, Z = k) = P(X = i)P(Y = j)P(Z = k)$$

neboli

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$$

Grafické znázornění modelu úplné nezávislosti najdeme na obrázku 3. [1, 4, 16, 40]

2.8 Hierarchické modely

Na obrázku 3 lze vidět, jak se dají jednotlivé typy modelů znázornit graficky. Znázornění párové závislosti není jednoznačné, jedná se buď o model párové závislosti, nebo o saturovaný model. Všechny modely, které jsme si uvedli, byly tzv. *hierarchické modely*. V přednáškách [16] najdeme definici, že hierarchický model je takový model, který s každou složitější interakcí obsahuje i všechny nižší interakce. Pokud tedy model obsahuje parametr λ_{ij}^{XY}, pak musí nutně obsahovat i oba parametry λ_i^X a λ_j^Y. Pokud je například parametr λ_k^Y nulový, pak již model není hierarchický. V knize [30] lze najít ekvivalentní definici. Pokud je v hierarchickém modelu parametr odpovídající určitému znaku (znakům) nulový, pak musí být nulové i všechny vyšší parametry, které odpovídají týmž znakům. Pokud tedy není v modelu parametr λ_{jk}^{XZ}, už v něm nemůže být ani λ_{ijk}^{XYZ}. V tabulce 3 najdeme příklady některých hierarchických modelů a v tabulce 4 přehled nehierarchických modelů. Hierarchičnost se vztahuje k bohatosti modelu (posloupnost $\lambda-$parametrů), nikoli k jednotlivým modelům vzájemně.

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

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

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

(4) $\ln(m_{ijk}) = \mu + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{YZ}$

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

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

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

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

(9) $\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$

Tabulka 3: Příklady hierarchických modelů

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

$$(2) \quad \ln(m_{ijk}) = \mu + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

$$(3) \quad \ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$$

$$(4) \quad \ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + +\lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$$

$$(5) \quad \ln(m_{ijk}) = \mu + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

$$(6) \quad \ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{jk}^{YZ}$$

Tabulka 4: Příklady nehierarchických modelů

3 Maximálně věrohodné odhady

3.1 Popis a odůvodnění metody

Maximálně věrohodné odhady (z anglického slova Maximum Likelihood Estimation, ozn. MLE) jsou snad nejklasičtější metodou pro odhad očekávaných četností m_{ijk}. Setkáme se s nimi jak v české, tak v zahraniční literatuře. Jmenujme např. Agresti [1], Stokes a spol. [30], Anděl [4], Prášková [26]. Pro trojrozměrné kontingenční tabulky s kladnými četnostmi platí, že maximálně věrohodné odhady parametrů loglineárního modelu existují a lze je jednoznačně vypočítat ze soustavy věrohodnostních rovnic. Tvrzení věty i s důkazem lze najít v knize Prášková [26]. Budeme uvažovat Poissonův model se střední hodnotou m_{ijk}. Pro multinomický model by odvození byla zbytečně složitá. Odhady MLE pro Poissonův model jsou navíc ekvivalentní MLE odhadům v multinomickém modelu, což se můžeme dočíst např. v knihách Agresti [1], Prášková [26] a Pecáková [25].

Sdružená Poissonova pravděpodobnost, že v každé buňce $\{i, j, k\}$ bude četnost n_{ijk} je

$$L(\mathbf{m}) = \prod_i \prod_j \prod_k \frac{e^{-m_{ijk}} m_{ijk}^{n_{ijk}}}{n_{ijk}!}, \tag{17}$$

kde $\mathbf{m}$ je vektor očekávaných četností $(m_{111}, m_{112}, \ldots, m_{IJK})$. Vztah (17) nazveme *Poissonovou věrohodnostní funkcí*. Pro výpočty častěji používanou logaritmická věrohodnostní funkci odvodíme následovně

$$\ln L(\mathbf{m}) = \ln \prod_i \prod_j \prod_k \frac{e^{-m_{ijk}} m_{ijk}^{n_{ijk}}}{n_{ijk}!}$$

$$\ln L(\mathbf{m}) = \sum_i \sum_j \sum_k \left(\ln e^{-m_{ijk}} + \ln m_{ijk}^{n_{ijk}} - \ln n_{ijk}! \right)$$

$$\ln L(\mathbf{m}) = \sum_i \sum_j \sum_k \left(-m_{ijk} + n_{ijk} \ln m_{ijk} - \ln n_{ijk}! \right). \tag{18}$$

A protože $\ln L$ je funkcí proměnné $\mathbf{m}$, lze člen $\ln n_{ijk}!$ nezávisející na m_{ijk} vynechat. Získáme tak logaritmickou věrohodnostní funkci ve tvaru

$$\ln L(\mathbf{m}) = \sum_i \sum_j \sum_k n_{ijk} \ln m_{ijk} - \sum_i \sum_j \sum_k m_{ijk}. \tag{19}$$

Ukažme si, jak se vztah (19) rozepíše pro saturovaný model (6), tj. pro model

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

Logaritmická věrohodnostní funkce je

$$\ln L(\mathbf{m}) = \sum_i \sum_j \sum_k n_{ijk} \left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \right) - \\ - \sum_i \sum_j \sum_k e^{\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}},$$

což lze upravit do tvaru

$$\ln L(\mathbf{m}) = n\mu + \sum_i n_{i++} \lambda_i^X + \sum_j n_{+j+} \lambda_j^Y + \sum_k n_{++k} \lambda_k^Z + \\ + \sum_i \sum_j n_{ij+} \lambda_{ij}^{XY} + \sum_i \sum_k n_{i+k} \lambda_{ik}^{XZ} + \sum_j \sum_k n_{+jk} \lambda_{jk}^{YZ} + \\ + \sum_i \sum_j \sum_k n_{ijk} \lambda_{ijk}^{XYZ} - \sum_i \sum_j \sum_k e^{\mu + \lambda_i^X + \cdots + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}}.$$

Pro další modely je logaritmická věrohodnostní funkce zcela analogická, přičemž jsou vždy vynechány některé konkrétní členy, které se v daném modelu nevyskytují.

Vztah (19) se derivuje podle každého neznámého parametru loglineárního modelu. Prášková ve své knize [26] tyto parametry pro zjednodušení značí symbolem θ. *Věrohodnostní rovnice* pak vypadají následovně

$$\frac{\partial \ln L}{\partial \theta} = -\sum_i \sum_j \sum_k \frac{\partial m_{ijk}}{\partial \theta} + \sum_i \sum_j \sum_k n_{ijk} \frac{\partial \ln m_{ijk}}{\partial \theta} =$$

$$= -\sum_i \sum_j \sum_k m_{ijk} \frac{\partial \ln m_{ijk}}{\partial \theta} + \sum_i \sum_j \sum_k n_{ijk} \frac{\partial \ln m_{ijk}}{\partial \theta} =$$

$$= \sum_i \sum_j \sum_k (n_{ijk} - m_{ijk}) \frac{\partial \ln m_{ijk}}{\partial \theta} = 0, \quad (20)$$

přičemž jsme použili rovnost $\frac{\partial \ln m_{ijk}}{\partial \theta} = \frac{1}{m_{ijk}} \frac{\partial m_{ijk}}{\partial \theta}$. Hodnota $\hat{\theta}$ parametru θ, která maximalizuje věrohodnostní funkci L, se nazývá *maximálně věrohodný odhad* parametru θ. Tedy při hledání maximálně věrohodného odhadu se obecně hledá extrém funkce L. Správnost maximalizace pomocí derivace (známého z matematické analýzy) zaručuje věta uvedená v knize pana Anděla [4] na str. 150. Maximálně věrohodné odhady parametrů loglineárního modelu jsou stejné pro výběr z multinomického rozdělení i pro výběr z Poissonova rozdělení. Ekvivalenci těchto rozdělení jsme si ukázali v závěru kapitoly 1.2. A jak jsme si uvedli na začátku této kapitoly, odvození odhadů v Poissonově modelu je jednodušší.

3.2 Model párové závislosti

V modelu párové závislosti se maximálně věrohodné odhady očekávaných četností odvodí z věrohodnostních rovnic následovně. Užitím vztahů (11) a (19) dostaneme logaritmickou věrohodnostní funkci

$$\ln L(\mathbf{m}) = n\mu + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z +$$

$$+ \sum_i \sum_j n_{ij+}\lambda_{ij}^{XY} + \sum_i \sum_k n_{i+k}\lambda_{ik}^{XZ} + \sum_j \sum_k n_{+jk}\lambda_{jk}^{YZ} -$$

$$- \sum_i \sum_j \sum_k e^{\mu+\lambda_i^X+\cdots+\lambda_{jk}^{YZ}}. \quad (21)$$

Derivací vztahu (21) podle parametru λ_{ij}^{XY} získáme věrohodnostní rovnici

$$\frac{\partial \ln L}{\partial \lambda_{ij}^{XY}} = n_{ij+} - \sum_k e^{\mu+\lambda_i^X+\cdots+\lambda_{jk}^{YZ}} = n_{ij+} - \sum_k m_{ijk} = n_{ij+} - m_{ij+}$$

a maximálně věrohodný odhad

$$\hat{m}_{ij+} = n_{ij+} \quad \forall i, j.$$

Derivací vztahu (21) podle parametrů λ_{jk}^{YZ} a λ_{ik}^{XZ} získáme podobné věrohodnostní rovnice a z nich odhady

$$\hat{m}_{i+k} = n_{i+k} \quad \forall i, k,$$

$$\hat{m}_{+jk} = n_{+jk} \quad \forall j, k.$$

Odhady parametrů hlavních efektů získáme týmž postupem z věrohodnostních rovnic. Odhad m_{i++} získáme z věrohodnostní rovnice

$$\frac{\partial \ln L}{\partial \lambda_i^X} = n_{i++} - \sum_j \sum_k m_{ijk} = 0.$$

Maximálně věrohodný odhad m_{i++} je

$$\hat{m}_{i++} = n_{i++} \quad \forall i.$$

Analogicky pro očekávané četnosti m_{+j+} a m_{++k} jsou odhady

$$\hat{m}_{+j+} = n_{+j+} \quad \forall j,$$

$$\hat{m}_{++k} = n_{++k} \quad \forall k.$$

V tomto modelu ovšem nelze explicitně vyjádřit m_{ijk} jako funkci marginálních očekávaných četností. K výpočtu se používá iterační proporční algoritmus, Newtonova Rapsonova metoda nebo „iterační metoda vyvážených nejmenších čtverců“. Postupy najdeme např. v literatuře [7, 4, 26].

3.3 Model podmíněné nezávislosti

K získání vztahu pro výpočet očekávaných četností m_{ijk} v modelu podmíněné nezávislosti (XZ,YZ) použijeme dříve zmíněných vztahů (5),(13), tj.

$$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}, \qquad \pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}.$$

Explicitní vztah pro očekávané četnosti m_{ijk} v tomto modelu je

$$m_{ijk} = \frac{m_{i+k}m_{+jk}}{m_{++k}}, \tag{22}$$

neboť platí

$$m_{ijk} = m_{+++}\pi_{ijk} = m_{+++}\frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} = m_{+++}\frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}\frac{m_{+++}}{m_{+++}} = \\ = \frac{m_{+++}\pi_{i+k}m_{+++}\pi_{+jk}}{\pi_{++k}m_{+++}} = \frac{m_{i+k}m_{+jk}}{m_{++k}}.$$

Je tedy zřejmé, že pro odhad očekávaných četností m_{ijk} budeme potřebovat odhady očekávaných četností m_{i+k}, m_{+jk} a m_{++k}. Užitím vztahů (12) a (19) dostaneme logaritmickou věrohodnostní funkci pro model podmíněné nezávislosti (XZ,YZ)

$$\ln L(\mathbf{m}) = n\mu + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z + \sum_i\sum_j n_{i+k}\lambda_{ik}^{XZ} + \\ + \sum_j\sum_k n_{+jk}\lambda_{jk}^{YZ} - \sum_i\sum_j\sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}}. \tag{23}$$

Derivací vztahu (23) podle parametru λ_{ik} získáme věrohodnostní rovnici a odhad očekávaných četností m_{i+k}.

$$\frac{\partial \ln L}{\partial \lambda_{ik}^{XZ}} = n_{i+k} - \sum_j e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}} = n_{i+k} - \sum_j m_{ijk} = n_{i+k} - m_{i+k} = 0$$

$$\hat{m}_{i+k} = n_{i+k} \quad \forall i,k.$$

Analogicky odhad m_{+jk} získáme z derivace vztahu (23) podle parametru λ_{jk}

$$\hat{m}_{+jk} = n_{+jk} \quad \forall j,k.$$

Věrohodnostní rovnici pro odhad očekávaných četností m_{++k} získáme opět derivací vztahu (23) tentokrát podle parametru λ_k, tj.

$$\frac{\partial \ln L}{\partial \lambda_k^Z} = n_{++k} - \sum_j \sum_j e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}} = n_{++k} - \sum_i \sum_j m_{ijk}.$$

Odtud

$$n_{++k} - m_{++k} = 0$$

a odhad očekávaných četností m_{++k} je

$$\hat{m}_{++k} = n_{++k} \quad \forall k.$$

Maximálně věrohodné odhady očekávaných četností m_{ijk} jsou ve tvaru

$$\hat{m}_{ijk} = \frac{\hat{m}_{i+k}\hat{m}_{+jk}}{\hat{m}_{++k}} = \frac{n_{i+k}n_{+jk}}{n_{++k}},$$

což je zřejmé dosazením jednotlivých odhadů do vztahu (22).

3.4 Model sdružené nezávislosti

Pravděpodobnosti a očekávané četnosti v modelu (XY, Z) jsou kvůli sdružené nezávislosti dané vztahem

$$\pi_{ijk} = \pi_{ij+}\pi_{++k},$$

což jsme si uvedli ve vztahu (15). Z odvození

$$m_{ijk} = m_{+++}\pi_{ijk} = m_{+++}\pi_{ij+}\pi_{++k} = \frac{m_{+++}\pi_{ij+}m_{+++}\pi_{++k}}{m_{+++}} = \frac{m_{ij+}m_{++k}}{m_{+++}}$$

je zřejmé, že explicitní vyjádření očekávaných četností m_{ijk} je ve tvaru

$$m_{ijk} = \frac{m_{ij+}m_{++k}}{m_{+++}}. \tag{24}$$

Ze vztahu (24) vidíme, že potřebujeme odhadnout tři sady očekávaných četností m_{ij+}, m_{++k} a m_{+++}. Užitím vztahů (12) a (19) dostaneme logaritmickou věrohodnostní funkci pro model sdružené nezávislosti (XY, Z)

$$\ln L(\mathbf{m}) = n\mu + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z + \\ + \sum_i \sum_j n_{ij+}\lambda_{ij}^{XY} - \sum_i \sum_j \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}}. \quad (25)$$

Věrohodnostní rovnice získáme postupně derivací vztahu (25) podle potřebných parametrů λ_{ij},λ_k a μ. Derivací podle parametru λ_{ij} získáme věrohodnostní rovnici

$$\frac{\partial \ln L}{\partial \lambda_{ij}^{XY}} = n_{ij+} - \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}} = n_{ij+} - \sum_k m_{ijk} = n_{ij+} - m_{ij+} = 0$$

a z ní odhad

$$\hat{m}_{ij+} = n_{ij+} \quad \forall i, j.$$

Již známým postupem dostaneme další dva odhady

$$\hat{m}_{++k} = n_{++k} \quad \forall k,$$

$$\hat{m}_{+++} = n.$$

Maximálně věrohodný odhad očekávaných četností m_{ijk} je v modelu sdružené nezávislosti dán vztahem

$$\hat{m}_{ijk} = \frac{\hat{m}_{ij+}\hat{m}_{++k}}{\hat{m}_{+++}} = \frac{n_{ij+}n_{++k}}{n},$$

což opět vidíme při dosazení odhadů do vztahu (24).

3.5 Model úplné nezávislosti

V tomto modelu vypočítáme neznámé pravděpodobnosti vzhledem k předpokladu nezávislosti znaků X, Y, Z jako

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k},$$

kde marginální pravděpodobnosti určíme z pozorovaných četností

$$\pi_{i++} = \frac{m_{i++}}{m_{+++}}, \qquad \pi_{+j+} = \frac{m_{+j+}}{m_{+++}}, \qquad \pi_{++k} = \frac{m_{++k}}{m_{+++}}. \quad (26)$$

Při použití (26) se očekávané četnosti m_{ijk} odvodí následovně

$$m_{ijk} = m_{+++}\pi_{ijk} = m_{+++}\pi_{i++}\pi_{+j+}\pi_{++k} = \\ = m_{+++}\pi_{i++}\frac{m_{+++}\pi_{+j+}}{m_{+++}}\frac{m_{+++}\pi_{++k}}{m_{+++}}.$$

Explicitní vztah pro očekávané četnosti m_{ijk} pak je

$$m_{ijk} = \frac{m_{i++}m_{+j+}m_{++k}}{m_{+++}^2}.$$

Opět užitím vztahů (12) a (19) dostaneme logaritmickou věrohodnostní funkci

$$\ln L(\mathbf{m}) = n\ln\mu + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z - \\ - \sum_i\sum_j\sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}.$$

Věrohodnostní rovnice je

$$\frac{\partial \ln L}{\partial \lambda_i^X} = n_{i++} - \sum_j\sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z} = n_{i++} - \sum_j\sum_k m_{ijk}$$

$$n_{i++} - m_{i++} = 0,$$

a maximálně věrohodný odhad pro m_{i++} bude

$$\hat{m}_{i++} = n_{i++} \quad \forall i.$$

Analogicky získáme další odhady

$$\hat{m}_{+j+} = n_{+j+} \quad \forall j,$$

$$\hat{m}_{++k} = n_{++k} \quad \forall k,$$

$$\hat{m}_{+++} = n.$$

Maximálně věrohodný odhad očekávaných četností m_{ijk} je

$$\hat{m_{ijk}} = \frac{\hat{m}_{i++}\hat{m}_{+j+}\hat{m}_{++k}}{\hat{m}_{+++}^2} = \frac{n_{i++}n_{+j+}n_{++k}}{n^2}.$$

3.6 IPFA versus metoda Newton-Raphson

Iterační postup při odhadování m_{ijk} nazývaný v anglické literatuře *The iterative proportional fitting algorithm (IPFA)* můžeme použít pro hierarchické modely. Jde o poměrně jednoduchou metodu, která nepotřebuje k výpočtu inverzní matice ani složité výpočty. Konverguje k maximálně věrohodným odhadům i v případě nulových četností v kontingenční tabulce. Nejspíš proto je algoritmus použit v řadě počítačových programů jako jsou SPSS Statistics[2], SAS[3], STATISTICA[4], TURNER.[5].

Postup je následující:

- počáteční aproximace se vždy volí $m_{ijk}^{(0)} = 1$ pro každé i, j, k. Je to proto, že volbou počáteční konstantní hodnoty se do tabulky nevnášejí případné interakce vyšších řádů, které by pak v dalších krocích nevymizely a výsledek by se znehodnotil.
- další aproximace počítáme dle vzorců pro $r = 0, 1, 2, \ldots$

$$\hat{m}_{ijk}^{(3r+1)} = \frac{\hat{m}_{ijk}^{(3r)}}{\hat{m}_{ij+}^{(3r)}} n_{ij+}$$

$$\hat{m}_{ijk}^{(3r+2)} = \frac{\hat{m}_{ijk}^{(3r+1)}}{\hat{m}_{i+k}^{(3r+1)}} n_{i+k}$$

$$\hat{m}_{ijk}^{(3r+3)} = \frac{\hat{m}_{ijk}^{(3r+2)}}{\hat{m}_{+jk}^{(3r+2)}} n_{+jk}.$$

Konvergence postupu zaručuje zisk maximálně věrohodných odhadů parametrů m_{ij+}, m_{i+k} a m_{+jk}, neboť platí

$$\lim_{t\to\infty} \hat{m}_{ij+}^{(t)} = n_{ij+} \qquad \lim_{t\to\infty} \hat{m}_{i+k}^{(t)} = n_{i+k} \qquad \lim_{t\to\infty} \hat{m}_{+jk}^{(t)} = n_{+jk}.$$

[4]

[2] *www.ccsr.ac.uk/staff/Ludi/documents/ProfGeogSimpsonTranmerNov*03*Full*_000.*doc*
[3] *http* : *//support.sas.com*
[4] http://www.statsoft.com
[5] http://rosuda.org/turner/

Nevýhodou IPFA je aplikovatelnost primárně na modely, ve kterých se ve věrohodnostních rovnicích rovnají pozorované a očekávané četnosti v marginálních tabulkách.

Pro lepší pochopení byl algoritmus naprogramován v programu MATLAB. Zdrojový kód najdeme v příloze č. 3.

Další metodou, jak vyřešit věrohodnostní rovnice a získat odhady parametrů v loglineárním modelu, může být *Newton-Raphsonova metoda (NR).*

Tato metoda dokáže řešit mnohem komplexnější systémy pravděpodobnostních rovnic. V každém kroku řeší systém rovnic, což může být problém ve vícerozměrných tabulkách s velkým vektorem parametrů. Mnohdy je tedy lepší metoda IPFA z hlediska času. Řád konvergence je kvadratický, tedy metoda NR je efektivnější než metoda IPFA, která má lineární řád konvergence. V SASu metodu NR používá procedura CATMOD. Jako vedlejší produkt u metody NR vzniká variační matice odhadů parametrů μ a $\lambda-$parametrů, příslušných konkrétnímu loglineárnímu modelu.[30]

Princip metody NR spočívá v Taylorově rozvoji logaritmické věrohodnostní funkce (19), přičemž se použijí pouze první dvě derivace. Další derivace se v rozvoji zanedbávají.

Ukažme si postup konkrétně opět pro model úplné nezávislosti (X, Y, Z). Použitím vztahu (19) pro model úplné nezávislosti (16) dostaneme

$$\ln L(\mathbf{m}) = \sum_i \sum_j \sum_k n_{ijk} \left(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z\right) - \sum_i \sum_j \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}. \quad (27)$$

První derivace dle všech čtyř parametrů rovnice (27) jsou

$$\left.\begin{aligned}
\frac{\partial \ln L}{\partial \mu} &= \frac{\partial^2 \ln L}{\partial \mu^2} = \textstyle\sum_i \sum_j \sum_k n_{ijk} - \sum_i \sum_j \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, \\
\frac{\partial \ln L}{\partial \lambda_i^X} &= \frac{\partial^2 \ln L}{\partial (\lambda_i^X)^2} = \textstyle\sum_j \sum_k n_{ijk} - \sum_j \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, \\
\frac{\partial \ln L}{\partial \lambda_j^Y} &= \frac{\partial^2 \ln L}{\partial (\lambda_j^Y)^2} = \textstyle\sum_i \sum_k n_{ijk} - \sum_i \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, \\
\frac{\partial \ln L}{\partial \lambda_k^Z} &= \frac{\partial^2 \ln L}{\partial (\lambda_k^Z)^2} = \textstyle\sum_i \sum_j n_{ijk} - \sum_i \sum_j e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}.
\end{aligned}\right\} \quad (28)$$

Druhé derivace podle všech dvojic parametrů jsou

$$\left.\begin{array}{ll}
\frac{\partial^2 \ln L}{\partial \mu^2} = -\sum_i \sum_j \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, & \frac{\partial^2 \ln L}{\partial (\lambda_i^X)^2} = -\sum_j \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, \\
\frac{\partial^2 \ln L}{\partial (\lambda_j^Y)^2} = -\sum_i \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, & \frac{\partial^2 \ln L}{\partial (\lambda_k^Z)^2} = -\sum_i \sum_j e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, \\
\frac{\partial^2 \ln L}{\partial \mu \partial \lambda_i^X} = -\sum_j \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, & \frac{\partial^2 \ln L}{\partial \mu \partial \lambda_j^Y} = -\sum_i \sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, \\
\frac{\partial^2 \ln L}{\partial \mu \partial \lambda_k^Z} = -\sum_i \sum_j e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, & \frac{\partial^2 \ln L}{\partial \lambda_i^X \partial \lambda_j^Y} = -\sum_k e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, \\
\frac{\partial^2 \ln L}{\partial \lambda_i^X \partial \lambda_k^Z} = -\sum_j e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}, & \frac{\partial^2 \ln L}{\partial \lambda_j^Y \partial \lambda_k^Z} = -\sum_i e^{\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z}.
\end{array}\right\} \quad (29)$$

Taylorův rozvoj věrohodnostní funkce $L(\mathbf{m})$ druhého řádu lze maticově zapsat ve tvaru

$$L(\mathbf{m}) \approx L(\mathbf{m}^s) + \left(\frac{\partial L(\mathbf{m})}{\partial \theta}\right)(\mathbf{m}-\mathbf{m}^s)^T + \frac{1}{2}(\mathbf{m}-\mathbf{m}^s)\frac{\partial^2 L(\mathbf{m})}{\partial \theta \partial \theta^T}(\mathbf{m}-\mathbf{m}^s)^T, \quad (30)$$

kde symbolem $\frac{\partial L(\mathbf{m})}{\partial \theta}$ rozumíme první derivaci funkce $L(\mathbf{m})$ podle složek vektoru $\theta = (\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z)$, tj. $\frac{\partial L(\mathbf{m})}{\partial \theta} = (\frac{\partial L(\mathbf{m})}{\partial \mu}, \frac{\partial L(\mathbf{m})}{\partial \lambda_{\mathbf{1}}^{\mathbf{X}}}, \ldots, \frac{\partial L(\mathbf{m})}{\partial \lambda_{\mathbf{I}}^{\mathbf{X}}}, \ldots, \frac{\partial L(\mathbf{m})}{\partial \lambda_{\mathbf{1}}^{\mathbf{Y}}}, \ldots, \frac{\partial L(\mathbf{m})}{\partial \lambda_{\mathbf{K}}^{\mathbf{Z}}})$. Symbolem $\frac{\partial^2 L(\mathbf{m})}{\partial \theta \partial \theta^T}$ rozumíme Hessovu matici, v našem případě matici rozměru 6×6 s mimodiagonálními prvky danými vztahy (29) a na diagonále jsou hodnoty získané pomocí vztahů (28).

Derivací Taylorova rozvoje (30) podle θ a položením derivace rovno nule, dostaneme rovnici

$$\frac{\partial L(\mathbf{m})}{\partial \theta} + \frac{\partial^2 L(\mathbf{m})}{\partial \theta \partial \theta^T}(\mathbf{m}-\mathbf{m}^s)^T = 0, \quad (31)$$

což lze přepsat do tvaru

$$\frac{\partial L(\mathbf{m})}{\partial \theta} + \frac{\partial^2 L(\mathbf{m})}{\partial \theta \partial \theta^T}\mathbf{m} = \frac{\partial^2 L(\mathbf{m})}{\partial \theta \partial \theta^T}\mathbf{m}^s.$$

Vynásobením rovnice zleva inverzní maticí k Hessově matici získáme rovnici

$$\left(\frac{\partial^2 L(\mathbf{m})}{\partial \theta \partial \theta^T}\right)^{-1}\frac{\partial L(\mathbf{m})}{\partial \theta} + \mathbf{m} = \mathbf{m}^s.$$

Odtud již přímo vyplývá vztah pro algoritmus NR

$$\mathbf{m}^{s+1} = \mathbf{m}^s + \left(\frac{\partial^2 L(\mathbf{m}^s)}{\partial\theta\partial\theta^T} \right)^{-1} \frac{\partial L(\mathbf{m}^s)}{\partial\theta}. \tag{32}$$

Ve vztahu (32) se předpokládá, že Hessova matice je regulární, aby k ní existovala inverze. Maximálně věrohodný odhad $\mathbf{m}$ je limitou $\mathbf{m}^s$ pro s rostoucí nade všechny meze.[1]

4 Metoda nejmenších čtverců

V této kapitole se budeme věnovat odvození soustav normálních rovnic, ze kterých následně získáme odhady očekávaných četností m_{ijk}. V některých případech bude možné četnosti m_{ijk} vypočítat explicitně, v jiných použijeme k výpočtu četností m_{ijk} iterační algoritmus. Ukážeme alternativní přístup k získávání očekávaných četností m_{ijk} pánů Deminga a Stephana, který publikovali v roce 1940 [7]. Jedná se o získání odhadů m_{ijk} pomocí metody nejmenších čtverců, přičemž se minimalizuje funkce

$$S = \sum_i \sum_j \sum_k \frac{(m_{ijk} - n_{ijk})^2}{n_{ijk}}. \tag{33}$$

V populaci jsou při vzniku kontingenční tabulky vždy dopředu známé některé marginální součty četností, což pak umožňuje okamžitě vypočítat příslušné očekávané marginální četnosti. Metodu si podrobně vysvětlíme pro případy dopředu známých součtů, které odpovídají již zmíněným loglineárním modelům párové závislosti, podmíněné, sdružené a úplné nezávislosti. V dalším textu budeme předpokládat, že z populace rozsahu N vybereme konkrétní vzorek rozsahu n, čímž získáme data pro trojrozměrnou kontingenční tabulku $I \times J \times K$ taktéž rozsahu n. Symbolem $\frac{N}{n}$ budeme značit výběrový poměr a pro populaci bude platit

$$N_{1++} + \cdots + N_{I++} = N_{+1+} + \cdots + N_{+J+} = N_{++1} + \cdots + N_{++K} = N.$$

Analogické rovnice platí i pro marginální součty četností ve výběru

$$n_{1++} + \cdots + n_{I++} = n_{+1+} + \cdots + n_{+J+} = n_{++1} + \cdots + n_{++K} = n.$$

4.1 Model párové závislosti

Známe-li v populaci, ze které vznikl výběr pro kontingenční tabulku $I \times J \times K$ součty populačních četností

$$N_{11+}, N_{12+} \ldots, N_{IJ+},$$

$$N_{1+1}, N_{1+2} \ldots, N_{I+K},$$

$$N_{+11}, N_{+12} \ldots, N_{+JK},$$

pak požadujeme, aby platily podmínky

$$\sum_k m_{ijk} = m_{ij+} = \frac{N_{ij+}n}{N}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \tag{34}$$

$$\sum_i m_{ijk} = m_{+jk} = \frac{N_{+jk}n}{N}, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K-1, \tag{35}$$

$$\sum_j m_{ijk} = m_{i+k} = \frac{N_{i+k}n}{N}, \quad i = 1, \ldots, I-1, \quad k = 1, \ldots, K-1. \tag{36}$$

Situace odpovídá modelu párové závislosti. K odvození podmínek (34), (35) a (36) používáme tzv. kódování dummy, které zaručuje jednoznačnost výsledku. Díky kódování není model tzv. přeurčený, tj. neobsahuje nadbytečné množství parametrů. Dummy kódování je druhý přístup k identifikaci parametrů v modelech. Oproti kódování efekt jsou efekty parametrů identifikovány k sobě navzájem. Jde o odchylky od jednoho zvoleného parametru (kategorie), který nazýváme *referenční*. Referenční kategorie lze volit zcela libovolně. Zde byly zvoleny jako referenční kategorie poslední kategorie, tj. $m_{+jK} \quad \forall j$, $m_{i+K} \quad \forall i$ a $m_{I+k} \quad \forall k$ a parametry λ_{+jK}, λ_{i+K} a λ_{I+k} jsou tedy pro tyto kategorie rovny nule.[16] Podmínky (34), (35) a (36) jsou jen jednou z možností, jak zavést dummy kódování pro model párové závislosti. Obecně existuje šest možných variant, jak zvolit poslední kategorie jako referenční. Aby byly parametry jednoznačně identifikovány, musí počet podmínek odpovídat počtu parametrů v příslušném modelu. V našem případě modelu párové závislosti, kde je počet parametrů $(IJ+JK+IK-I-J-K+1)$. Počet podmínek (34), (35) a (36) je

$$IJ + J(K-1) + (I-1)(K-1), \tag{37}$$

což odpovídá počtu parametrů pro model párové závislosti. Další možností může být například

$$\sum_k m_{ijk} = m_{ij+} = \frac{N_{ij+}n}{N}, \quad i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1,$$

$$\sum_i m_{ijk} = m_{+jk} = \frac{N_{+jk}n}{N}, \quad j = 1, \ldots, J-1, \quad k = 1, \ldots, K,$$

$$\sum_j m_{ijk} = m_{i+k} = \frac{N_{i+k}n}{N}, \quad i = 1, \ldots, I, \quad k = 1, \ldots, K$$

nebo

$$\sum_k m_{ijk} = m_{ij+} = \frac{N_{ij+}n}{N}, \quad i = 1, \ldots, I-1, \quad j = 1, \ldots, J,$$

$$\sum_i m_{ijk} = m_{+jk} = \frac{N_{+jk}n}{N}, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K,$$

$$\sum_j m_{ijk} = m_{i+k} = \frac{N_{i+k}n}{N}, \quad i = 1, \ldots, I-1, \quad k = 1, \ldots, K-1.$$

Nyní chceme získat soustavu normálních rovnic, ze které se vypočítají odhady očekávaných četností. Minimalizujeme tedy vztah (33). Minimalizací rozumíme hledání extrému „funkce“ (33) vzhledem k podmínkám (34), (35), (36). Tato situace se řeší tzv. metodou Lagrangeových multiplikátorů. *Lagrangeova funkce* je tvaru

$$\Phi(m_{ijk}) = \sum_i \sum_j \sum_k \frac{(m_{ijk} - n_{ijk})^2}{n_{ijk}} - 2\sum_k \lambda_{ijk} m_{ijk} - \\ - 2\sum_i \lambda_{ijk} m_{ijk} - 2\sum_j \lambda_{ijk} m_{ijk},$$

kde λ_{ijk} jsou Lagrangeovy multiplikátory. V dalších krocích derivujeme vztah Φ dle konkrétní četnosti m_{ijk}

$$\frac{d\Phi(m_{ijk})}{dm_{ijk}} = 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\sum_k \lambda_{ijk} - 2\sum_i \lambda_{ijk} - 2\sum_j \lambda_{ijk} =$$

$$= 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\lambda_{ij+} - 2\lambda_{+jk} - 2\lambda_{i+k} \quad \forall i, j, k.$$

Derivaci položíme rovno 0 a můžeme vypočítat očekávané četnosti m_{ijk}

$$\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - \lambda_{ij+} - \lambda_{+jk} - \lambda_{i+k} = 0 \quad \forall i, j, k,$$

$$\frac{m_{ijk} - n_{ijk}\,(1 + \lambda_{ij+} + \lambda_{+jk} + \lambda_{i+k})}{n_{ijk}} = 0 \quad \forall i, j, k,$$

$$m_{ijk} - n_{ijk}\,(1 + \lambda_{ij+} + \lambda_{+jk} + \lambda_{i+k}) = 0 \quad \forall i, j, k,$$

$$m_{ijk} = n_{ijk}(1 + \lambda_{ij+} + \lambda_{+jk} + \lambda_{i+k}), \quad \forall i, j, k, \tag{38}$$

přičemž vzhledem ke kódování dummy je λ_{+jK} pro každé j je rovno nule, λ_{I+k} pro každé k je rovno nule a λ_{i+K} pro každé i jsou taktéž rovny nule. Jakmile budeme znát Lagrangeovy multiplikátory, vypočítáme okamžitě očekávané četnosti m_{ijk}. K výpočtu použijeme rovnice (34), (35), (36) a (38). Soustavu $(IJ + JK + IK - I - J - K + 1)$ normálních rovnic o stejném počtu neznámých Lagrangeových multiplikátorech získáme sčítáním přes jednotlivé varianty znaků X, Y a Z, tj. přes řádky (i), sloupce (j) a vrstvy (k) ve vztahu (38), čímž získáme systém rovnic

$$\sum_k m_{ijk} = \sum_k n_{ijk}\,(1 + \lambda_{ij+} + \lambda_{+jk} + \lambda_{i+k})\,, \tag{39}$$

$$\sum_i m_{ijk} = \sum_i n_{ijk}\,(1 + \lambda_{ij+} + \lambda_{+jk} + \lambda_{i+k})\,, \tag{40}$$

$$\sum_j m_{ijk} = \sum_j n_{ijk}\,(1 + \lambda_{ij+} + \lambda_{+jk} + \lambda_{i+k})\,. \tag{41}$$

Sumarizací přes příslušné indexy lze rovnice přepsat do tvaru

$$m_{ij+} = n_{ij+} + \sum_k n_{ijk}\lambda_{ij+} + \sum_k n_{ijk}\lambda_{+jk} + \sum_k n_{ijk}\lambda_{i+k}, \tag{42}$$

$$m_{+jk} = n_{+jk} + \sum_i n_{ijk}\lambda_{ij+} + \sum_i n_{ijk}\lambda_{+jk} + \sum_i n_{ijk}\lambda_{i+k}, \tag{43}$$

$$m_{i+k} = n_{i+k} + \sum_j n_{ijk}\lambda_{ij+} + \sum_j n_{ijk}\lambda_{+jk} + \sum_j n_{ijk}\lambda_{i+k}. \tag{44}$$

Výsledná soustava normálních rovnic je

$$m_{ij+} - n_{ij+} = n_{ij+}\lambda_{ij+} + \sum_k n_{ijk}\lambda_{+jk} + \sum_k n_{ijk}\lambda_{i+k}, \tag{45}$$

$$m_{+jk} - n_{+jk} = \sum_i n_{ijk}\lambda_{ij+} + n_{+jk}\lambda_{+jk} + \sum_i n_{ijk}\lambda_{i+k}, \tag{46}$$

$$m_{i+k} - n_{i+k} = \sum_j n_{ijk}\lambda_{ij+} + \sum_j n_{ijk}\lambda_{+jk} + n_{i+k}\lambda_{i+k}. \tag{47}$$

Rovnice (39), (42) a (45) probíhají přes $i = 1, \ldots, I, j = 1, \ldots, J$, rovnice (40), (43) a (46) probíhají přes $j = 1, \ldots, J, k = 1, \ldots, K-1$ a rovnice (41), (44) a (47) probíhají přes $i = 1, \ldots, I-1, k = 1, \ldots, K-1$.

Soustavu normálních rovnic řešíme iteračně. Iterační postup bude demonstrován v kapitole (4.5) pro model úplné nezávislosti. Pro model párové závislosti bychom postupovali analogicky.

4.2 Model podmíněné nezávislosti

V populaci se oproti kapitole (4.1) předpokládá, že jsou známé pouze součty populačních četností

$$N_{1+1}, N_{1+2} \ldots, N_{I+K}$$

$$N_{+11}, N_{+12} \ldots, N_{+JK},$$

což odpovídá modelu podmíněné nezávislosti (XZ, YZ), kdy jsou znaky X a Y podmíněně nezávislé na znaku Z. Požadavkem pak je

$$\sum_j m_{ijk} = m_{i+k} = \frac{N_{i+k}n}{N}, \quad i = 1, \ldots, I, \quad k = 1, \ldots, K, \tag{48}$$

$$\sum_i m_{ijk} = m_{+jk} = \frac{N_{+jk}n}{N}, \quad j = 1, \ldots, J-1, \quad k = 1, \ldots, K. \tag{49}$$

Minimalizací vztahu (33) vzhledem k podmínkám (48), (49) metodou Lagrangeových multiplikátorů získáme *Lagrangeovu funkci*

$$\Phi(m_{ijk}) = \sum_i \sum_j \sum_k \frac{(m_{ijk} - n_{ijk})^2}{n_{ijk}} - 2\sum_i \lambda_{ijk} m_{ijk} - 2\sum_j \lambda_{ijk} m_{ijk},$$

kde λ_{ijk} jsou Lagrangeovy multiplikátory. V dalších krocích derivujeme vztah Φ dle konkrétní četnosti m_{ijk}

$$\frac{d\Phi(m_{ijk})}{dm_{ijk}} = 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\sum_i \lambda_{ijk} - 2\sum_j \lambda_{ijk} =$$

$$= 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\lambda_{+jk} - 2\lambda_{i+k} \quad \forall i, j, k$$

$$\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - \lambda_{+jk} - \lambda_{i+k} = 0 \quad \forall i, j, k$$

$$\frac{m_{ijk} - n_{ijk}(1 + \lambda_{+jk} + \lambda_{i+k})}{n_{ijk}} = 0 \quad \forall i, j, k$$

$$m_{ijk} - n_{ijk}(1 + \lambda_{+jk} + \lambda_{i+k}) = 0 \quad \forall i, j, k$$

$$m_{ijk} = n_{ijk}(1 + \lambda_{i+k} + \lambda_{+jk}), \quad \forall i, j, k, \tag{50}$$

přičemž λ_{+Jk} pro každé k je rovno 0. Poslední kategorie znaku Y je opět kvůli dummy kódování zvolena jako referenční. Očekávané četnosti m_{ijk} vypočítáme, pokud budeme znát Lagrangeovy multiplikátory. K výpočtu použijeme rovnice

(48), (49) a (50). Soustavu ($K(I + J - 1)$) normálních rovnic získáme sčítáním rovnic (50) přes řádky a sloupce, tj.

$$\sum_j m_{ijk} = \sum_j n_{ijk} \left(1 + \lambda_{i+k} + \lambda_{+jk}\right) \quad i = 1, \ldots, I, \quad k = 1, \ldots, K,$$

$$\sum_i m_{ijk} = \sum_i n_{ijk} \left(1 + \lambda_{i+k} + \lambda_{+jk}\right) \quad j = 1, \ldots, J-1, \quad k = 1, \ldots, K.$$

Odtud

$$m_{i+k} = n_{i+k} + \sum_j n_{ijk}\lambda_{i+k} + \sum_j n_{ijk}\lambda_{+jk} \quad i = 1, \ldots, I, \quad k = 1, \ldots, K,$$

$$m_{+jk} = n_{+jk} + \sum_i n_{ijk}\lambda_{i+k} + \sum_i n_{ijk}\lambda_{+jk} \quad j = 1, \ldots, J-1, \quad k = 1, \ldots, K.$$

Výslednou soustavu normálních rovnic lze přepsat do tvaru

$$m_{i+k} - n_{i+k} = n_{i+k}\lambda_{i+k} + \sum_j n_{ijk}\lambda_{+jk} \quad i = 1, \ldots, I, \quad k = 1, \ldots, K,$$

$$m_{+jk} - n_{+jk} = \sum_i n_{ijk}\lambda_{i+k} + n_{+jk}\lambda_{+jk} \quad j = 1, \ldots, J-1, \quad k = 1, \ldots, K.$$

Očekávané četnosti m_{ijk} opět najdeme iteračním postupem.

4.3 Model sdružené nezávislosti

Nyní v populaci známe pouze součty populačních četností

$$N_{11+}, N_{12+} \ldots, N_{IJ+},$$

což odpovídá modelu sdružené nezávislosti (XY, Z), kdy jsou oba znaky X a Y sdruženě nezávislé na Z. Požadujme tedy

$$\sum_k m_{ijk} = m_{ij+} = \frac{N_{ij+}n}{N}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J. \tag{51}$$

Lagrangeova funkce je ve tvaru

$$\Phi(m_{ijk}) = \sum_i \sum_j \sum_k \frac{(m_{ijk} - n_{ijk})^2}{n_{ijk}} - 2 \sum_k \lambda_{ijk} m_{ijk},$$

kde λ_{ijk} jsou Lagrangeovy multiplikátory. V dalších krocích derivujeme vztah Φ dle konkrétní četnosti m_{ijk}

$$\frac{d\Phi(m_{ijk})}{dm_{ijk}} = 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\sum_k \lambda_{ijk} = 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\lambda_{ij+} \quad \forall i, j, k,$$

$$\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - \lambda_{ij+} = 0 \quad \forall i, j, k,$$

$$\frac{m_{ijk} - n_{ijk}\,(1 + \lambda_{ij+})}{n_{ijk}} = 0 \quad \forall i, j, k,$$

$$m_{ijk} - n_{ijk}\,(1 + \lambda_{ij+}) = 0 \quad \forall i, j, k,$$

$$m_{ijk} = n_{ijk}\,(1 + \lambda_{ij+}) \quad \forall i, j, k. \tag{52}$$

Sčítáním přes k získáme

$$\sum_k m_{ijk} = \sum_k n_{ijk}\,(1 + \lambda_{ij+}) \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

$$m_{ij+} = n_{ij+}\,(1 + \lambda_{ij+}) \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

$$\frac{m_{ij+}}{n_{ij+}} = 1 + \lambda_{ij+} \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$

Odtud, dosazením do rovnice (52) získáme

$$\hat{m}_{ijk} = n_{ijk}\,(1 + \lambda_{ij+}) = n_{ijk}\frac{m_{ij+}}{n_{ij+}} \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

čímž jsme tentokrát obdrželi přímo vztah pro výpočet očekávaných četností m_{ijk}.

4.4 Model úplné nezávislosti

Nyní jsou známé všechny tři sady marginálních součtů populace

$$N_{1++}, N_{2++}, \ldots, N_{I++}$$

$$N_{+1+}, N_{+2+}, \ldots, N_{+J+}$$

$$N_{++1}, N_{++2}, \ldots, N_{++K}.$$

Tato situace odpovídá modelu úplné nezávislosti. Požadavky na očekávané četnosti jsou formulované takto

$$\sum_j \sum_k m_{ijk} = m_{i++} = \frac{N_{i++}n}{N}, \quad i = 1, \ldots, I, \tag{53}$$

$$\sum_i \sum_k m_{ijk} = m_{+j+} = \frac{N_{+j+}n}{N}, \quad j = 1, \ldots, J-1, \tag{54}$$

$$\sum_i \sum_j m_{ijk} = m_{++k} = \frac{N_{++k}n}{N}, \quad k = 1, \ldots, K-1. \tag{55}$$

Opět zavádíme kódování dummy. Vztahy (53), (54) a (55) dávají $I + (J-1) + (K-1) = I + J + K - 2$ podmínek/rovnic. Lagrangeova funkce je

$$\Phi(m_{ijk}) = \sum_i \sum_j \sum_k \frac{(m_{ijk} - n_{ijk})^2}{n_{ijk}} - \\ - 2\sum_j \sum_k \lambda_{ijk} m_{ijk} - 2\sum_i \sum_k \lambda_{ijk} m_{ijk} - 2\sum_i \sum_j \lambda_{ijk} m_{ijk},$$

kde λ_{ijk} jsou Lagrangeovy multiplikátory. V dalších krocích derivujeme vztah Φ dle konkrétní četnosti m_{ijk}

$$\frac{d\Phi(m_{ijk})}{dm_{ijk}} = 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\sum_j \sum_k \lambda_{ijk} - 2\sum_i \sum_k \lambda_{ijk} - 2\sum_i \sum_j \lambda_{ijk} = \\ = 2\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - 2\lambda_{i++} - 2\lambda_{+j+} - 2\lambda_{++k} \quad \forall i, j, k,$$

$$\frac{m_{ijk} - n_{ijk}}{n_{ijk}} - \lambda_{i++} - \lambda_{+j+} - \lambda_{++k} = 0 \quad \forall i, j, k,$$

$$\frac{m_{ijk} - n_{ijk}\left(1 + \lambda_{i++} + \lambda_{+j+} + \lambda_{++k}\right)}{n_{ijk}} = 0 \quad \forall i, j, k,$$

$$m_{ijk} - n_{ijk}\left(1 + \lambda_{i++} + \lambda_{+j+} + \lambda_{++k}\right) = 0 \quad \forall i, j, k,$$

$$m_{ijk} = n_{ijk}\left(1 + \lambda_{i++} + \lambda_{+j+} + \lambda_{++k}\right), \tag{56}$$

přičemž λ_{+J+} a λ_{++K} jsou rovny 0. Soustava normálních rovnic se odvodí následovně

$$\sum_j \sum_k m_{ijk} = \sum_j \sum_k n_{ijk}\left(1 + \lambda_{i++} + \lambda_{+j+} + \lambda_{++k}\right), \tag{57}$$

$$\sum_i \sum_k m_{ijk} = \sum_i \sum_k n_{ijk}\left(1 + \lambda_{i++} + \lambda_{+j+} + \lambda_{++k}\right), \tag{58}$$

$$\sum_i \sum_j m_{ijk} = \sum_i \sum_j n_{ijk}\left(1 + \lambda_{i++} + \lambda_{+j+} + \lambda_{++k}\right), \tag{59}$$

$$m_{i++} = n_{i++} + \sum_j \sum_k n_{ijk}\lambda_{i++} + \sum_j \sum_k n_{ijk}\lambda_{+j+} + \sum_j \sum_k n_{ijk}\lambda_{++k}, \tag{60}$$

$$m_{+j+} = n_{+j+} + \sum_i \sum_k n_{ijk}\lambda_{i++} + \sum_i \sum_k n_{ijk}\lambda_{+j+} + \sum_i \sum_k n_{ijk}\lambda_{++k}, \tag{61}$$

$$m_{++k} = n_{++k} + \sum_i \sum_j n_{ijk}\lambda_{i++} + \sum_i \sum_j n_{ijk}\lambda_{+j+} + \sum_i \sum_j n_{ijk}\lambda_{++k}, \tag{62}$$

a tedy

$$m_{i++} - n_{i++} = n_{i++}\lambda_{i++} + \sum_j n_{ij+}\lambda_{+j+} + \sum_k n_{i+k}\lambda_{++k}, \tag{63}$$

$$m_{+j+} - n_{+j+} = \sum_i n_{ij+}\lambda_{i++} + n_{+j+}\lambda_{+j+} + \sum_k n_{+jk}\lambda_{++k}, \tag{64}$$

$$m_{++k} - n_{++k} = \sum_i n_{i+k}\lambda_{i++} + \sum_j n_{+jk}\lambda_{+j+} + n_{++k}\lambda_{++k}. \tag{65}$$

Rovnice (57), (60) a (63) probíhají $i = 1, \ldots, I$, rovnice (58), (61) a (64) probíhají $j = 1, \ldots, J-1$ a rovnice (59), (62) a (65) probíhají $k = 1, \ldots, K-1$. Soustava se opět řeší iteračně. Podrobnější popis si ukážeme v následující kapitole.

4.5 Iterační algoritmus výpočtu očekávaných četností v modelu úplné nezávislosti

Nyní si ukážeme metodu, jak získat očekávané četnosti m_{ijk} v případě modelu úplné nezávislosti. Tato metoda se v literatuře obvykle nazývá anglickým názvem The Iterative Proportionate Fitting Algorithm. Český překlad by mohl být Iterační proporční algoritmus.

Z kapitoly 4.4 jsme získali soustavu normálních rovnic danou rovnicemi (63)-(65). Soustava normálních rovnic se řeší explicitním vyjádření λ_{i++} z rovnice (63) a dosazením do dalších rovnic (64) a (65). Stále přitom uvažujeme předpoklady z kódování dummy, tj. referenční kategorie $\lambda_{+J+} = 0$ i $\lambda_{++K} = 0$. Vyjádříme λ_{i++} z (63)

$$\lambda_{i++} = \frac{1}{n_{i++}} \left(m_{i++} - n_{i++} - \sum_j n_{ij+}\lambda_{+j+} - \sum_k n_{i+k}\lambda_{++k} \right),$$

$$\lambda_{i++} = \frac{1}{n_{i++}} \left(m_{i++} - \sum_j n_{ij+}\lambda_{+j+} - \sum_k n_{i+k}\lambda_{++k} \right) - 1$$

a dosadíme do rovnice (56) pro očekávané četnosti m_{ijk} a dostaneme

$$m_{ijk} = n_{ijk} \left[\frac{1}{n_{i++}} \left(m_{i++} - \sum_j n_{ij+}\lambda_{+j+} - \sum_k n_{i+k}\lambda_{++k} \right) + \lambda_{+j+} + \lambda_{++k} \right].$$

Analogicky se získají podobné vztahy vyjádřením λ_{+j+} z (64) a λ_{++k} z (65), tj.

$$m_{ijk} = n_{ijk} \left[\frac{1}{n_{+j+}} \left(m_{+j+} - \sum_i n_{ij+}\lambda_{i++} - \sum_k n_{+jk}\lambda_{++k} \right) + \lambda_{i++} + \lambda_{++k} \right],$$

$$m_{ijk} = n_{ijk} \left[\frac{1}{n_{++k}} \left(m_{++k} - \sum_i n_{i+k}\lambda_{i++} - \sum_j n_{+jk}\lambda_{+j+} \right) + \lambda_{i++} + \lambda_{+j+} \right].$$

K nalezení očekávaných četností m_{ijk} metodou nejmenších čtverců potřebujeme vztahy

$$m_{ijk}^{(1)} = n_{ijk} \left(\frac{m_{i++}}{n_{i++}} \right),$$

$$m_{ijk}^{(2)} = m_{ijk}^{(1)} \left(\frac{m_{+j+}}{m_{+j+}^{(1)}} \right),$$

$$m_{ijk}^{(3)} = m_{ijk}^{(2)} \left(\frac{m_{++k}}{m_{++k}^{(2)}} \right),$$

které nám poslouží jako algoritmus, který se opakuje celý nebo částečný dokud nezískáme m_{ijk} splňující podmínky (53), (54) a (55).

Podobným postupem se z normálních rovnic získají očekávané četnosti i pro složitější modely.

Algoritmus pro výpočet očekávaných četností v modelu úplné nezávislosti byl vyzkoušen softwarem MATLAB. M-file se zdrojovým kódem najdeme v příloze č. 4.

5 Posuzování kvality modelu

5.1 Testování vhodnosti modelu

Ve chvíli, kdy máme k dispozici hierarchický model s odhadnutými parametry, je nutné otestovat kvalitu modelu (vhodnost) pro daná data. K tomu slouží dvě základní statistiky, které porovnávají pozorované a očekávané četnosti. Jsou jimi statistika X^2 odpovídající Pearsonově chí-kvadrát statistice pro testy dobré shody a statistika G^2 označovaná jako *deviance*, která vychází z testu věrohodnostním poměrem. Statistiky jsou ve tvaru

$$G^2 = 2 \sum_i \sum_j \sum_k n_{ijk} \ln \left(\frac{n_{ijk}}{\hat{m}_{ijk}} \right), \tag{66}$$

$$X^2 = \sum_i \sum_j \sum_k \frac{(n_{ijk} - \hat{m}_{ijk})^2}{\hat{m}_{ijk}}, \tag{67}$$

kde n_{ijk} a m_{ijk} označují pozorované a očekávané četnosti. Pokud model správně vyrovnává data, mají obě statistiky asymptoticky χ^2 rozdělení se stupni volnosti, které odpovídají počtu buněk v tabulce zmenšených o počet nezávislých parametrů. Počet stupňů volnosti odpovídá rozdílu dimenzí alternativní a nulové hypotézy. Dimenzí alternativní hypotézy se v tomto případě rozumí počet parametrů v saturovaném modelu a nulovou hypotézou počet parametrů pro zvolený model. [1, 4, 30]

Nejprve si podrobněji ukážeme odvození statistiky G^2. Označme $L_n(\mathbf{m})$ logaritmickou věrohodnostní funkci Poissonova rozdělení. Maximum této funkce označme $L_n(\hat{\mathbf{m}})$. Maximální možnou hodnotu této funkce, kterou jsme schopni spočítat z dat, označme $L_n(\mathbf{x})$. Hodnota $L_n(\mathbf{x})$ odpovídá hodnotě logaritmické věrohodnostní funkce v saturovaném modelu. Deviance je definována následovně

$$G^2(\hat{\mathbf{m}}) = -2 \left[L_n(\hat{\mathbf{m}}) - L_n(\mathbf{x}) \right].$$

Užitím vztahu (18) pro logritmickou věrohodnostní funkci získáme

$$G^2(\hat{\mathbf{m}}) = -2\left[L_n(\hat{\mathbf{m}}) - L_n(\mathbf{x})\right] = -2[\sum_i \sum_j \sum_k (-\hat{m}_{ijk} + n_{ijk} \ln \hat{m}_{ijk} - \ln n_{ijk}!) -$$

$$- \sum_i \sum_j \sum_k (-n_{ijk} + n_{ijk} \ln n_{ijk} - \ln n_{ijk}!)] =$$

$$= -2[\sum_i \sum_j \sum_k (-\hat{m}_{ijk}) + \sum_i \sum_j \sum_k n_{ijk} \ln \hat{m}_{ijk} -$$

$$- \sum_i \sum_j \sum_k (-n_{ijk}) - \sum_i \sum_j \sum_k n_{ijk} \ln n_{ijk}.$$

Uvědomme si, že $\sum_i \sum_j \sum_k \hat{m}_{ijk} = n$. Potom lze poslední výraz upravit a získáme

$$-2[-n + \sum_i \sum_j \sum_k n_{ijk} \ln \hat{m}_{ijk} - (-n) - \sum_i \sum_j \sum_k n_{ijk} \ln n_{ijk}] =$$

$$= -2 \sum_i \sum_j \sum_k n_{ijk} \ln \hat{m}_{ijk} + 2 \sum_i \sum_j \sum_k n_{ijk} \ln n_{ijk} =$$

$$= -2 \sum_i \sum_j \sum_k n_{ijk} \left(\ln \hat{m}_{ijk} - \ln n_{ijk}\right) =$$

$$= 2 \sum_i \sum_j \sum_k n_{ijk} \left(\ln n_{ijk} - \ln \hat{m}_{ijk}\right) =$$

$$= 2 \sum_i \sum_j \sum_k n_{ijk} \frac{\ln n_{ijk}}{\hat{m}_{ijk}} = G^2.$$

Nyní se zaměříme přímo na testování podmodelů a uvedeme si nulové hypotézy. Při testování podmodelů se vychází, jako u všech statistických testů, z nulové hypotézy proti alternativě. Nulovou hypotézou se vždy rozumí typ nezávislosti příslušného podmodelu, který chceme testovat a alternativou je saturovaný model. Budeme-li mít například model úplné nezávislosti a budeme-li chtít otestovat, zda tento model odpovídá datům, použijeme jako nulovou hypotézu úplnou nezávislost a statistikou G^2 nebo χ^2 testujeme proti alternativě saturovaného modelu. Nulové hypotézy pro jednotlivé modely jsou vypsány v tabulce 5. [41]

$$
\begin{array}{ll}
(X,Y,Z) & H_0 : \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \\
(XY,Z) & H_0 : \pi_{ijk} = \pi_{ij+}\pi_{++k} \\
(XZ,Y) & H_0 : \pi_{ijk} = \pi_{i+k}\pi_{+j+} \\
(YZ,X) & H_0 : \pi_{ijk} = \pi_{i++}\pi_{+jk} \\
(XY,XZ) & H_0 : \pi_{ijk} = \frac{\pi_{ij+}\pi_{i+k}}{\pi_{i++}} \\
(XY,YZ) & H_0 : \pi_{ijk} = \frac{\pi_{ij+}\pi_{+jk}}{\pi_{+j+}} \\
(YZ,XZ) & H_0 : \pi_{ijk} = \frac{\pi_{+jk}\pi_{i+k}}{\pi_{++k}}
\end{array}
$$

Tabulka 5: Nulové hypotézy pro jednotlivé modely

5.2 Porovnávání modelů

Při volbě vhodného modelu je vždy důležité najít správný kompromis mezi přesností a úsporností modelu. Nejpřesnějším modelem je samozřejmě saturovaný model, který má však příliš mnoho parametrů. Nejúspornějším modelem z hlediska parametrů je model úplné nezávislosti. Ten naopak nemusí obsahovat podstatné informace o vztazích v datech. Proto se hledá model, který bude vhodný z obou hledisek. Takových modelů, které odpovídají datům a splňují předpoklady na přesnost i úspornost, existuje většinou více. Je tedy potřeba určit, který model lépe odpovídá. K tomuto posuzování se používají statistiky (66) a (67), díky nimž je pak snažší rozhodnout, který model bude vhodnější.

Statistika G^2 má dvě výhody oproti X^2. Zaprvé je odvozená metodou maximální věrohodnosti a zadruhé je vhodná přímo k porovnání dvou modelů. Jsou tedy k dispozici dva modely M_1 a M_2, přičemž druhý model je speciální případ prvního (druhý model tedy obsahuje podmnožinu λ-parametrů obsažených v prvním modelu). Prvnímu modelu odpovídá statistika $G^2(M_1)$ s v_1 stupni volnosti a druhému statistika $G^2(M_2)$ s v_2 stupni volnosti. Protože model M_2 je jednodušší než model M_1, platí $G^2(M_2) \geq G^2(M_1)$ a $v_2 \geq v_1$. Pro porovnávání modelů M_1 a M_2 se užívá statistika

$$
G^2(M_2|M_1) = G^2(M_2) - G^2(M_1) = -2\left[L_n(\hat{\mathbf{m}}_{M_2}) - L_n(\hat{\mathbf{m}}_{M_1})\right], \qquad (68)
$$

která má asymptoticky χ^2 rozdělení s $(v_2 - v_1)$ stupni volnosti, za předpokladu

správnosti obou modelů. Jak je vidět ze vztahu (68), rozdíl deviancí vede na test věrohodnostním poměrem. [30]

Dalším přístupem k porovnávání několika modelů mohou být tzv. *informační kritéria*. Získáme-li více modelů, které vhodně popisují chování dat, je potřeba zjistit, který model je lepší. Lepší v tomto případě znamená lépe odpovídající reálnému světu, nikoli přesnější v matematickém smyslu. Kritéria mohou tedy posoudit bohatost informace z dat, kterou v sobě nese model. Patří mezi ně *Bayesovo* a *Akaikeho*. V případě velkých souborů dat statistické významnosti selhávají, protože se prakticky každý model ukáže jako významný (dle statistik G^2 a χ^2). Je tedy těžké je porovnat a vhodným nástrojem jsou právě informační kritéria. Za lepší model pak považujeme ten, který má nejmenší hodnotu informačního kritéria.[16, 22]

Prvním kritériem, které si uvedeme, je tzv. Bayesovo informační kritérium, které se počítá dle vzorce

$$BIC = -2\ln L + p\ln(n),$$

kde n je rozsah výběru a p je počet nezávislých parametrů v modelu. Lze používat u porovnávací BIC, které srovnává BIC testovaného modelu a saturovaného modelu (budeme označovat BIC_p). Odvození porovnávacího BIC je následující

$$BIC_{sat.model} = -2\ln L_s + n\ln(n)$$

$$BIC_{test.model} = -2\ln L + p\ln(n)$$

$$BIC_p = BIC_{test.model} - BIC_{sat.model} = \underbrace{2\ln L_s - 2\ln L}_{G^2} - \underbrace{n\ln(n) + p\ln(n)}_{k\ln(n)}$$

$$BIC_p = G^2 - k\ln(n),$$

kde k je počet stupňů volnosti.

Dalším často používaným kritériem je Akaikeho informační kritérium. Dle definice se počítá následovně

$$AIC = -2\ln L + 2p,$$

kde p je počet nezávislých parametrů v modelu. Analogicky jako u BIC lze použít porovnávacího AIC (označení AICp), které srovnává AIC testovaného modelu a saturovaného modelu, tj.

$$AIC_{sat.model} = -2 \ln L_s + 2n$$

$$AIC_{test.model} = -2 \ln L + 2p$$

$$AIC_p = AIC_{test.model} - AIC_{sat.model} = \underbrace{2 \ln L_s - 2 \ln L}_{G^2} - \underbrace{2n + 2p}_{2k} \qquad (69)$$

$$AIC_p = G^2 - 2k,$$

k je opět počet stupňů volnosti.

Korekce pro konečný výběr, která je vhodná, když je rozsah n malý nebo naopak počet parametrů v modelu p velký. [1, 16, 47]

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}.$$

Pro zkoumání odlišnosti dat od modelu lze použít i tzv. *index odlišnosti* neboli dissimilarity index. Nechť pro pozorované četnosti n_{ijk} platí $n_{ijk} = np_{ijk}$ a pro očekávané četnosti m_{ijk} platí $\hat{m}_{ijk} = n\hat{\pi}_{ijk}$. Vzorec pro výpočet indexu je

$$\Delta = \sum_i \sum_j \sum_k \frac{|n_{ijk} - \hat{m}_{ijk}|}{2n} = \sum_i \sum_j \sum_k \frac{|p_{ijk} - \hat{\pi}_{ijk}|}{2}.$$

Hodnota indexu odlišnosti je mezi nulou a jedničkou a čím je menší, tím model lépe respektuje data. Malá hodnota indexu odpovídá malému rozdílu pozorovaných a očekávaných četností. Velká hodnota indexu naopak značí velký rozdíl v očekávaných a pozorovaných četnostech. Když je $\Delta = 0$, pak model odpovídá perfektně datům. V praxi může být index odlišnosti roven nule pouze pro saturovaný model. Hodnota indexu odlišnosti udává o jaký poměr musí být pozorované četnosti změněny, aby přesně odpovídaly modelovým četnostem. [1, 21]

Další metodou vhodnou k porovnávání modelů je R^2 známý z lineárních modelů a jeho adjustovaná verze. Svoji analogii má i v loglineárních modelech. Hodnota R^2 udává, jak moc je celková variabilita dat vysvětlena použitým modelem. Testovaný model musí vysvětlovat více variability než jeho podmodely, což je podmínkou pro realizaci výpočtu R^2. Adjustované R^2 mění původní R^2 tím, že penalizuje větší model za to, že vysvětluje více variability v datech. V loglineárních modelech hraje statistika G^2 (tj. testování modelu vůči saturovanému

modelu) stejnou roli jako součty čtverců (tj. reziduální součet čtverců a celkový součet čtverců) v lineárních modelech. Je tedy namístě R^2 z lineárních modelů modifikovat právě pomocí G^2. V loglineárním modelu je R^2 definován vztahem

$$R^2 = \frac{G^2(\text{model úplné nezávislosti}) - G^2(\text{test. model})}{G^2(\text{model úplné nezávislosti})}. \tag{70}$$

Testovaný model je porovnáván s nejméně zajímavým modelem (tj. model úplné nezávislosti), neboť pak je G^2(model úplné nezávislosti) mírou celkové variability v datech. Rozdíl v čitateli vztahu (70) je mírou variability vysvětlované testovaným modelem. Pro saturovaný model je R^2 rovno 1, neboť G^2 je rovno 0. Důležitou podmínkou pro použití R^2 je stejný počet stupňů volnosti v testovaném modelu a v modelu nejméně významným.

Adjustované R^2 je klasické R^2 s korekcí na velikost souboru, tj.

$$Adj.R^2 = 1 - \frac{IJK - \text{DF}(\text{model úplné nezávislosti})}{IJK - \text{DF}(\text{test. model})(1 - R^2)},$$

kde DF značí stupně volnosti modelu, I je počet řádků v tabulce, J počet sloupců a K počet hladin. Dosazením (70) do adjustovaného R^2 získáme

$$\begin{aligned} Adj.R^2 &= 1 - \frac{IJK - \text{DF}(\text{model úplné nezávislosti})}{IJK - \text{DF}(\text{test. model})} \cdot \\ &\quad \cdot \left[1 - \frac{G^2(\text{model úplné nezávislosti}) - G^2(\text{test. model})}{G^2(\text{model úplné nezávislosti})}\right] = \\ &= 1 - \frac{IJK - \text{DF}(\text{model úplné nezávislosti})}{IJK - \text{DF}(\text{test. model})} \cdot \frac{G^2(\text{test. model})}{G^2(\text{model úplné nezávislosti})} = \\ &= 1 - \frac{\frac{G^2(\text{test.model})}{IJK - \text{DF}(\text{test.model})}}{\frac{G^2(\text{model úplné nezávislosti})}{IJK - \text{DF}(\text{model úplné nezávislosti})}}. \end{aligned}$$

Pro lepší přehlednost označme IJK jako n, model úplné nezávislosti M_0 a jeho stupně volnosti DF_{M_0} a testovaný model označme M_1 a jeho stupně volnosti DF_{M_1}. Získáme vztah

$$Adj.R^2 = 1 - \frac{\frac{G^2(M_1)}{n-DF_{M_1}}}{\frac{G^2(M_0)}{n-DF_{M_0}}}.$$

Velká hodnota $Adj.R^2$ značí, že testovaný model (tj. M_1) datům dobře odpovídá. [13]

6 Šance a poměry šancí

6.1 Šance a poměry šancí v kontingenčních tabulkách

Dalším přístupem k analyzování kontingenčních tabulek jsou poměry šancí, které jsou známé pod anglickým názvem odds ratio. Pro jednoduchost budeme tedy poměr šancí značit symbolem OR. V prvé řadě je potřeba si uvědomit, že šance není to samé jako pravděpodobnost. Obecně však platí úvaha, že čím větší je šance na nějakou událost, tím větší je pravděpodobnost té události. Tedy určitý vztah mezi šancí a pravděpodobností existuje. Pojem šance definujme následovně. *Šancí výskytu jevu A se rozumí podíl pravděpodobnosti výskytu jevu A a pravděpodobnosti výskytu jevu opačného k jevu A.* Zmíněný vztah mezi šancí a pravděpodobností si nyní uvedeme.

Nechť π je pravděpodobnost jevu A. Pak šance lze zapsat vztahem

$$\Omega = \frac{\pi}{1-\pi}, \tag{71}$$

kde Ω je vždy nezáporné číslo. Je-li $\Omega > 1$, pak je jev A mnohem pravděpodobnější než jev opačný k jevu A. Čím je šance blíže nule, tím je pravděpodobnost výskytu jevu A menší. Ze vzorce pro šanci (71) lze lehce odvodit vzorec pro výpočet pravděpodobnosti π výskytu jevu A, tj.

$$\begin{aligned}
\Omega &= \frac{\pi}{1-\pi} \qquad /(1-\pi) \\
\Omega(1-\pi) &= \pi \\
\Omega - \Omega\pi - \pi &= 0 \\
\pi(1+\Omega) &= \Omega \\
\pi &= \frac{\Omega}{1+\Omega}.
\end{aligned}$$

Jev A budeme nadále nazývat *úspěch* a jev opačný k jevu A budeme nazývat *neúspěch.*

Na následujícím jednoduchém příkladu si ukážeme, co šance znamenají. Mějme sto osob, které trpí určitou nemocí (data jsou ilustrační, lze předpokládat například rakovinu) a některé léčíme určitou metodou (například chemoterapií). V tabulce 6 vidíme, počty vyléčených (resp. nevyléčených) pacientů v závislosti na

léčbě chemoterapií. Znakem X budeme rozumět léčbu chemoterapií s kategoriemi léčeno/neléčeno a znakem Y výsledek léčby s kategoriemi vyléčen/nevyléčen.

	vyléčeno	nevyléčeno
léčeno	43	7
neléčeno	18	32

Tabulka 6: Léčba rakoviny chemoterapií

Pravděpodobnost, že se pacient *vyléčí*, pokud *podstoupí* chemoterapii je

$$\frac{n_{11}}{n_{11}+n_{12}} = \frac{43}{50} \doteq 0,86.$$

Pravděpodobnost, že se pacient *vyléčí*, pokud *nepodstoupí* chemoterapii je

$$\frac{n_{21}}{n_{21}+n_{22}} = \frac{18}{50} \doteq 0,36.$$

Pravděpodobnost, že se pacient *nevyléčí*, pokud *podstoupí* chemoterapii je

$$\frac{n_{12}}{n_{11}+n_{12}} = \frac{7}{50} \doteq 0,14.$$

Pravděpodobnost, že se pacient *nevyléčí*, pokud *nepodstoupí* chemoterapii je

$$\frac{n_{22}}{n_{21}+n_{22}} = \frac{32}{50} \doteq 0,64.$$

Nyní již můžeme vypočítat šance. Šance na vyléčení pacienta při léčbě chemoterapií je

$$\Omega = \frac{\frac{n_{11}}{n_{11}+n_{12}}}{\frac{n_{12}}{n_{11}+n_{12}}} = \frac{n_{11}}{n_{12}} = \frac{43}{7} \doteq 6,14. \tag{72}$$

Analogicky šance na vyléčení pacienta, když pacient nepodstupuje chemoterapii je

$$\Omega = \frac{\frac{n_{21}}{n_{21}+n_{22}}}{\frac{n_{22}}{n_{21}+n_{22}}} = \frac{n_{21}}{n_{22}} = \frac{18}{32} \doteq 0,56. \tag{73}$$

Pro pacienta však bude určitě důležité, jak si tyto šance stojí proti sobě. Jednoduchým nástrojem k porovnání těchto šancí je právě poměr šancí. Poměrem

šancí rozumíme podíl úspěchu jedné metody a úspěchu druhé metody. Vztah pro poměr šancí v klasické dvourozměrné kontingenční tabulce 2×2 je definován jako

$$OR = \frac{\frac{n_{11}}{n_{12}}}{\frac{n_{21}}{n_{22}}},$$

z čehož úpravou získáme již zmíněný vztah pro poměr šancí

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{74}$$

Jedná se o poměr šancí na úspěch dvou různých metod (přístupů). Na pozici n_{11} se nachází úspěch první metody, tedy léčení chemoterapií. Naopak na pozici n_{21} je úspěch při druhé metodě, tedy pacienti se neléčí chemoterapií. Číslo z, kterému se pak OR bude rovnat, budeme reprezentovat jako $z-$násobná šance na úspěch první metody (chemoterapie) proti druhé metodě (bez chemoterapie). Poměr šancí může nabývat libovolné kladné hodnoty. Pro lepší interpretaci je možno vyjadřovat jej v procentech. Bude-li například $OR = 2$, pak je dvakrát větší šance na úspěch určité události. Je to o 100% větší šance na úspěch oproti neúspěchu, tedy na 200 pozorování s úspěchem připadá 100 pozorování s neúspěchem. Naopak vyjde-li $OR = 0,5$ je šance na úspěch poloviční proti neúspěchu. Tedy 50% úspěšných pozorování. Bude-li poměr šancí 1, pak jsou stejné šance na úspěch i neúspěch. Úspěch je v tomto případě jen termín. Paradoxně „úspěchem“ v tomto kontextu může být úmrtí při nějaké nemoci. Z tabulky 6 spočítáme poměr šancí dle vzorce (74)

$$OR = \frac{43 \times 32}{18 \times 7} = 10,92. \tag{75}$$

Pacient má 10,92 krát větší šanci na vyléčení rakoviny, když se bude léčit chemoterapií, než když se léčit nebude.

Přirozenou vlastností poměru šancí je, že pokud zvětšíme (resp. zmenšíme) každou buňku kontingenční tabulky o stejné číslo, zůstane poměr šancí taktéž stejný. Je to proto, že zvětšíme-li (resp. zmenšíme) rozsah výběru dvakrát, pak se hodnoty při dosazení do vzorce zkrátí a poměr šancí zůstane zachován. Stejné platí i pro změny vůči marginálním součtům. Podrobněji v textu [16].

Dosud jsme hovořili o šancích a poměrech šancí v tzv. čtyřpolní tabulce (kontingenční tabulka 2×2.) Nyní se podívejme, jak se situace změní pro obecnou dvourozměrnou tabulku $I \times J$. V tabulce, kde mají znaky více variant, existuje poměrů šancí několik. Některé poměry jsou však nadbytečné, neboť k popisu vztahů v tabulce $I \times J$ stačí pouze $(I-1)(J-1)$ poměrů šancí, tzv. neredundantních. Zbylé poměry šancí lze odvodit právě z neredundantních poměrů šancí. V kontingenční tabulce $I \times J$ je $ij-$tý poměr šancí definován pro každé $i = 1, \ldots, I-1$ a $j = 1, \ldots, J-1$ [16]

$$OR_{ij} = \frac{n_{ij} n_{(i+1)(j+1)}}{n_{i(j+1)} n_{(i+1)j}}. \tag{76}$$

V případě, že chceme počítat libovolné poměry šancí nebo se nechceme omezovat jen na sousední dvojice, můžeme vzorec zevšeobecnit pro libovolné dva řádky i', i'' a dva sloupce j' a j'', tj.

$$\tilde{OR}_{ij} = \frac{n_{i'j'} n_{i''j''}}{n_{i'j''} n_{i''j'}}. \tag{77}$$

	smrt	těžké zranění	lehké zranění	bez zranění
dětský věk	**8**	**15**	**22**	**20**
produktivní věk	**40**	**32**	**18**	**11**
důchodový věk	**12**	**36**	**34**	**6**

Tabulka 7: Počty osob zraněných při autonehodě v závislosti na jejich věku

Uvažujme kontingenční tabulku 7, kde se jedná o ilustrační data lidí, kteří měli autonehodu. Pro takovou tabulku 3×4 existuje k popisu vztahů 18 poměrů šancí (pro 1. a 2. řádek, 1. a 3. řádek a 2. a 3. řádek vždy po 6 možnostech volby dvou sloupců). Obecně pro tabulku $I \times J$ platí pro počet poměrů šancí vzorec

$$\binom{I}{2}\binom{J}{2} = \frac{I(I-1)J(J-1)}{4}.$$

Přidáme-li i reciproční poměry šancí, pak je pro tabulku 7 celkově 72 poměrů šancí (tj. $18 \cdot 4$). Číslo 4 je tam proto, že v každém poměru šancí můžeme přehodit sloupce, řádky případně obojí a získáme další tři poměry šancí, které jsou však nadbytečné. Počet poměrů šancí lze vždy zmenšit na počet neredundantních poměrů. V případě, že chceme poměr šancí 2. a 3. řádku a 2. a 4. sloupce, pak

jej můžeme vypočítat dle vzorce (76) nebo použít zjednodušení násobením jiných poměrů šancí.

Toto zjednodušení si popišme blíže. Poměr šancí 2. a 3. řádku a 3. a 4. sloupce je $\frac{n_{23}n_{34}}{n_{33}n_{24}}$. Poměr šancí 2. a 3. řádku a 2. a 3. sloupce je $\frac{n_{22}n_{33}}{n_{23}n_{32}}$. Vynásobením těchto dvou poměrů získáme kýžený poměr 2. a 3. řádku a 2. a 4. sloupce $\frac{n_{22}n_{34}}{n_{32}n_{24}}$.

Nyní zpět k tabulce 7 a spočítejme si názorněji šanci na vyváznutí bez zranění. Při počítání použijeme vzorce (77). Šance, že dítě přežije autonehodu bez zranění je $\frac{20}{45} \doteq 0,444$ proti ostatním variantám výsledku autonehody. Šance, že vyvázne bez zranění (než že nastane horší varianta) dospělý jedinec (myšleno osoby v produktivním věku) je $\frac{11}{90} \doteq 0,1222$ a šance, že autonehodu bez zranění přežije důchodce je $\frac{6}{82} \doteq 0,0732$. Největší šanci, že člověk přežije autonehodu bez zranění, má v dětském věku. Poměr šancí dítěte na vyváznutí bez zranění je oproti dospělým lidem (produktivní i důchodový věk) $\frac{20 \cdot 172}{45 \cdot 17} \doteq 4,5$. Dítě má tedy oproti dospělým 4,5 krát větší šanci na vyváznutí z autonehody bez zranění. Jinými slovy připadá na 100 dospělých, kteří přežijí nehodu bez zranění, připadá 450 dětí, kteří přežijí bez zranění.

Budeme-li používat přímo vzorec (77) a zvolíme-li kategorie následovně. Kategorie i' bude dětský věk, kategorie i'' bude důchodový věk, kategorie j' bude smrt a kategorie j'' bude bez zranění. Přímým dosazením do (77) dostaneme, $\frac{8 \cdot 6}{12 \cdot 20} \doteq 0,2$. Dítě má tedy oproti důchodci o pětinu menší šanci, že zemře při autonehodě než že přežije. Pro důchodce je tedy oproti dítěti o 80% jistější, že při autonehodě zemře, než že vyvázne bez zranění.

Nyní přejděme k šancím a poměrům šancí v trojrozměrných případech. Nejprve si uveďme, co budeme rozumět pod označením parciální a marginální tabulky. Sledujeme-li pevně zvolenou hladinu znaku Z a zkoumáme-li vztah X a Y vzhledem k této hladině, pak z výsledků jednotlivých dvojrozměrných tabulek lze odvodit chování celého znaku Z. Jedná se o tzv. *parciální kontingenční tabulky.* Počet parciálních tabulek pro X a Y je roven počtu hladin znaku Z. V parciálních tabulkách „kontrolujeme“ znak Z. Znakem Z budeme rozumět rozlišení pacientů z tabulky 6 dle pohlaví. Znaky X a Y zůstávají stejné jako v tabulce 6. Dostaneme tak tabulku (resp. tabulky) 8. *Marginální tabulka* vznikne v případě, že kombinujeme parciální tabulky. V buňkách marginální tabulky budou tedy součty četností parciálních tabulek na stejných pozicích. Z tabulky 8 získáme eliminací znaku Z

marginální tabulku součtem přes znak Z, tj. pohlaví. Jedná se přímo o tabulku 6.

		vyléčeno	nevyléčeno
ženy	léčeno	**23**	**2**
	neléčeno	**15**	**11**
		vyléčeno	nevyléčeno
muži	léčeno	**20**	**5**
	neléčeno	**3**	**21**

Tabulka 8: Parciální tabulka pro ženy a parciální tabulka pro muže

Dalším nezbytným pojmem budou tzv. *marginální šance*, kterými rozumíme šance na konkrétní úspěch a počítáme je z marginálních četností. Marginální šance na vyléčení pacienta při léčbě je stejná jako v tabulce 6, tj. 6,14. Stejně tak marginální šance na vyléčení pacienta při neléčení chemoterapií, tj. 0,56. Marginální poměr šancí na vyléčení rakoviny při léčbě chemoterapií oproti neléčení chemoterapií v tabulce 6 je právě 10,92, což plyne z (75).

Zjišťujeme-li vztah mezi dvěma znaky při současném eliminování třetí proměnné, pak jde o tzv. *podmíněné šance*. Podmíněné šance jsou vztahy v parciálních tabulkách. Výpočty se provádí dle stejných vzorců jako u výpočtu šancí a poměrů šancí v dvojrozměrných kontingenčních tabulkách, viz vzorce (72), (73). Rozdíl mezi nimi je pouze ten, že podmíněný poměr šancí se počítá pro jednotlivé podskupiny. Tedy v tabulce 8 budeme počítat šance zvlášť pro ženy a zvlášť pro muže. Čím více se podmíněné šance na jednu věc v rámci kategorií jiné proměnné od sebe odlišují, tím je mezi zkoumanými znaky silnější vztah. Rozdělíme-li osoby v tabulce 6 opět dle pohlaví na muže a ženy, získáme již zmíněnou tabulku 8. Podmíněný poměr šancí ženy na vyléčení rakoviny (při léčení chemoterapií oproti neléčení chemoterapií) je $\frac{23\times 11}{2\times 15} \doteq 8,43$ a muže $\frac{20\times 21}{5\times 3} = 28$. Z čehož je zřejmé, že u mužů je chemoterapie úspěšnější než u žen. [1, 16, 44]

Uvažujme kontingenční tabulku $2 \times 2 \times 2$. V tabulce 6 lze spočítat poměr šancí (resp. poměr poměrů šancí) na vyléčení při léčbě chemoterapií mezi ženami a muži tak, že dáme do poměru parciální poměry šancí pro muže a ženy a získáme $\frac{8,43}{28} \doteq 0,301$, což, jsme získali z následujícího vztahu

$$OR_3 = \frac{\frac{n_{111}n_{221}}{n_{121}n_{211}}}{\frac{n_{112}n_{222}}{n_{122}n_{212}}}. \tag{78}$$

Poměr (78) interpretujeme vzhledem k třetí proměnné, v tomto případě vzhledem k pohlaví. Udává kolikrát jsou větší (resp. menší) šance na událost první kategorie znaku Z oproti druhé kategorii znaku Z. Čím je OR_3 vzdálenější od 1, tím jsou větší rozdíly v kategoriích znaku Z. Pokud by OR_3 bylo rovno 1, pak by rozdíl mezi kategoriemi nebyl naprosto žádný a hovoříme o tzv. *homogennosti podmíněných poměrů šancí*. Znak Z v takové situaci nehraje v kontingenční tabulce roli a není potřeba jej interpretovat. [16] Vzhledem k této úvaze mají ženy téměř třetinovou šanci na vyléčení rakoviny, pokud podstoupí chemoterapii, oproti mužům. Podobným postupem lze testovat vliv druhého (resp. prvního) znaku v kontingenční tabulce tím, že zaměníme pořadí znaků. Poměr OR_3 však vyjde naprosto stejný, z čehož plyne, že na pořadí znaků v kontingenční tabulce nezáleží. Je tedy jedno, zda je pohlaví znakem Z nebo znakem Y. Může nás totiž například zajímat otázka, zda je pro ženu větší šance na vyléčení rakoviny, při podstoupení chemoterapie proti tomu, že chemoterapii nepodstoupí. Znakem X je nyní pohlaví (řádky), znakem Y je vyléčení (sloupce) a znakem Z je podstoupení chemoterapie (hladiny). Číslo OR_3 bude $\frac{\frac{23\cdot 5}{20\cdot 2}}{\frac{15\cdot 21}{3\cdot 11}} \doteq 0,301$. Opět mají ženy oproti mužům třetinovou šanci na vyléčení rakoviny při podstoupení chemoterapie nebo jinak řečeno, muži mají téměř 3 a půl krát větší šanci na vyléčení rakoviny než muži ($\frac{1}{0,301} \doteq 3,32$).

Mějme nyní tabulku $I \times J \times K$. Zvolme pro jednoduchost opět tabulku 7, avšak rozšiřme ji o informaci o roku, ve kterém byly nehody sledovány. Ilustrační data najdeme v tabulce 9.

K smysluplnému popsání vztahů mezi třemi znaky nám stačí méně poměrů šancí, než je počet buněk v tabulce. Poměr šancí lze určit pro libovolnou dvojici řádků a sloupců nebo řádků a vrstev, případně sloupců a vrstev. Celkový počet možností pro výběr je:

$$\begin{array}{rlll} \text{řádků} & \dots & \binom{I}{2} & = \frac{I(I-1)}{2}, \\ \text{sloupců} & \dots & \binom{J}{2} & = \frac{J(J-1)}{2}, \\ \text{vsrtev} & \dots & \binom{K}{2} & = \frac{K(K-1)}{2}. \end{array}$$

rok	věk	smrt	těžké zranění	lehké zranění	bez zranění
2000	dětský věk	**4**	**13**	**12**	**8**
	produktivní věk	**23**	**16**	**13**	**6**
	důchodový věk	**9**	**16**	**16**	**2**
		smrt	těžké zranění	lehké zranění	bez zranění
2010	dětský věk	**4**	**2**	**10**	**12**
	produktivní věk	**17**	**16**	**5**	**5**
	důchodový věk	**3**	**20**	**18**	**4**

Tabulka 9: Výsledky autonehod v letech 2000 a 2010 rozlišeny dle věku

Poměrů šancí v trojrozměrné kontingenční tabulce bude dohromady

$$\binom{I}{2}\binom{J}{2}\binom{K}{2} = \frac{IJK(I-1)(J-1)(K-1)}{8}.$$

Poměr šancí v trojrozměrných kontingenčních tabulkách $I \times J \times K$ lze počítat z parciálních tabulek, popisujících podmíněnou závislost. Získáme tím skupinu poměrů pro posouzení závislosti dvou znaků vzhledem k pevně zvolené kategorii třetího znaku. O homogenitě znaků X a Y hovoříme tehdy, je-li podmíněný poměr šancí dvou kategorií znaku X a dvou kategorií znaku Y stejný vzhledem ke každé kategorii Z.[1]

Podmíněný poměr šancí znaků X a Y při fixní kategorii k znaku Z spočítáme dle (v $i'-$tém a $i''-$tém řádku a $j'-$tém a $j''-$tém sloupci)

$$\tilde{OR}_{XY(Z)} = \frac{n_{i'j'k}n_{i''j''k}}{n_{i'j''k}n_{i''j'k}}.$$

Využijeme-li možnosti dopočítávání poměrů šancí z již vypočítaných poměrů, pak lze užít vzorec

$$OR_{XY(Z)} = \frac{n_{ijk}n_{(i+1)(j+1)k}}{n_{i(j+1)k}n_{(i+1)jk}}.$$

Marginální poměr šancí dvou znaků X a Y při eliminování třetího znaku Z spočítáme dle následujícího vztahu

$$\tilde{OR}_{XY} = \frac{n_{i'j'+}n_{i''j''+}}{n_{i'j''+}n_{i''j'+}}$$

a analogicky pro výpočet nejmenšího počtu poměrů šancí

$$OR_{XY} = \frac{n_{ij+}n_{(i+1)(j+1)+}}{n_{i(j+1)+}n_{(i+1)j+}}.$$

Označíme-li $i'-$tý řádek jako kategorii dítě, $i''-$tý řádek jako kategorie dospělý, $j'-$tý sloupec jako kategorii bez zranění a $j''-$tý sloupec jako kategorii smrt, můžeme pro dvě fixní kategorie roku spočítat dva poměry šancí. Z tabulky 9 lze tedy nejprve spočítat šance na vyváznutí dítěte z autonehody bez zranění v roce 2000 vůči úmrtí, tj. $\frac{8}{4} \doteq 2$. Analogicky pro rok 2010 bude taková šance $\frac{12}{4} \doteq 3$. Vůči úmrtí při autonehodě je vyváznutí bez zranění pravděpodobnější v roce 2010. Zlepšení je o dvě třetiny. Dáme-li šance do poměru vůči dospělým (ne důchodcům) získáme $OR_{2000} = \frac{8 \cdot 23}{6 \cdot 4} \doteq 7,666$ s $OR_{2010} = \frac{12 \cdot 17}{5 \cdot 4} \doteq 10,2$ a spočítáme-li poměr těchto poměrů dostaneme číslo $0,752$. Takový výsledek nám říká, že šance na vyváznutí bez nehody (oproti úmrtí) byla pro dítě (oproti dospělému) v roce 2000 více než poloviční oproti roku 2010. Neboli stejná šance je v roce 2010 je 1,33 krát větší než v roce 2000 ($\frac{1}{0,752} \doteq 1,33$).

6.2 Šance a poměry šancí v loglineárních modelech

Poměry šancí najdou své využití i v loglineární analýze. Splňují-li poměry šancí v kontingenční tabulce jisté podmínky, pak určují nezávislost resp. závislost proměnných, které tabulku tvoří. Budeme zde hovořit o modelech uvedených v kapitole 2.

Mějme model podmíněné nezávislosti znaků X a Y, tj. model (XZ, YZ), kde jsou interakce 2. řádu pro každé i, j, k rovny nule a interakce 1. řádu $\lambda_{ij}^{XY} = 0$ pro každé i, j. V literatuře [40] se ale dočteme, že předpoklad na nulovost konkrétní interakce 1. řádu nemusí být jediný. V případě, že interakce 2. řádu je nulová pro každé i, j, k a podmíněný poměr šancí

$$OR_{XY(Z)} = \frac{m_{ijk}m_{i'j'k}}{m_{i'jk}m_{ij'k}}$$

je roven 1, pak taktéž vznikne model podmíněné nezávislosti (XZ, YZ). Jinými slovy v tabulkách $X \times Y$ nesmí být pro žádné $k = 1, \ldots K$ podmíněný poměr šancí statisticky významný od 1.

Symetricky se předpoklad přenese i pro modely podmíněné nezávislosti (X, Z) a (Y, Z).

Analogická situace nastane pro modely sdružené nezávislosti. Mějme konkrétně model sdružené nezávislosti (XY, Z). V literatuře [40] je opět uvedena druhá varianta pro vznik modelu. Model (XY, Z) vznikne v případě, že

$$\lambda_{ijk}^{XYZ} = 0 \quad \forall i, j, k$$

a podmíněné poměry šancí

$$OR_{XZ(Y)} = \frac{m_{ijk}m_{i'jk'}}{m_{i'jk}m_{ijk'}},$$

$$OR_{YZ(X)} = \frac{m_{ijk}m_{ij'k'}}{m_{ij'k}m_{ijk'}},$$

jsou rovny jedné.

Uvažujme nyní model homogenní asociace (XY, YZ, XZ). Použijeme interakcí 1. řádu k popisu podmíněného poměru šancí. Nejdříve si uvědomme, že podmíněná závislost (asociace) znaků X a Y při pevně dané k-té hladině Z je popsána $(I-1)(J-1)$ lokálními podmíněnými poměry šancí. Takže pro jakoukoli dvourozměrnou tabulku $I \times J$ v $k-$ té hladině je poměr šancí

$$OR_{XY(Z)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}}, \tag{79}$$

pro $i = 1, ..., I-1$ a $j = 1, ..., J-1$.

Podobně popisuje $(I-1)(K-1)$ poměrů šancí podmíněnou závislost (asociaci) XZ při pevném j a $(J-1)(K-1)$ poměrů šancí podmíněnou závislost (asociaci) YZ při pevném i.

Zpět tedy k myšlence, že interakce 1. řádu odpovídají podmíněnému poměru šancí nezávislému na $k-$té hladině znaku Z, což ukazuje následující odvození. Ve výpočtu využijeme toho, že poměr šancí počítán z pravděpodobností je naprosto totožný s poměrem šancí počítaným z očekávaných četností, neboť platí $m_{ijk} = n\pi_{ijk}$ a tedy všechny n se zkrátí. Logaritmováním vztahu (79) získáme

$$\ln(OR_{XY(Z)}) = \ln \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}} = \ln \frac{m_{ijk}m_{i+1,j+1,k}}{m_{i,j+1,k}m_{i+1,j,k}} =$$

$$= \ln m_{ijk} + \ln m_{i+1,j+1,k} - \ln m_{i,j+1,k} - \ln m_{i+1,j,k} =$$

$$= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} +$$

$$+ \lambda + \lambda_{i+1}^X + \lambda_{j+1}^Y + \lambda_k^Z + \lambda_{i+1,j+1}^{XY} + \lambda_{j+1,k}^{YZ} + \lambda_{i+1,k}^{XZ} -$$

$$- \lambda - \lambda_i^X - \lambda_{j+1}^Y - \lambda_k^Z - \lambda_{i,j+1}^{XY} - \lambda_{j+1,k}^{YZ} - \lambda_{i,k}^{XZ} -$$

$$- \lambda - \lambda_{i+1}^X - \lambda_j^Y - \lambda_k^Z - \lambda_{i+1,j}^{XY} - \lambda_{j,k}^{YZ} - \lambda_{i+1,k}^{XZ} =$$

$$= \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}$$

nebo obecněji

$$\ln OR_{XY(Z)} = \ln \frac{m_{ijk}m_{i',j',k}}{m_{i,j',k}m_{i',j,k}} = \lambda_{ij}^{XY} + \lambda_{i',j'}^{XY} - \lambda_{i,j'}^{XY} - \lambda_{i',j}^{XY}. \tag{80}$$

Vztah (80) opět nezávisí na volbě hladiny, ale pouze na volbě řádků i, i' a sloupců j a j'.

Absence tří faktorových interakcí určuje ekvivalenci modelu homogenní asociace s

$$OR_{XY(1)} = OR_{XY(2)} = \cdots = OR_{XY(K)},$$

což platí pro všechny kategorie znaků X a Y a také ekvivalenci s

$$OR_{X(1)Z} = OR_{X(2)Z} = \cdots = OR_{X(J)Z},$$

což platí pro všechny kategorie X a Z a ekvivalenci s

$$OR_{(1)YZ} = OR_{(2)YZ} = \cdots = OR_{(I)YZ}$$

pro všechny kategorie znaků Y a $.Z$

Pokud budeme uvažovat saturovaný model, tedy kompletní model včetně interakcí druhého řádu, pak právě interakce druhého řádu jsou interpretovatelné pomocí poměru poměrů šancí ve vybraných řádcích (resp. vybereme $i-$tý řádek a tím je již vybraný $(i+1)-$ní řádek), sloupcích a hladinách. Tří faktorová interakce tedy popisuje, jak se poměr šancí dvou znaků mění v kategoriích třetího

znaku. Napíšeme-li si jak vypadá poměr poměrů šancí vůči dvěma kategoriím znaku Z, pak se po rozepsání očekávaných četností pomocí parametrů logline-árního saturovaného modelu odečtou všechny interakce 1. řádu, hlavní efekty i celkové průměry. K interpretaci pak zůstávají pouze interakce 2. řádu.

$$\begin{aligned}
\ln \frac{OR_{XY(Z=k)}}{OR_{XY(Z=k+1)}} &= \ln \frac{m_{ijk}m_{i+1,j+1,k}}{m_{i,j+1,k}m_{i+1,j,k}} \Big/ \frac{m_{i,j,k+1}m_{i+1,j+1,k+1}}{m_{i,j+1,k+1}m_{i+1,j,k+1}} = \\
&= \ln m_{ijk} + \ln m_{i+1,j+1,k} - \ln m_{i,j+1,k} - \ln m_{i+1,j,k} - \ln m_{i,j,k+1} - \\
&\quad - \ln m_{i+1,j+1,k+1} + \ln m_{i,j+1,k+1} + \ln m_{i+1,j,k+1} = \\
&= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} + \\
&\quad + \lambda + \lambda_{i+1}^X + \lambda_{j+1}^Y + \lambda_k^Z + \lambda_{i+1,j+1}^{XY} + \lambda_{i+1,k}^{XZ} + \lambda_{j+1,k}^{YZ} + \lambda_{i+1,j+1,k}^{XYZ} - \\
&\quad - \lambda - \lambda_i^X - \lambda_{j+1}^Y - \lambda_k^Z - \lambda_{i,j+1}^{XY} - \lambda_{ik}^{XZ} - \lambda_{j+1,k}^{YZ} - \lambda_{i,j+1,k}^{XYZ} - \\
&\quad - \lambda - \lambda_{i+1}^X - \lambda_j^Y - \lambda_k^Z - \lambda_{i+1,j}^{XY} - \lambda_{i+1,k}^{XZ} - \lambda_{jk}^{YZ} - \lambda_{i+1,j,k}^{XYZ} + \\
&\quad + \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{k+1}^Z + \lambda_{ij}^{XY} + \lambda_{i,k+1}^{XZ} + \lambda_{j,k+1}^{YZ} + \lambda_{i,j,k+1}^{XYZ} + \\
&\quad + \lambda + \lambda_{i+1}^X + \lambda_{j+1}^Y + \lambda_{k+1}^Z + \lambda_{i+1,j+1}^{XY} + \lambda_{i+1,k+1}^{XZ} + \lambda_{j+1,k+1}^{YZ} + \lambda_{i+1,j+1,k}^{XYZ} - \\
&\quad - \lambda - \lambda_i^X - \lambda_{j+1}^Y - \lambda_{k+1}^Z - \lambda_{i,j+1}^{XY} - \lambda_{i,k+1}^{XZ} - \lambda_{j+1,k+1}^{YZ} - \lambda_{i,j+1,k+1}^{XYZ} - \\
&\quad - \lambda - \lambda_{i+1}^X - \lambda_j^Y - \lambda_{k+1}^Z - \lambda_{i+1,j}^{XY} - \lambda_{i+1,k+1}^{XZ} - \lambda_{j,k+1}^{YZ} - \lambda_{i+1,j,k+1}^{XYZ} = \\
&= \lambda_{ijk}^{XYZ} + \lambda_{i+1,j+1,k}^{XYZ} - \lambda_{i,j+1,k}^{XYZ} - \lambda_{i+1,j,k}^{XYZ} + \lambda_{i,j,k+1}^{XYZ} + \lambda_{i+1,j+1,k}^{XYZ} - \lambda_{i,j+1,k+1}^{XYZ} - \lambda_{i+1,j,k+1}^{XYZ}.
\end{aligned}$$

Nyní lze ilustrativně všechny „plus-první“ kategorie znaků položit nule, čímž zbyde pouze jediný neredundantní (tj. nezbytný) parametr λ_{ijk}^{XYZ}. Zajímavé je, že v případě, kdy bude tento parametr nulový, získáme model homogenní asociace znaků X a Y, neboť $OR_{XY(Z=k)} = OR_{XY(Z=k+1)}$. [1, 14, 42]

7 Výzkumná část

V této části se budeme zabývat závislostmi mezi braním drog, kouřením a dalšími faktory. Cílem bude najít loglineární modely, které budou vhodně reprezentovat data z připraveného dotazníku. V získaných modelech pak bude možné najít vztahy mezi modelovanými veličinami a díky nim bude možné spočítat vhodné šance na závislosti vyplývající z loglineární analýzy. Data budou zpracována pomocí tří statistických softwarů, SAS, STATISTICA a R. Řešení v uvedených softwarech bude rozčleněno do jednotlivých kapitol. V závěru poslední kapitoly bude uvedeno srovnání těchto tří softwarů.

Dosud jsme se zabývali pouze teoretickou částí, potřebnou k pochopení dané problematiky, což nám nyní pomůže při řešení konkrétních situací. Data sesbíraná pomocí vlastního anonymního dotazníku (viz příloha) zkusíme analyzovat podle uvedené teorie. Dotazník obsahoval 17 otázek, z toho 15 otázek, kdy se dala zvolit pouze jedna možnost z nabízených. Dále 2 otázky, kde bylo možno vybrat odpovědí více. Všechny otázky byly cíleně voleny jako kategoriální proměnné s několika možnými odpověďmi. Na otázky nebylo možné odpovídat vlastními slovy. Otázky obsahovaly od dvou do deseti možných odpovědí (kategorií). Dotazník se velmi obecně zaměřoval na popis respondentů.

Celkový počet respondentů byl 340, přestože oslovených lidí bylo mnohem více. Číslo 340 je zcela náhodné a nebylo dopředu známo. Dotazník byl vystaven na internetu a odkaz na něj dostali jen vybraní lidé. Vybraní respondenti se jevili jako dostatečně variabilní vzorek populace. Byli osloveni lidé obou pohlaví (58,83% ženy), všech věkových kategorií (2,67% nad 61 let, 5,73% do 18 let), z různě velkých měst a vesnic a různých oblastí České Republiky (2 respondenti byli slovenské národnosti žijící v ČR). Respondenti pocházeli ze všech krajů České Republiky s výjimkou tří, Libereckého, Plzeňského a Karlovarského. Mezi respondenty byli pracující lidé (58,83%), studenti (34,12%), důchodci (3,53%) i lidé evidovaní na pracovním úřadě (3,24%). Respondenti byli vybíráni i dle vzdělání. Nejčetnější skupinou bylo středoškolské s maturitou (35,99%). Vzhledem k tomu, že původně vybraní lidé se vzděláním doktorským a vyšším byli pouze přírodovědného a medicínského směru, byli dále osloveni i odborní pracovníci, docenti a profesoři na vysokých školách s humanitním zaměřením.

Z dotazníku se nyní budeme snažit najít odpověď na často polemizovanou

otázku spojitosti alkoholu, kouření a drog a jejich souvislostech s dalšími informacemi o respondentech. Již ze surových dat však bylo zřejmé, že otázku „Pití alkoholu“ budeme muset z dotazníku vyřadit, případně ji použít pouze pro určité typy otázek. Na otázku „Pití alkoholu“ odpověděla převážná většina respondentů (85,3%) *příležitostně*. Zbylých (14,8%) pokrývalo další tři možné odpovědi a jeden respondent na otázku alkoholu neodpověděl, což je příliš málo dat pro analýzu trojrozměrných kontingenčních tabulek. Mnoho četností v buňkách tabulek by byly nulové. Ani při seskupení skupin *příležitostně a často* a skupin *vůbec a dříve ano, nyní ne* se situace nezlepšila. Stále bylo mnoho četností nulových a navíc přímo v kombinaci „alkohol drogy kouření“ byla i jedna marginální četnost nulová, což už je zásadní problém při modelování.

Zaměřme se tedy na vztah mezi kouřením, zkušeností s drogami a všemi dalšími otázkami z dotazníku. Po vynechání otázky „Pití alkoholu.“ zbylo 14 otázek a tedy 14 kategoriálních znaků. Vytvořili jsme 14 loglineárních modelů, kde vždy dva znaky byly „drogy“ a „kouření“ a třetí znak byl variabilní. Pro jednotlivé trojice byly vždy sestaveny všechny hierarchické modely uvedené v teoretické části v kapitole 2.8. Způsobů, jak se dopracovat k tzv. nejlepšímu modelu, je více. Je možné začít modelem úplné nezávislosti a postupně přidávat interakce. Možný je i opačný postup, ze saturovaného modelu postupně interakce ubírat. Třetí možností je vycházet z modelu, kde předpokládáme určitý vztah mezi znaky. Do modelu pak postupně přidáváme nebo z něj ubíráme parametry. Výsledné modely je vždy nutné porovnávat rozdílem G^2 nebo informačními kritériemi (v kapitole 5.2). Pomoci rozhodnout se pro určitý model nám však může i významnost interakcí a hlavních efektů nebo tabulka reziduí. Vzhledem k tomu, že se neuvádí, který postup je pro hledání vhodného modelu lepší, nechali jsme vytvořit všechny modely, jež jsme následně mezi sebou porovnali a vybrali nejvhodnější. Z důvodu málo početných skupin a snahy vyhnout se nulovým četnostem, byla většina kategorií znaků (otázek) seskupena do dvou, tří nebo čtyř skupin. Pět skupin bylo použito pro otázku „Bydliště do 18 let“. Většina modelů vyšla jako nehierarchické kvůli nevýznamnosti hlavního efektu kouření. Přesto byl efekt kouření zahrnut do modelů a modely se tak staly hierarchické. Stalo se tak z důvodu významnosti interakcí s kouřením. Ve všech modelech se projevila jako významná interakce kouření s drogami. Jediná interakce druhého řádu, která vyšla významná, je v modelu, kde bylo bráno v úvahu, zda člověk má děti. Model byl přesto zredukován na model podmíněné nezávislosti, neboť saturovaný model není pro modelování

vhodný. Potřebujeme data popsat jednodušším modelem a model podmíněné nezávislosti jsme zvolili proto, že vyšel významnější než model párové závislosti.

K analýze uvedeného dotazníku bylo použito statistických softwarů STATISTICA, SAS a R. Přehled vybraných modelů i s odhady jejich parametrů je uveden v příloze. Jak jsme si již uvedli, ve všech modelech se projevila vazba mezi braním drog (zkušeností s drogami) a kouřením cigaret. Další významné interakce s kouřením se objevily v souvislosti se vzděláním, věkem, zaměstnáním, rodinným stavem, pitím kávy a známkami z matematiky a českého jazyka. V případě vlivu vzdělání na kouření a zkušenost s drogou byla prokázána párová závislost. Stejně tak s věkem. Tedy ovlivňuje se kouření a braní drog, kouření a vzdělání (resp. věk) a v neposlední řadě vzdělání (resp. věk) a braní drog. Zdrojové kódy a výstupy jsou uvedeny v příloze. Zde si podrobněji zanalyzujeme pouze jeden loglineární model, ukážeme si a popíšeme zdrojový kód a vysvětlíme si výstupy ze softwaru SAS. Následně se podíváme na řešení v softwaru STATISTICA a nakonec na řešení v statistickém programu R.

Pomocí softwaru STATISTICA byly překódovány kategorie, následně sestaveny trojrozměrné tabulky a pomocí softwaru Excel byly tabulky upraveny do podoby, kterou zná software SAS. Kdybychom nechtěli používat softwaru STATISTICA, šlo by překódovat kategorie i v Excelu nebo v prostředí SAS Enteprice Guide. STATISTICA je v tomhle ohledu ale nejšikovnější a nejrychlejší.

7.1 Software SAS 9.3

Zdrojový kód pro SAS si rozeberme podrobněji pro trojici „drogy, kouření, zvíře“. Znakem „zvíře“ je myšlena otázka, zda respondent vlastní domácí zvíře. Odpověď na otázku bylo možno vybrat ze sedmi odpovědí, případně i jejich kombinací. K analýze však bylo potřeba kategorie seskupit z důvodu málopočetných skupin a získali jsme pouze dvě kategorie, tj. osoba vlastní zvíře, ať venku nebo doma a osoba nevlastní žádné zvíře. Kategorie pro znak „zvíře“ jsou tedy ANO/NE. Ve znacích „drogy“ a „kouření“ došlo taktéž k seskupení kategorií do dvou. Kategorie ANO obsahuje jakoukoli zkušenost s drogou resp. kuřák, odpověď NE značí, že člověk nebere drogy ani příležitostně, resp. nekouří a nikdy nekouřil. Získáme trojrozměrnou tabulku $2 \times 2 \times 2$ celkem o 8 buňkách. Teoreticky vychází na jednu buňku přibližně 42 jedinců, což splňuje doporučené kritérium alespoň 5 jedinců na buňku kontingenční tabulky. Reálné napozorované četnosti tomu pak odpovídají.

Data znázorňující počty osob, u kterých se sledovalo, zda kouří, zda mají zkušenost s drogami a zda vlastní domácí zvíře, najdeme v tabulce 10.

zvíře	kouření	drogy ano	drogy ne
ne	ano	**27**	**15**
ne	ne	**26**	**45**
		drogy ano	drogy ne
ano	ano	**45**	**49**
ano	ne	**28**	**88**

Tabulka 10: Kontingenční tabulka

Nyní blíže ke zdrojovému kódu. Software SAS nerozlišuje velká a malá písmenka a je tedy na uživateli, zda si příkazy bude chtít zvýrazňovat velkými písmeny. Po prvotním načtení dat lze rovnou psát příkaz na konstrukci loglineárního modelu. Načtení dat je možné provést několika způsoby. V této práci byla zvolena metoda přímého vepsání jednotlivých kombinací kategorií znaků kouření, drogy a zvíře a četnostmi. SAS používá pro loglineární analýzu proceduru CATMOD nebo GENMOD. Proceduru CATMOD si popíšeme důkladně a nakonec přidáme i proceduru GENMOD, která nám pomůže získat další informace o loglineárním modelu. Procedura CATMOD se používá nejen pro loglineární modely. Je proto nutné SASu říci, který model budeme používat. Příkaz pro použití loglineárního modelu se provede přidáním `_RESPONSE_` za definováním vstupních proměnných. Konkrétní model se pak volí pomocí `LOGLIN` a následným výpisem efektů a interakcí. Interakce se značí symbolem | . Seznam zadání interakcí do SASu najdeme v tabulce 11.

typ loglineárního modelu	**zápis**	**v programu**	**SAS**
saturovaný model		X|Y|Z	
model párové závislosti		X|Y X|Z Y|Z	
model podmíněné nezávislosti	X|Y X|Z	resp. X|Y Y|Z	resp. X|Z Y|Z
model sdružené nezávislosti	X Y|Z	resp. X|Y Z	resp. X|Z Y
model úplné nezávislosti		X Y Z	

Tabulka 11: Syntaxe typu loglineárního modelu

Procedura CATMOD povoluje několik volitelných parametrů, kterými si zjednodušíme práci. Pokud jsou v kontingenční tabulce nuly, použijeme parametr `ZERO=sampling` nebo `ZERO=structural` podle toho, jestli jsou nuly důsledkem výběrovým, tj. když nám nikdo nespadá do dané kategorie, ale určitě v populaci

takoví jedinci existují, nebo důsledkem strukturálním, tj. že taková varianta kategorií nemůže nastat. Strukturální nulou může být například při analyzování typu rakoviny, že rakovinu prostaty nemůže dostat žena. Parametr `ZERO=sampling` jsme využili v modelu `alkohol drogy kouření.` Parametr `ZERO=structural` jsme nevyužili v žádném modelu. Dalším parametrem je volba iterační metody. Implicitně SAS používá metodu Newton-Raphson, ale pokud chceme raději algoritmus IPF, změníme jej příkazem `ML=IPF.`

Dalším volitelným parametrem, který jsme využili je `NOPROFILE`, což zabraňuje neustálému vypisování všech možných variant kategorií a počtu pozorování. Profil dat jsme nechali vypsat pouze u prvního, tj. saturovaného, modelu. Ve všech modelech jsem použili `P=freq,` což vypíše pozorované a očekávané četnosti včetně příslušných směrodatných odchylek od pozorovaných i očekávaných četností a v posledním sloupci uvede výpis reziduí očekávaných a pozorovaných četností. Pokud bychom chtěli znát pravděpodobnost, s jakou se jedinec může ocitnout v konkrétní buňce, použili bychom příkazu `P=prob.` Z těchto pravděpodobností se pak jednoduše přenásobením celkového počtu respondentů získají očekávané četnosti.

Datový soubor ve tvaru kontingenční tabulky se do SASu vloží následovně. Nejdříve datový soubor pojmenujeme pomocí příkazu `DATA`, poté nadefinujeme názvy znaků vstupujících do analýzy pomocí příkazu `INPUT`. Nakonec musíme zadat styl zadávaných dat, tj. jak má SAS rozeznat jednotlivé řádky naší kontingenční tabulky. Pro zadávání dat po řádcích použijeme příkaz `DATALINES`. Dále se vypíše seznam všech možných kombinací kategorií vstupních znaků a jejich napozorovaných četností. Zdrojový kód tedy vypadá následovně.

```
% pojmenujeme dataset
DATA zvíře;
% vložíme názvy proměnných v tabulce a název pro četnosti
INPUT drogy $ zvíře $ kouření $ počet;

% vložíme kontingenční tabulku s četnostmi
DATALINES;
ano ne  ano 27
ano ne  ne 26
ano ano ano 45
ano ano ne 28
ne ne  ano 15
```

```
ne ne  ne 45
ne ano ano 49
ne ano ne 88
;
```

Nejdříve sestavíme saturovaný model a model úplné nezávislosti (v kapitole 2).

```
% sestavení saturovaného modelu

% použijeme proceduru CATMOD
PROC CATMOD ORDER=data;

% při použití WEIGHT se proměnná "počet" použije jako četnosti
WEIGHT počet;

% sestavíme model ze tří proměnných
% _RESPONSE_ určí, že půjde o loglineární model
MODEL drogy * zvíře * kouření = _response_

% P vypíše očekávané a modelové četnosti
P=freq;

% určíme typ modelu (saturovaný)
LOGLIN drogy|zvíře|kouření;
run;

% model úplné nezávislosti
PROC CATMOD order=data;
WEIGHT počet;
MODEL drogy * zvíře * kouření = _RESPONSE_

% NOPROFILE  procedura nevypíše profil dat
/ NOPROFILE P=freq;
LOGLIN drogy zvíře kouření;
run;
```

V dalším kroku sestavíme modely podmíněné a sdružené nezávislosti a model párové závislosti (v kapitole 2). Syntaxe modelu je naprosto analogická, jako při sestavování modelů saturovaného a modelu úplné nezávislosti. Mění se jen typ modelu dle níže uvedených možností.

```
% sestavení modelů sdružené nezávislosti
LOGLIN drogy|kouření zvíře;
LOGLIN drogy|zvíře kouření;
LOGLIN drogy zvíře|kouření;

% sestavení modelů podmíněné nezávislosti
LOGLIN drogy|zvíře zvíře|kouření;
LOGLIN drogy|zvíře drogy|kouření;
LOGLIN zvíře|kouření drogy|kouření;

% sestavení modelu párové závislosti
LOGLIN drogy|zvíře zvíře|kouření drogy|kouření;
```

Nyní si popišme jednotlivé výstupy ze SASu při použití procedury CATMOD.

Z výstupu na obrázku 4 vidíme, že do analýzy vstupuje 323 osob a existuje 8 možností, jak tyto osoby klasifikovat dle znaků „drogy, pohlaví a zvíře“. Ve výstupu na obrázku 5 vidíme významnosti parametrů vyskytujících se v saturovaném modelu. Ve sloupci nazvaném *Source* najdeme vstupní parametry, ve sloupci *DF* najdeme počet stupňů volnosti, ve sloupci *Chi-Square* najdeme Waldovu testovou statistiku (viz například [1]), která testuje nulovost parametru v modelu. V posledním sloupci najdeme příslušnou p-hodnotu. V případě p-hodnoty menší než zvolená hladina významnosti (např. 0,05) je parametr významný a je tedy vhodné, aby v modelu zůstal, neboť jej statisticky významně ovlivňuje. Na posledním řádku najdeme Likelihood Ratio, což je analogické devianci G^2 (v kapitole 5.1). LR udává hodnotu testu věrohodnostním poměrem, kdy porovnáváme daný model s modelem saturovaným. Z toho důvodu není pro saturovaný model hodnota LR nikdy ve výstupu uvedena, neboť nelze porovnat saturovaný model se saturovaným modelem. Hodnota LR příslušného modelu je vždy vůči saturovanému modelu.

Ve výstupu na obrázku 6 vidíme významnosti parametrů vyskytujících se v modelu párové závislosti (Chí-kvadrát statistikou je Waldova statistika) a ve výstupu na obrázku 7 vidíme významnosti parametrů v modelu podmíněné nezávislosti (Chí-kvadrát statistikou je opět Waldova statistika). Model vhodně modeluje data, pokud je p-hodnota u testu LR větší než zvolená hladina významnosti. Jak je vidět, oba modely jak párové závislosti, tak podmíněné nezávislosti,

Data Summary			
Response	drogy*zvire*koureni	Response Levels	8
Weight Variable	pocet	Populations	1
Data Set	ZVIRE	Total Frequency	323
Frequency Missing	0	Observations	8

Population Profiles	
Sample	Sample Size
1	323

Response Profiles			
Response	drogy	zvire	koureni
1	ano	ne	ano
2	ano	ne	ne
3	ano	ano	ano
4	ano	ano	ne
5	ne	ne	ano
6	ne	ne	ne
7	ne	ano	ano
8	ne	ano	ne

Obrázek 4: Profil dat

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
drogy	1	5.58	0.0181
zvire	1	23.42	<.0001
drogy*zvire	1	6.34	0.0118
koureni	1	5.41	0.0201
drogy*koureni	1	18.98	<.0001
zvire*koureni	1	3.55	0.0595
drogy*zvire*koureni	1	0.02	0.8796
Likelihood Ratio	0	.	.

Obrázek 5: Saturovaný model

D K Z	DK KZ DZ	DZ ZK	**DZ DK**	ZK DK	DK Z	DZ K	ZK D
AICp	AICp	AICp	AICp	AICp	AICp	AICp	AICp
19,4	-1,98	17,13	**-0,3**	2,47	2,22	16,88	19,65
Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
2,8E-08	2,5E-08	4,6E-08	**0**	2,5E-08	0	1,2E-07	4,2E-08
BICp	BICp	BICp	BICp	BICp	BICp	BICp	BICp
4,29	-5,76	9,58	**-7,86**	-5,09	-9,11	5,55	8,32
LR	LR	LR	LR	LR	LR	LR	LR
27,4	0,02	21,13	**3,7**	6,47	8,22	22,88	25,65
$Adj.R^2$	$Adj.R^2$	$Adj.R^2$	$Adj.R^2$	$Adj.R^2$	$Adj.R^2$	$Adj.R^2$	$Adj.R^2$
0	0,99	0,49	**0,91**	0,84	0,76	0,33	0,25

Tabulka 12: Porovnávací kritéria pro modely vztahů drogy, zvíře a kouření

vhodně datům odpovídají. Je ale nutné zvolit pouze jeden model. Přistoupíme tedy k porovnávání modelů. Nejlepším modelem popisujícím data o kouření, drogách a zvířeti se jeví model podmíněné nezávislosti znaků „zvíře a kouření“. Je to zřejmé z následného porovnání hodnot LR.

Chceme-li zjistit, zda je model párové závislosti lepší než model saturovaný, podíváme se na hodnotu LR v modelu párové závislosti (LR = 0,02) a porovnáme s kvantilem Chí-kvadrát rozdělení o $1 - 0$ stupních volnosti DF (tj. 3,842). Hodnota LR je menší než 3.842 a tudíž se model zlepšil odstraněním interakce druhého řádu. Dále srovnáme analogicky s modely podmíněné nezávislosti. Porovnáním modelu `drogy|zvíře drogy|kouření` s modelem párové závislosti získáme hodnotu rozdílů LR $3,7 - 0,02$, tj. 3,68, což je opět menší než hodnota 3,842 (rozdíl DF je 2-1) a tedy model se zlepšil odstraněním interakce `kouření|zvíře`. U dalších dvou modelů podmíněné nezávislosti se zlepšení neprojevilo. Výstupy je možné vidět v příloze, zde uvedeme pouze hodnoty rozdílů LR (bez interakce `drogy|kouření` 21,11 a bez interakce `drogy|zvíře` 6,45). Další zlepšení se opět nepotvrdilo, tedy nemá smysl odstraňovat další parametry z modelu. Rozdíly LR byly u všech modelů sdružené nezávislosti větší než kvantily Chí-kvadrát rozdělení. (v kapitole 5.2)

Podobného závěru je možné dosáhnout i po spočítání AICp pro jednotlivé modely. Nejnižší hodnoty AICp byly u modelu párové závislosti (-1,98) a u modelu

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
drogy	1	5.94	0.0148
zvire	1	24.84	<.0001
drogy*zvire	1	6.39	0.0115
koureni	1	5.44	0.0197
zvire*koureni	1	3.60	0.0576
drogy*koureni	1	20.36	<.0001
Likelihood Ratio	1	0.02	0.8795

Obrázek 6: Model párové závislosti

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
drogy	1	6.89	0.0086
zvire	1	23.22	<.0001
drogy*zvire	1	4.52	0.0335
koureni	1	3.55	0.0597
drogy*koureni	1	18.70	<.0001
Likelihood Ratio	2	3.70	0.1571

Obrázek 7: Model podmíněné nezávislosti

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
drogy	ano	-0.1608	0.0613	6.89	0.0086
zvire	ne	-0.2865	0.0594	23.22	<.0001
drogy*zvire	ano ne	0.1264	0.0594	4.52	0.0335
koureni	ano	-0.1109	0.0589	3.55	0.0597
drogy*koureni	ano ano	0.2548	0.0589	18.70	<.0001

Obrázek 8: Odhady parametrů v modelu podmíněné nezávislosti

Maximum Likelihood Predicted Values for Frequencies							
			Observed		Predicted		
drogy	zvire	koureni	Frequency	Standard Error	Frequency	Standard Error	Residual
ano	ne	ano	27	4.974237	30.28571	4.463856	-3.28571
ano	ne	ne	26	4.889491	22.71429	3.687413	3.285714
ano	ano	ano	45	6.223395	41.71429	5.367203	3.285714
ano	ano	ne	28	5.056951	31.28571	4.553612	-3.28571
ne	ne	ano	15	3.781984	19.49239	3.027256	-4.49239
ne	ne	ne	45	6.223395	40.50761	5.125987	4.492386
ne	ano	ano	49	6.447214	44.50761	5.405823	4.492386
ne	ano	ne	88	8.001548	92.49239	7.540241	-4.49239

Obrázek 9: Očekávané a pozorované četnosti, odchylky a rezidua v modelu podmíněné nezávislosti

podmíněné nezávislosti (-0,3), který jsme vybrali jako nejlepší, viz příloha. Hodnoty AICp a dalších statistik v tabulce 12 byly vypočítány v Excelu dle definice (v kapitole 5.2). Procedura CATMOD tento výstup neumožňuje.

Odhady parametrů pro model podmíněné nezávislosti „kouření“ a „zvířete“ najdeme ve výstupu na obrázku 8.

Ve výstupu na obrázku 9 vidíme očekávané i pozorované četnosti pro model podmíněné nezávislosti. Seznam očekávaných četností pro všechny typy modelů znaků „drogy, kouření, zvíře“ najdeme v tabulce 13, kdy pro jednoduchost značíme znak „drogy“ symbolem D, znak „kouření“ symbolem K a znak „zvíře“ symbolem Z.

Použijeme-li již zmíněného parametru `P=prob`, získáme přehled očekávaných pravděpodobností pro jednotlivé buňky v trojrozměrné kontingenční tabulce, viz

Maximum Likelihood Predicted Values for Probabilities							
			Observed		Predicted		
drogy	zvire	koureni	Probability	Standard Error	Probability	Standard Error	Residual
ano	ne	ano	0.0836	0.0154	0.0938	0.0138	-0.01
ano	ne	ne	0.0805	0.0151	0.0703	0.0114	0.0102
ano	ano	ano	0.1393	0.0193	0.1291	0.0166	0.0102
ano	ano	ne	0.0867	0.0157	0.0969	0.0141	-0.01
ne	ne	ano	0.0464	0.0117	0.0603	0.0094	-0.014
ne	ne	ne	0.1393	0.0193	0.1254	0.0159	0.0139
ne	ano	ano	0.1517	0.02	0.1378	0.0167	0.0139
ne	ano	ne	0.2724	0.0248	0.2864	0.0233	-0.014

Obrázek 10: Pravděpodobnosti v buňkách tabulky pro model podmíněné nezávislosti

výstup na obrázku 10. Z těchto očekávaných pravděpodobností lze jednoduše přenásobením celkového počtu respondentů (v tomto případě číslem 323) získat očekávané četnosti v buňkách kontingenční tabulky. Očekávané četnosti získané tímto postupem přesně odpovídají očekávaným četnostem ve výstupu na obrázku 9.

Z tabulky 13 lze odhadnout četnosti jednotlivých kategorií pro celou populaci České Republiky. Za předpokladu, že jsme dotazníkem získali dostatečně reprezentativní vzorek populace ČR a při uvážení, že ČR má přibližně 11000000 obyvatel (k 1.1.2012 je 10721315 obyvatel)[6], získáme celorepublikové odhady uvedené v tabulce 15. Tyto odhady spočítáme pouze pro model podmíněné nezávislosti (DZ, DK). Odhady očekávaných četností vztažené k ČR spočítáme podle „nesmrtelné“ trojčlenky, stejně tak odchylky očekávaných četností. Z tabulky 15 je vidět, že například lidí, kteří mají zkušenost s drogou, nekouří a vlastní domácí zvíře, je v České Republice odhadováno na 1 005000. Celkově například zarytých nekuřáků by mělo být $1005000 + 754000 + 647000 + 1345000 = 3751000$. Analogicky se dají spočítat počty lidí majících stejnou vlastnost (třeba právě zálibu v kouření). Pokud si však dohledáme počet kuřáků v ČR, zjistíme počet 2300000.[7] Dle úvahy by tedy mělo být $11000000 - 230000 = 10770000$ nekuřáků. Odlišnost by mohla být způsobena odchylkou očekávaných četností u nekuřáků, která však

[6] http://www.mvcr.cz/clanek/statistiky-pocty-obyvatel-v-obcich.aspx

[7] http://www.kurakovaplice.cz/kouření_cigaret/zajimavosti-a-statistiky/statistiky-tykajici-se-kouření/10-statistiky-tykajici-se-kouření-cigaret.html

D	K	Z	n_{ijk}	DKZ	DK KZ DZ	DZ KZ	DZ DK
ano	ne	ano	27	27	26,699	19,699	30,286
ano	ne	ne	26	26	26,301	33,301	22,714
ano	ano	ano	45	45	45,301	32,676	41,714
ano	ano	ne	28	28	27,699	40,324	31,286
ne	ne	ano	15	15	15,301	22,301	19,492
ne	ne	ne	45	45	44,699	37,699	40,508
ne	ano	ano	49	49	48,699	61,324	44,508
ne	ano	ne	88	88	88,301	75,676	92,492

D	K	Z	DK KZ	KZ D	DZ K	DK Z	D K Z
ano	ne	ano	22,235	16,384	22,316	25,189	18,560
ano	ne	ne	20,503	27,697	30,684	18,892	25,520
ano	ano	ano	49,765	36,669	30,737	46,811	34,492
ano	ano	ne	33,498	45,251	42,263	35,108	47,427
ne	ne	ano	19,765	25,616	25,263	22,390	29,019
ne	ne	ne	50,497	43,303	34,737	46,529	39,901
ne	ano	ano	44,235	57,331	57,684	41,60	53,929
ne	ano	ne	82,503	70,749	79,316	86,471	74,152

Tabulka 13: Očekávané četnosti v jednotlivých modelech

dá v těhle číslech minimální rozdíl. Když přičteme k našemu odhadu přibližně půlmilionu lidí (součet kladných odchylek u počtu nekuřáků), ani se nepřiblížíme oficiálně odhadovanému počtu nekuřáků, nehledě na nesmyslnost pouhého připočítávání maximálních odchylek. Pokud si ale uvědomíme, že v 11000000 obyvatel jsou i děti, které jsme do dotazníku nezahrnovali, pak po odečtení dětí do 15 let, získáme $11000000 - 1615000$[8] $= 9385000$ osob. I tak se však odhad příliš nezlepší. Je to nejspíše z důvodu různého chápání kuřáků a nekuřáků. V našem dotazníku je nekuřák člověk, který nikdy neměl cigaretu v ústech nebo to o sobě aspoň tvrdí. V oficiálních statistikách je definován kuřák nejspíš odlišně.

V tabulce 14 je model podmíněné nezávislosti znaků kouření a zvířete rozepsán pro jednotlivé očekávané četnosti. Použité hodnoty najdeme ve výstupech obrázcích 8 a 9. Hodnota „mean“ je vypočítána jako průměr z přirozených logaritmů očekávaných četností.

Dle výše uvedených výsledků by se dalo usuzovat, že to zda člověk kouří,

[8] přesně je dětí do 15 let v ČR 1615100
http://www.czso.cz/csu/redakce.nsf/i/vekova_skladba_obyvatelstva_cr

ln (30,28571) = 3,587747 + (-0,1608) + (-0,2865) + (-0,1109) + 0,1264 + 0,2548
mean + D ano + Z ne + K ano + DZ ano ne + DK ano ano

ln (22,71429) = 3,587747 + (-0,1608) + (-0,2865) + 0,1109 + 0,1264 + (-0,2548)
mean + D ano + Z ne + K ne + DZ ano ne + DK ano ne

ln (41,71429) = 3,587747 + (-0,1608) + 0,2865 + (-0,1109) + (-0,1264) + 0,2548
mean + D ano + Z ano + K ano + DZ ano ano + DK ano ano

ln (31,28571) = 3,587747 + (-0,1608) + 0,2865 + 0,1109 +(-0,1264) + (-0,2548)
mean + D ano + Z ano + K ne + DZ ano ano + DK ano ne

ln (19,49239) = 3,587747 + 0,1608 + (-0,2865) + (-0,1109) + (-0,1264) + (-0,2548)
mean + D ne + Z ne + K ano + DZ ne ne + DK ne ano

ln (40,50761) = 3,587747 + 0,1608 + (-0,2865) + 0,1109 + (-0,1264) + 0,2548
mean + D ne + Z ne + K ne + DZ ne ne + DK ne ne

ln (44,50761) =3,587747 + 0,1608 + 0,2865 + (-0,1109) + 0,1264 + (-0,2548)
mean + D ne + Z ano + K ano + DZ ne ano + DK ne ano

ln (92,49239) =3,587747 + 0,1608 + 0,2865 + 0,1109 + 0,1264 + 0,2548
mean + D ne + Z ano + K ne + DZ ne ano + DK ne ne

Tabulka 14: Výpočet očekávaných četností

D	K	Z	n_{ijk}	DZ DK	%	celorepublikový odhad
ano	ne	ano	27	30,29	9,38	1005000
ano	ne	ne	26	22,71	7,03	754000
ano	ano	ano	45	41,71	12,91	1385000
ano	ano	ne	28	31,29	9,69	1039000
ne	ne	ano	15	19,49	6,03	647000
ne	ne	ne	45	40,51	12,54	1345000
ne	ano	ano	49	44,51	13,78	1477000
ne	ano	ne	88	92,49	28,64	3070000

Tabulka 15: Celorepublikové očekávané četnosti v modelu (DZ, DK)

nemá vliv na to, zda vlastní nějaké zvíře a naopak. Jinými slovy zvířata vlastní jak kuřáci, tak nekuřáci. Souvisí však spolu to, zda člověk kouří nebo bere drogy a to, zda bere drogy a vlastní nějaké zvíře. Jestli je větší šance u kouřícího člověka

na braní drog nebo naopak, zjistíme z poměrů šancí (v kapitole 6.1). Stejně tak u vztahu „drogy“ a „zvíře“.

Z tabulky 10 zjistíme, že u kuřáků je šance, že budou brát drogy $\frac{(27+45)}{(15+49)} = 1,125$. Šance, že nekuřáci budou brát drogy je $\frac{(26+28)}{(45+88)} = 0,406$. Poměr šancí $OR = \frac{(27+45)\cdot(45+88)}{(26+28)\cdot(15+49)} = 1,125/0,406 = 2,771$. Pokud tedy budu mít například syna, který kouří, bude mít 2,771 krát větší šanci na to, že bude brát i drogy (alespoň příležitostně) oproti tomu, kdyby nekouřil vůbec.

Podobně můžeme spočítat šanci na braní drog, když člověk vlastní zvíře, oproti tomu, že zvíře nevlastní. Tedy šance, že člověk, který vlastní zvíře, bude brát drogy, je $\frac{45+28}{49+88} = 0,533$. Šance, že člověk, který žádné zvíře nemá, bude brát drogy, je $\frac{27+26}{15+45} = 0,883$. Tedy $OR = \frac{0,533}{0,883} = 0,603$. Tedy šance, že můj imaginární syn bude brát drogy, je 0.603 krát menší, když mu pořídím domácí zvíře. Z obrácené hodnoty OR lze spočítat šance, že bude brát drogy, když nebude mít žádné domácí zvíře, tj. $\frac{1}{0,603} = 1,658$. Z čehož plyne, že ta šance je více než jeden a půl krát větší.

Závěrem si stručně shrňme, které znaky se vzájemně ovlivňují. Prvním důležitým poznatkem je, že se významně ovlivňují znaky „drogy“ a „kouření.“ Kuřáci častěji propadnou drogám (zkouší drogy) než nekuřáci. Dalšími významnými faktory jsou vzdělání a věk. Oba jsou závislé na faktorech drogy i kouření. Faktory jako jsou hra na hudební nástroj, sourozenci a bydliště do 18 let nejsou přínosnými faktory. Faktor kouření má významnou vazbu na faktory pití kávy a známky z matematiky a českého jazyka. Naopak s drogami úzce souvisí rodinný stav, bezdětnost (resp. počet dětí), pohlaví a bydliště respondentů a to, zda člověk vlastní zvíře a jestli pracuje. Faktor alkohol je těžce interpretovatelný kvůli nevhodnému počtu respondentů (nuly v kontingenční tabulce), avšak jako významná se projevila vazba drog a alkoholu, což určitě není překvapením.

Ve stručnosti si zde ještě uveďme proceduru GENMOD, která nám oproti CATMOD vypočítá kritéria, na jejichž základě lze porovnat modely výše uvedené modely. Načtení dat se provede stejně jako v proceduře CATMOD. Hned po načtení dat následuje zdrojový kód. Použité příkazy jsou vysvětlené přímo ve zdrojovém kódu.

```
DATA zvíře;
```

```
INPUT drogy $ zvíře $ kouření $ počet;
DATALINES;
ano ne  ano 27
ano ne  ne 26
ano ano ano 45
ano ano ne 28
ne ne  ano 15
ne ne  ne 45
ne ano ano 49
ne ano ne 88
;

% sestavení modelu podmíněné nezávislosti
PROC GENMOD data=zvíře;

% CLASS nastaví proměnné jako kategoriální
% PARAM=effect použije kódování typu efekt
CLASS drogy zvíře kouření / PARAM=effect;

% MODEL určí model (nutné vypsat všechny parametry v modelu)
MODEL  počet = drogy zvíře kouření drogy*zvíře drogy*kouření

% DIST=poisson sestaví Poissonův model
% LINK=log nastaví link funkci na logaritmickou
% TYPE3 výpočítá významnosti pro každý parametr v modelu
% WALD souvisí s TYPE3, k tetsování použije Waldovu statistiku
/ DIST=poisson LINK=log TYPE3 WALD;
run;

% sestavení saturovaného modelu
PROC GENMOD data=zvíře;
CLASS drogy zvíře kouření
/ param=effect;
MODEL  počet = drogy kouření zvíře drogy*kouření kouření*zvíře
drogy*zvíře drogy*zvíře*kouření
/ dist=poisson link=log type3 wald;
run;
```

Ve výstupu na obrázku 11 vidíme již zmíněná porovnávací kritéria pro model podmíněné nezávislosti a pro saturovaný model totéž na obrázku 12. Před-

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	2	3.7012	1.8506
Scaled Deviance	2	3.7012	1.8506
Pearson Chi-Square	2	3.6408	1.8204
Scaled Pearson X2	2	3.6408	1.8204
Log Likelihood		910.0745	
Full Log Likelihood		-23.5116	
AIC (smaller is better)		59.0232	
AICC (smaller is better)		143.0232	
BIC (smaller is better)		59.4998	

Obrázek 11: Porovnávací kritéria v modelu podmíněné nezávislosti

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	.	0.0000	.
Scaled Pearson X2	.	0.0000	.
Log Likelihood		911.9250	
Full Log Likelihood		-21.6610	
AIC (smaller is better)		59.3220	
AICC (smaller is better)		.	
BIC (smaller is better)		59.9576	

Obrázek 12: Porovnávací kritéria v saturovaném modelu

stavu o vhodnosti modelu podmíněné nezávislosti si uděláme odečtením hodnoty AICp pro model podmíněné nezávislosti od modelu saturovaného, tj. $59,322 - 59,0232 \doteq 0,3$. Ke stejnému číslu jsme došli i počítáním AIC dle definice z teorie v kapitole (5.2). Ve výstupech vidíme hodnotu deviance, Pearsonovy Chí-kvadrát statistiky, logaritmickou věrohodnostní funkci, příslušné stupně volnosti a hodnoty kritérií Akaikeho, korigovaného Akaikeho a Bayessovo.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	3.5877	0.0613	3.4677	3.7078	3430.62	<.0001
drogy	ano		1	-0.1608	0.0613	-0.2809	-0.0408	6.89	0.0086
zvire	ano		1	0.2865	0.0594	0.1699	0.4030	23.22	<.0001
koureni	ano		1	-0.1109	0.0589	-0.2264	0.0045	3.55	0.0597
drogy*zvire	ano	ano	1	-0.1264	0.0594	-0.2429	-0.0099	4.52	0.0335
drogy*koureni	ano	ano	1	0.2548	0.0589	0.1393	0.3703	18.70	<.0001
Scale			0	1.0000	0.0000	1.0000	1.0000		

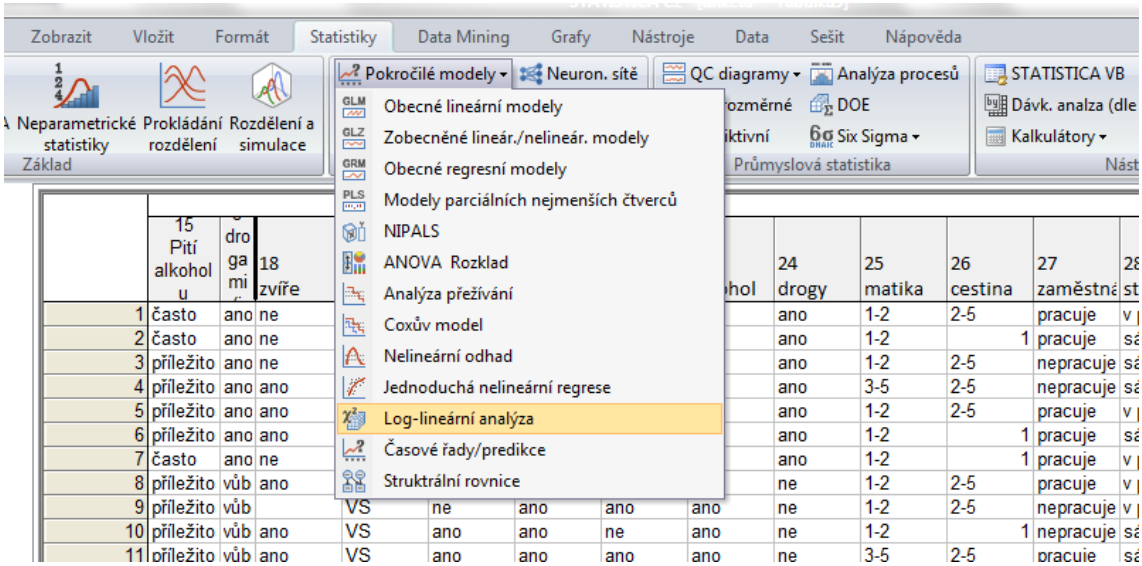
Obrázek 13: Odhady parametrů v modelu podmíněné nezávislosti

Wald Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
drogy	1	6.89	0.0086
zvire	1	23.22	<.0001
koureni	1	3.55	0.0597
drogy*zvire	1	4.52	0.0335
drogy*koureni	1	18.70	<.0001

Obrázek 14: Významnosti parametrů v modelu podmíněné nezávislosti

Ve výstupech na obrázku 13 vidíme odhady parametrů v modelu podmíněné nezávislosti, včetně chyb, Waldových testových statistik a jejich významnosti. Oproti proceduře CATMOD získáme i 95% intervaly spolehlivosti pro odhady parametrů. Procedura GENMOD navíc přímo vypíše odhad i pro celkový průměr a není tedy nutné jej vypočítávat z očekávaných četností.

Poslední výstup, který si rozebereme, je na obrázku 14 a týká se sasovských parametrů `TYPE3` a `WALD`, které spolu souvisí. Parametrem `TYPE3` určujeme, že chceme vypsat významnost každého parametru (připomeňme, že testujeme nulovost parametru a pokud je tedy p menší než zvolená hladina významnosti, pak je parametr významný, tudíž nenulový), který se v modelu vyskytuje, včetně stupňů volnosti a testových statistik. Parametrem `WALD` určíme, že požadujeme právě Waldovy statistiky. Protože jsme zvolili kódování efekt (`PARAM=effect`), získáváme naprosto totožné výsledky, jako u procedury CATMOD.

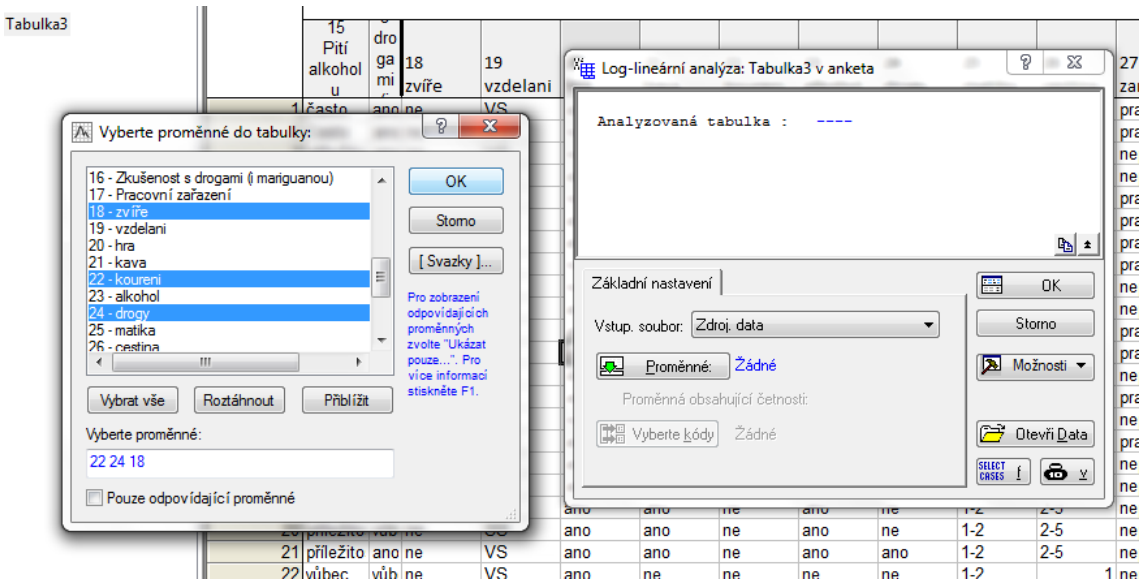


Obrázek 15: Použití loglineární analýzy ve STATISTICE

7.2 StatSoft STATISTICA 10

Stejného závěru při volbě nejvhodnějšího modelu pro znaky „drogy“, „kouření“ a „zvíře“ dosáhneme i při použití softwaru STATISTICA. Import dat z Excelu se nám nabídne hned po otevření softwaru STATISTICA. Po načtení excelovské tabulky s daty (je možné použít přímo odpovědi jednotlivých respondentů, není potřeba zadávat data jako četnosti) můžeme rovnou použít záložku *Statistika*, v níž najdeme *Pokročilé modely* a zvolíme volbu *Loglineární analýza*. Postup vidíme na obrázku 15. Následně se nám otevře okno *Log-lineární analýza Tabulka 3 v anketa.*

Nyní je potřeba zadat, v jakém formátu se data nachází. Chceme použít přímo kódované odpovědi respondentů (nikoli četnosti) a proto zvolíme možnost *Zdroj.data*. Kliknutím na tlačítko *Proměnné*, zvolíme znaky, které budou vstupovat do modelu. V našem případě „kouření“, „drogy“ a „zvíře.“ Uvedený postup najdeme na obrázku 16. Zvolit proměnné lze dvěma způsoby. Buď myší vybereme názvy sloupců, které chceme do analýzy vložit (za současného držení klávesy Ctrl) nebo přímo do řádku vypíšeme čísla sloupců, které budeme používat. Výběr potvrdíme tlačítkem *OK*, čímž se vrátíme do okna *Log-lineární analýza Tabulka 3 v anketa.* Pokud bychom chtěli použít jen některé kódy pro vstupní proměnné, provedli bychom to přes tlačítko *Vyberte kódy*. My tuto akci provádět nepotřebujeme, neboť v proměnných „kouření“, „drogy“ a „zvíře“ máme vždy jen dvě



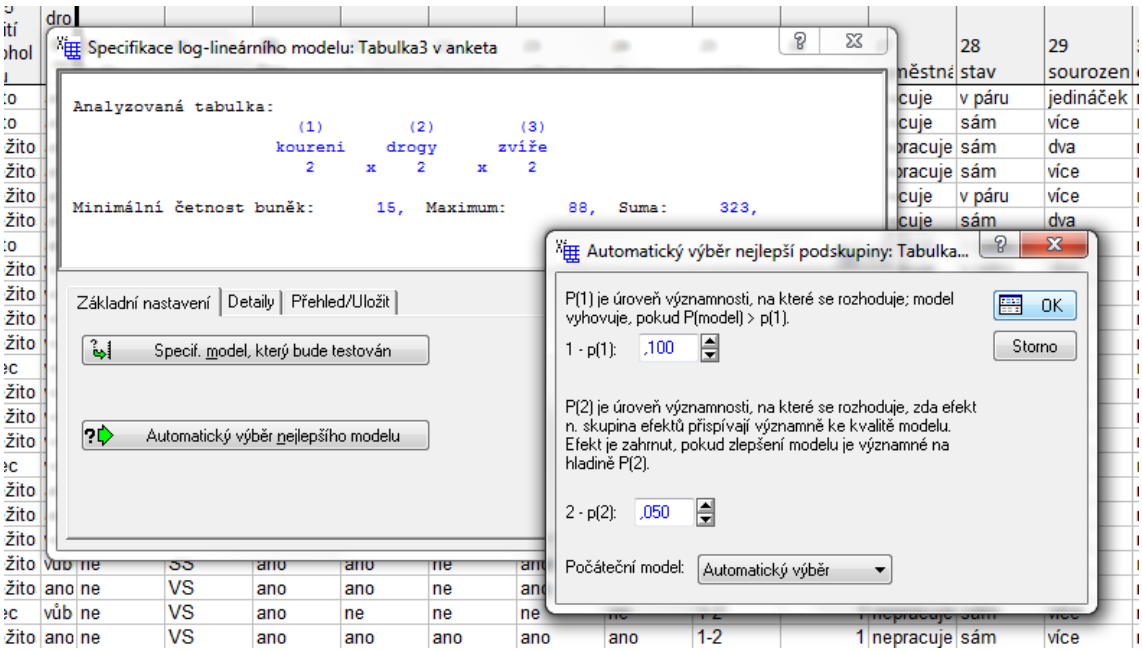
Obrázek 16: Výběr vstupních proměnných

kategorie ANO/NE. Možnost výběru kódů je vhodná v případě, že některé respondenty nechceme v analýze použít. Například, pokud bychom chtěli pracovat s proměnnou „vzdelani“ a nechtěli bychom v analýze respondenty se základním vzděláním, pak bychom zvolili pouze kódy (kategorie) SS a VS.

Nyní je již možné potvrdit výběr tlačítkem *OK*, čímž se dostaneme do okna *Specifikace loglineárního modelu.* Na obrázku 17 vidíme vypsané vstupní proměnné, označené čísly 1-3, včetně počtu příslušných kategorií, celkový počet respondentů (Suma) a minimální a maximální četnosti v tabulce.

V tomto okně je možné již přímo zvolit Automatický výběr nejlepšího modelu. V záložce *Detaily* je možné specifikovat model, testovat marginální a parciální asociace a ovlivňovat iterační algoritmus (lze měnit počet iterací a kritérium konvergence). V záložce *Přehled/Uložit* je možné nechat vypsat tabulku pozorovaných četností a uložit ji. V *Základním nastavení* lze konkrétně specifikovat model, který chceme testovat. Pokud tedy chceme testovat párovou závislost, můžeme zadat rovnou model „12 23 13“, což značí model „kouření*drogy + drogy*zvíře + kouření*zvíře“. My vybereme možnost *Automatického výběru modelu*, kde přijmeme nabídnuté hodnoty pro testování významnosti. Zde je možné i ručně zvolit počáteční model. Popsaný postup je vidět na obrázku 17.

Právě se dostáváme do okna „Automatický výběr nejlepší podskupiny,“ kde vidíme počáteční model a nejlepší výsledný model a postup najdeme na obrázku



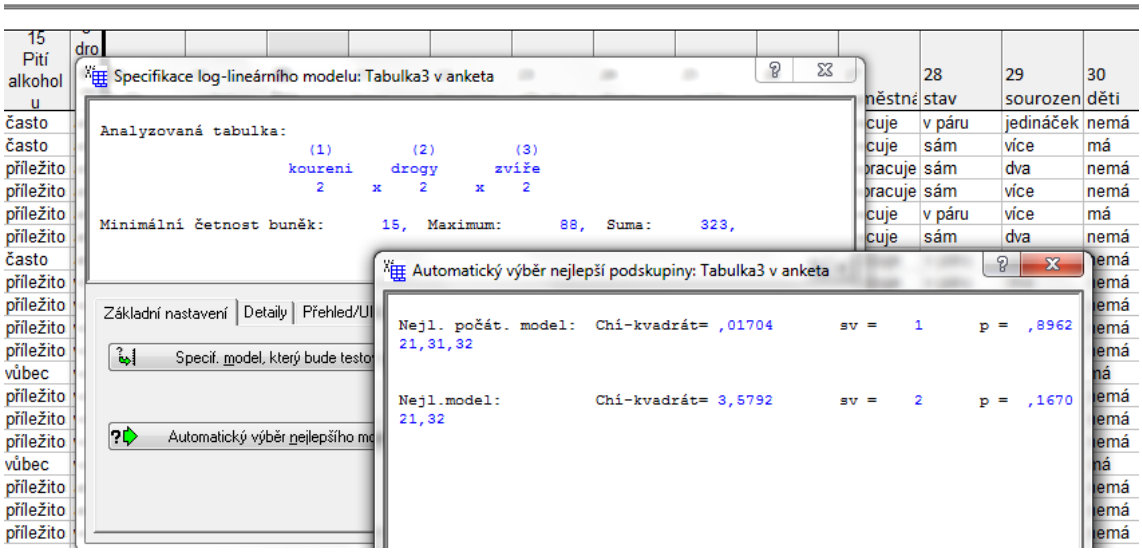
Obrázek 17: Specifikace modelu

18. Počáteční model byl zvolen jako model párové závislosti a nejlepším modelem vyšel model podmíněné nezávislosti znaků „kouření“ a „zvíře.“ Závěr je stejný jako při použití Softwaru SAS. Významnosti parametrů v modelu najdeme pod již zmíněnou možností testování párových a marginálních závislostí.

Software STATISTICA zvolil jako nejlepší přesně tytéž modely, které jsme pomocí softwaru SAS vybrali za nejlepší i my. Jediný rozdíl nastal v případě analyzování trojice „kouření, drogy, děti“, kdy STATISTICA zvolila za nejlepší model saturovaný. Jak jsme si již uváděli, je vždy snaha získat model jednodušší než saturovaný, proto jsme zvolili spíše model podmíněné nezávislosti znaků „kouření“ a „deti.“

7.3 Program R 2.13.2

Jako poslední ke zpracování dat „drogy, kouření a zvíře“ použijeme statistický program R, verzi 2.13.2. Program R nabízí k loglineární analýze několik možností. První z nich může být příkazem `loglm`, který najdeme v balíku MASS. Pokud nemáme balík MASS v R nainstalovaný, provedeme instalaci tak, že na horní liště zvolíme PACKAGES, vybereme INSTALL PACKAGES. Dále zvolíme tzv. CRAN mirror (zemi odkud chceme instalovat balík) a pak už jen vybe-



Obrázek 18: Výsledný model, který dle soft. STATISTICA nejlépe vystihuje data

reme příslušný balík. Příkaz `loglm` neposkytuje žádné rozsáhlé výsledky. Prakticky jediným výstupem je hodnota Likelihood Ratia (deviance) a Pearsonova Chí-kvadrátu, včetně stupňů volnosti a významnosti statistik. Devianci můžeme zjistit i dalším příkazem `deviance(model)`. Zdrojový kód lze zapsat následujícím způsobem.

```
# načtení kontingenční tabulky z textového dokumentu
# parametr header určuje, že první řádek obsahuje názvy
# proměnných
> DZK = read.table("data_zvíře_kouření_drogy.txt",
header = TRUE)

# vypsaná data
> DZK

  drogy zvíře kouření počet
1   ano    ne     ano    27
2   ano    ne      ne    26
3   ano   ano     ano    45
4   ano   ano      ne    28
5    ne    ne     ano    15
6    ne    ne      ne    45
7    ne   ano     ano    49
```

```
8    ne   ano      ne    88

# otevření balíku MASS
> library(MASS)

# sestavíme model podmíněné nezávislosti
# parametr data určuje, odkud se berou názvy proměnných
> model <- loglm( počet ~  zvíře * drogy + drogy * kouření,
data=DZK )

# necháme si vypsat výsledek
> model

Call:
loglm(formula = počet ~ zvíře * drogy + drogy * kouření,
data = data)

Statistics:
                      X^2 df  P(> X^2)
Likelihood Ratio 3.701152  2 0.1571466
Pearson          3.640849  2 0.1619570
```

Vhodnějším přístupem k loglineárním modelům je příkaz `glm`, který je přímo v základu R. Nabízí mnohem rozsáhlejší spektrum možností v loglineární analýze. Načtení dat se prování totožně jako v modelování pomocí `loglm`, poté se sestaví model a příkazem `summary(model.podm.nez.)` získáme informace o odhadech parametrů v modelu, chybách, testových statistikách a významnostech. V tomto výstupu získáme dokonce i hodnotu AIC, což je tentokráte hodnota AIC přímo pro daný model, nikoli rozdíl AIC testovaného modelu a saturovaného modelu. Další příkaz, který můžeme následně použít je `anova(model.podm.nez.)`. Tento příkaz uvede významnosti parametrů v modelu, deviance, stupně volnosti a rozdíly deviancí mezi parametry i rozdíly stupnů volnosti mezi parametry. Zdrojový kód včetně výstupů může vypadat následovně.

```
# model podmíněné nezávislosti
# parametr family= určuje, z jakého data pocházejí rozdělení
# parametr data= určuje, odkud brát názvy parametrů v modelu
> model.podm.nez. <- glm(počet ~  zvíře * drogy +
```

```
kouření * drogy, family=poisson, data=DZK)

# vypíše použitý model, odhady parametrů
# a dokonce hodnotu AIC
> summary(model.podm.nez.)

Call:
glm(formula = počet ~ zvíře * drogy + kouření * drogy,
family = poisson, data = DZK)

Deviance Residuals:
      1        2        3        4        5        6
-0.6084   0.6737   0.5023  -0.5982  -1.0609   0.6934

 7        8
0.6625  -0.4710

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.73084    0.14018  26.614  < 2e-16 ***
zvířene          -0.32017    0.18046  -1.774   0.0760 .
drogyne           0.06482    0.19365   0.335   0.7378
kouřeníne        -0.28768    0.18002  -1.598   0.1100
zvířene:drogyne  -0.50547    0.23777  -2.126   0.0335 *
drogyne:kouřeníne  1.01915    0.23569   4.324 1.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 80.8059  on 7  degrees of freedom
Residual deviance:  3.7012  on 2  degrees of freedom
AIC: 59.023

Number of Fisher Scoring iterations: 4

# vypíše přehled parametrů, stupně volnosti, devianci
# a významnost parametrů
> anova(model.podm.nez., test="Chisq")

Analysis of Deviance Table
```

```
Model: poisson, link: log

Response: počet

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                             7     80.806
zvíře          1  29.5845         6     51.221 5.353e-08 ***
drogy          1  15.7350         5     35.486 7.286e-05 ***
kouření        1   8.0864         4     27.400   0.00446 **
zvíře:drogy    1   4.5174         3     22.883   0.03355 *
drogy:kouření  1  19.1814         2      3.701 1.189e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Z výše uvedeného je vidět, že hodnota deviance rozdílu testovaného modelu a saturovaného modelu je 3,7 a má 2 stupně volnosti. Stejný výsledek jsme obdrželi i z procedury CATMOD a GENMOD.

7.4 Diskuze

Z výsledků, uvedených v předchozí kapitole nelze jednoznačně říci, který použitý statistický software je ke zpracování dat pomocí loglineárních modelů nejvhodnější. Nejspíše záleží na uživateli, jaký si představuje výsledek nebo které prostředí je pro něj vhodnější. Pokud preferujeme „klikací“ prostředí, pak je nanejvýš vhodný software STATISTICA, v němž jsme ale omezeni předem nadefinovanými možnostmi softwaru. Chceme-li být variabilnější a vybírat si, které výstupy jsou pro nás vhodné a co naopak nepotřebujeme, je lepší použít software SAS nebo R, v nichž lze volit typ kódování nebo výpočetní algoritmus. Bezespornou výhodou softwaru R oproti všem ostatním je, že je dostupný zdarma. Naopak velkou nevýhodou softwaru R je velmi slabá nápověda a časově dosti náročné vyhledávání příkazů a jejich pochopení. Dle zkušeností z této práce je nejvhodnější kombinování všech zmíněných softwarů. Software SAS má také výhodu, že velice elegantně pracuje s chybějícími hodnotami, nulami v kontingenční tabulce a pře-

devším jeho výstupy jsou vhodné k publikování. Výstupy z programu R je nutné graficky upravit, aby byly „líbivější“.

Závěr

Po přečtení této diplomové práce snad čtenář získá ucelený přehled o loglineárních modelech a o šancích v kontingenčních tabulkách, což se vše dočte v části teoretické. Z výzkumné (praktické) části by mělo být zřejmé, jak konkrétní data zpracovat pomocí některého ze tří uvedených softwarů, případně vhodnou kombinací dosáhnout kýžených výstupů.

Kontingenční tabulky jsou elegantním nástrojem k analýze kategoriálních dat, což lze vidět na výsledcích z dotazníku uvedeného ve výzkumné části. Trojrozměrné kontingenční tabulky lze využít nejen pro sociologické průzkumy, ale i pro analýzu medicínských dat, kdy zjišťujeme například vztahy mezi typem onemocnění pacienta, metodou léčby a výsledkem léčby. Další využití najdeme například v dopravě, kdy na určitém úseku komunikace víme, kolik nastalo nebo nenastalo přírodních katastrof (povodeň, polom a sněhová kalamita) a chceme zjišťovat, zda spolu tyto katastrofy nějakým způsobem souvisí. V moderním statistickém softwaru jako SAS, STATISTICA, SPSS Statistics nebo program R jsou implementovány příkazy a procedury na práci s trojrozměrnými kontingenčními tabulkami, konkrétně loglineárními modely a umožňují i zpracování velkých objemů dat. Tím se možnosti takové analýzy ještě prohlubují. A není ani nutné omezovat se vždy jen na tři dimenze.

Tato práce obsahuje jen zlomek toho, jak rozsáhle jde s kontingenčními tabulkami pracovat. V této práci jsme se zaměřili pouze na kategoriální data, ale loglineární modely lze upravit i pro analyzování spojitých dat, lze využít znalostí markovových bází, využít loglineární modely pro analýzu přežívání nebo se zabývat grafickým znázorněním vztahů v kontingenčních tabulkách, což jsou už ale problematiky vysoce přesahující rámec této práce. Loglineární modelování je poměrně nové statistické odvětví, které nachází uplatnění především na zahraničních univerzitách a ústavech. Do povědomí českých vysokých škol se dostává zatím pouze sporadicky. Nejen z toho důvodu byla práce zpracována v českém jazyce, neboť existuje velmi málo česky psaných materiálů obsahujících ucelené informace o trojrozměrných tabulkách.

Příloha

Příloha 1: strany 99-102

Dotazník, který obdrželo přibližně 400 lidí byl vypracován pomocí aplikace Google chrome. Dotazník byl posílán online pod adresou https://docs.google.com/spreadsheet/viewform?formkey=dDhsYTF1Nks1ZkhmSXVrVEtIVjVQbmc6MQ.

Příloha 2: strany 103 a 104

V tabulkách na straně 103 a 104 se nachází vybrané loglineární modely. U každého modelu je uveden odhad parametrů.

Příloha 3: strany 105-107

Zdrojový kód pro metodu maximální věrohodnosti.

Příloha 4: strany 108 a 109

Zdrojový kód pro metodu minimalizace.

Příloha 5: loglinearni_modely.xlsx na přiloženém CD

V tabulkách se nachází 15 variant loglineárních modelů (vždy po devíti modelech od každé trojice znaků). V každém modelu se vyskytují znaky „kouření“ a „drogy“ a vždy jeden další znak. Ke každé trojici znaků jsou vždy všechny hierarchické loglineární modely (jedna stránka pro jeden typ modelu.
Ve sloupci „source“ se nachází parametry modelu, ve sloupci „DF“ stupně volnosti, ve sloupci „Chi-square“ je testová Waldova statistika, ve sloupci „Pr>ChiSq“ je významnost Waldovy statistiky. U modelů, které nejsou saturované, je další sloupec s Akaikeho kritériemi. Pod saturovaným modelem je modře vyznačený nejvhodnější model. Červeně jsou vyznačeny NEVÝZNAMNÉ parametry.
Pod modely, které vstupovaly do porovnávání, jsou uvedeny hodnoty Chí-kvadrát rozdělení se stupni volnosti rozdílu DF v porovnávaných modelech hned ve vedlejší buňce hodnota testové statistiky (rozdíl Likelihood Ratia porovnávaných testů). Zeleně jsou vyznačeny hodnoty nevýznamné (model se odebráním para-

metrů nezlepšil) a modře jsou zvýrazněné významné hodnoty (v modelu nastalo zlepšení odebráním parametrů).

Příloha 6: odhady_v_modelech.xlsx na CD

V tabulkách najdeme odhady parametrů vybraných loglineárních modelů.

Příloha 7: data.xlsx na CD

Obsahuje surová data z dotazníku.

Anketa

Diky tehle ankete mi muzete pomoci pri vytvareni me diplomove prace. Budu Vam za to moc vdecna. Udaje budou slouzit ke statistickemu zpracovani, nikde nebudou uvedena jmena ani emailove adresy. Odpovidejte prosim popravde, aby me zavery pak byly spravne. Moc dekuji a omlouvam se za chybejici diaktritiku. Zde nebylo mozne psat hacky a carky.

Pohlaví

- muž
- žena

Věk

- 0 - 14 let
- 15 - 18 let
- 19 - 25 let
- 26 - 30 let
- 31 - 40 let
- 41 - 50 let
- 51 - 60 let
- 61 let a více

Bydliště nyní

- vesnice
- malé město do 30 tis. obyvatel
- střední město do 100 tis. obyvatel
- velké město nad 100 tis. obyvatel

Bydliště do 18 let. Platí pro ty, kteří nyní bydlí v jiném domě / bytě než do 18 let. Tahle otázka má zjistit, kde jste bydleli v dětství / mládí. Je možné více variant, pokud se někdo častěji stěhoval.

- malé město do 30 tis. obyvatel stejné jako v otázce Bydliště nyní
- malé město do 30 tis. obyvatel jiné než v otázce Bydliště nyní
- střední město do 100 tis. obyvatel stejné jako v otázce Bydliště nyní
- střední město do 100 tis. obyvatel jiné než v otázce Bydliště nyní
- velké město nad 100 tis. obyvatel stejné jako v otázce Bydliště nyní
- velké město nad 100 tis. obyvatel jiné než v otázce Bydliště nyní
- vesnice stejná jako v otázce Bydliště nyní
- vesnice jiná než v otázce Bydliště nyní
- stále bydlím ve stejném bytě / domě
- ještě mi není 18

Vzdělání, nejvyšší dosažené vzdělání

- ○ základní škola
- ○ učiliště, odborné učiliště, střední bez maturity
- ○ střední škola s maturitou
- ○ bakalářský titul nebo VOŠ
- ○ titul Mgr. nebo Ing.
- ○ Doktorát a vyšší

Hra na hudební nástroj

- ○ dříve ano, nyní vůbec
- ○ dříve ano, nyní někdy
- ○ hraji stále
- ○ nehrál / a jsem nikdy

Domácí zvíře. Je možné více odpovědí.

- ☐ pes v bytě / domě
- ☐ pes na zahradě
- ☐ kočka v bytě / domě
- ☐ kočka na zahradě
- ☐ rybičky
- ☐ hlodavec (králík, křeček, morče, činčila...)
- ☐ žádné zvíře nemáme

Nejčastější známka na vysvědčení z matematiky. Pokud se to střídalo, udělejte jakýsi průměr.

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

Nejčastější známka na vysvědčení z českého jazyka. Pokud se to střídalo, udělejte jakýsi průměr.

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

Rodinný stav

- svobodný / á
- druh / družka
- ženatý / vdaná
- rozvedený / á
- vdovec / vdova
- ženatý / vdaná vícekrát

Počet sourozenců (jedináčci nevyplňují)

- 1
- 2
- 3
- více než 3

Počet dětí

- 1
- 2
- 3
- více než 3
- žádné
- sám / sama jsem ještě dítě

Pití kávy

- několikrát denně
- jednou denně
- několikrát do týdne
- jednou za týden
- méně často
- vůbec
- dříve ano, nyní ne

Kouření cigaret

- pravidelně
- příležitostně
- vůbec
- dříve ano, nyní ne

Pití alkoholu

- často
- příležitostně
- vůbec
- dříve ano, nyní ne

Zkušenost s drogami (i marihuanou). To je kvůli srovnání s výzkumem z USA

- ano, často
- ano, jen párkrát
- vůbec

Pracovní zařazení

- student
- běžný zaměstnanec
- vedoucí pozice
- podnikatel / zaměstnavatel
- důchodce

Odeslat

ln m	=	**drogy + deti + koureni + drogy∗deti + drogy∗koureni**
ln m	=	3,4287 + (-0,3981) + (0,5303) + (-0,1028) + 0,4336 + 0,2629
		mean + D ano + De nema + K ano + DDe ano nema + DK ano ano
ln m	=	**drogy + pohlavi + koureni + drogy∗pohlavi + drogy∗koureni**
ln m	=	3,6030 + (-0,1579) + 0,1573 + (-0,0970) + (-0,2125) + 0,2487
		mean + D ano + P ano + K ano + DP ano zena + DK ano ano
ln m	=	**drogy + kava + koureni + koureni∗kava + drogy∗koureni**
ln m	=	3,2511 + (-0,2095) + 0,894 + (-0,409) + 0,445 + 0,2524
		mean + D ano + Ka ano + K ano + KKa ano ano + DK ano ano
ln m	=	**drogy + bydliste + koureni + drogy∗bydliste + drogy∗koureni**
ln m	=	2,8749 + (-0,2667) + 0,4807 + (-0,0982) + 0,266 + 0,2558
		mean + D ano + B velke_mesto + K ano + DB ano velke_mesto + DK ano ano
ln m	=	**drogy + alkohol + koureni + drogy∗alkohol + drogy∗koureni**
ln m	=	0,35# + (-2,7203)# + 3,9673 + (-0,0935) + 2,5374# + 0,2604
		mean + D ano + A ano + K ano + DA ano ano + DK ano ano
ln m	=	**drogy + koureni + matika + drogy∗koureni + matika∗koureni**
ln m	=	3,5884 + (-0,2095) + (-0,0423) + 0,322 + 0,2524 + (-0,1928)
		mean + D ano + K ano + M 1_2 + DK ano ano + MK 1_2 ano
ln m	=	**cestina + koureni + drogy + cestina∗koureni + drogy∗koureni**
ln m	=	3,5873 + 0,3488 + (-0,1541) + (-0,2061) + 0,1679 + 0,2558
		mean + C 2_5 + K ano + D ano + CK 2_5 ano + DK ano ano
ln m	=	**drogy + zvire + koureni + drogy∗zvire + drogy∗koureni**
ln m	=	3,5877 + (-0,1608) + (-0,2865) + (-0,1109) + 0,1264 + 0,2548
		mean + D ano + Z ne + K ano + DZ ano ne + DK ano ano
ln m	=	**drogy + stav + koureni + drogy∗stav + drogy∗koureni**
ln m	=	3,5556 + (-0,3056) + (-0,3768) + (-0,0982) + (-0,2829) + 0,2558
		mean + D ano + S v_paru + K ano + DS ano v_paru + DK ano ano
ln m	=	**drogy + koureni + zamestnani + drogy∗koureni + drogy*zamestnani**
ln m	=	3,6429 + (-0,19) + (-0,0982) + (-0,144) + 0,2558 + 0,1127
		mean + D ano + K ano + Z nepracuje + DK ano ano + DZ ano nepracuje

ln m	=	**drogy + vzdelani + koureni + drogy∗vzdelani + vzdelani∗koureni + drogy∗koureni**
ln m	=	2,8462 + (-0,2607) + 0,7348 + (-0,0172) + 0,249 + (-0,3118) + 0,3098
		mean + D ano + V VS + K ano + DV ano VS + VK VS ano + DK ano ano
ln m	=	**drogy + vek + koureni + drogy∗vek + vek∗koureni + drogy∗koureni**
ln m	=	2,9411 + (-0,4824) + 0,6218 + (-0,0354) + 0,4816 + (-0,1588) + 0,3165
		mean + D ano + V do25 + K ano + DV ano do25 + VK do25 ano + DK ano ano
ln m	=	**drogy + koureni + hra + drogy∗koureni**
ln m	=	3,6433 + (-0,1967) + (-0,104) + 0,1164 + (0,2615)
		mean + D ano + K ano + H ano + DK ano ano
ln m	=	**drogy + koureni + sourozenci + drogy∗koureni**
ln m	=	3,0330 + (-0,2061) + (-0,0982) + (-0,986) + 0,2558
		mean + D ano + K ano + S jedinacek + DK ano ano
ln m	=	**drogy + koureni + bydliste do 18+ drogy∗koureni**
ln m	=	2,5801 + (-0,1831) + (-0,0936) + (-0,0509) + 0,2447
		mean + D ano + K ano + B stredni_mesto + DK ano ano

```
% Zdrojový kód pro metodu maximální věrohodnosti
% zadejte kontingenční tabulku
tabulka = [38 27 6;33 19 1] % první hladina tabulky
tabulka(:,:,2) = [16 32 7;72 43 10] % druhá hladina tabulky

% zaveďme proměnné jako nulové matice
radky = zeros(J,K);
sloupce = zeros(I,K);
vrstvy = zeros(I,J);

% zjistíme rozměr tabulky a uložíme jako proměnné I,J,K
[I,J,K] = size(tabulka)

% součet přěs řádky n_(+jk)
for k = 1: K
    for j = 1: J
    radky(j,k) = sum(tabulka(:,j,k));
    end
end

% součet přes sloupce n_(i+k)
for i = 1: I
    for k = 1: K
    sloupce(i,k) = sum(tabulka(i,:,k));
    end
end

% součet přes hladiny n_(ij+)
for i = 1: I
    for j = 1: J
    vrstvy(i,j) = sum(tabulka(i,j,:));
    end
end

% počáteční aproximace tj. každé m_(ijk) = 1
M = ones(I,J,K)

M1 = M;
iterace = 1;
```

```
% algoritmus
    for i = 1: I
        for j = 1: J
            for k = 1: K

 M(i,j,k) = M1(i,j,k)*vrstvy(i,j)/sum(M1(i,j,:));

            end
        end
    end

 M1 = M;

    for i = 1: I
        for j = 1: J
            for k = 1: K

  M(i,j,k) = M1(i,j,k)*sloupce(i,k)/sum(M1(i,:,k));

            end
        end
    end

M1 = M;

    for i = 1: I
        for j = 1: J
            for k = 1: K

  M(i,j,k) = M1(i,j,k)*radky(j,k)/sum(M1(:,j,k));

             end
        end
    end

% kritérium pro zastavení algoritmu
epsilon = 1;
iterace = 1;
while (max(max(max(abs(mr - m)))) > epsilon)
&& (iterace < 10)
```

```
% opět algoritmus
M1 = M;
    for i = 1: I
        for j = 1: J
            for k = 1: K

  M(i,j,k) = M1(i,j,k)*vrstvy(i,j)/sum(M1(i,j,:));

            end
        end
    end

 M1 = M;

    for i = 1: I
        for j = 1: J
            for k = 1: K

  M(i,j,k) = M1(i,j,k)*sloupce(i,k)/sum(M1(i,:,k));

            end
        end
    end

M1 = M;

    for i = 1: I
        for j = 1: J
            for k = 1: K

  M(i,j,k) = M1(i,j,k)*radky(j,k)/sum(M1(:,j,k));

             end
        end
    end

% pokud nebyl algoritmus zastaven, přidá další iteraci
iterace = iterace + 1

end
```

```
% Zdrojový kód pro metodu minimalizace
% zadejte trojrozměrnou tabulku
tabulka(:,:,1) = [ ]; %první hladina tabulky
tabulka(:,:,2) = [ ]; %druhá hladina tabulky

% zadejte očekávané marginální četnosti
mi = [ ]      % vektor m_{i++}
mj = [ ];     % vektor m_{+j+}
mk = [ ];     % vektor m_{++k}

% zjistíme velikost tabulky a uložíme do proměnných I,J,K
[I,J,K] = size(tabulka);

% zadáme počáteční matice jako nulové
M = zeros(I,J,K);
M1 = M;

% algoritmus
for(krok = 1:10) % počet iterací
    for(i = 1:I)
        for(j = 1:J)
            for(k = 1:K)

  M(i,j,k) = tabulka(i,j,k).*mi(i)./sum(sum(tabulka(i,:,:)));

            end
        end
    end
    for(i = 1:I)
        for(j = 1:J)
            for(k = 1:K)

  M(i,j,k) = M(i,j,k).*mj(j)./sum(sum(M(:,j,:)));

            end
        end
    end
    for(i = 1:I)
        for(j = 1:J)
```

```
            for(k = 1:K)

  M(i,j,k) = M(i,j,k).*mk(k)./sum(sum(M(:,:,k)));

            end
        end
    end

% napočítá nové marginální četnosti matice M
    mi = sum(sum(M,2),3);
    mj = sum(sum(M,1),3);
    mk = sum(sum(M,1),2);

% zaokrouhlená (stačí nám celé četnosti) norma matice,
což slouží k posouzení konvergence metody
round(sum(sum(sum(M1-M))))
% uloží matici M do matice M1 a algoritmus pokračuje další iterací
    M1 = M;
end
```

Literatura

[1] Agresti, A.: *Categorical data analysis*, John Wiley and Sons, New Jersey, 2002

[2] Allison P. D.: *Logistic regression using the SAS system: theory and application*, SAS Institute Inc. and John Wiley and Sons, Inc., NY, 1999

[3] Anderson, C. J.: [online] *Log-linear Models for Contingency Tables* http://www.scribd.com/doc/52071342/6loglin1-ha-online, datum stažení: 2. říjen 2011

[4] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2007

[5] Brown, M. B., Fuchs, C.: článek *On maximum likelihood estimation in sparse contingency tables*, Computational Statistics & Data Analysis 1, str. 3-15, březen 1983

[6] Cressie, N., Read, T., R., C.: článek *Multinomial Goodness-of-fit tests*, Journal of the Royal Statistical Socienty, Volume 46, No. 3, str. 440-464, 1984

[7] Deming, W. E., Stephan, F. F.: článek *On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known*, The Annals of Mathematical Statistics, Vol. 11, č. 4 (prosinec, 1940), str. 427-444

[8] Estrada, V.,Lagier, R., Univerzita Kalifornie (East Bay): [online] *Loglinear Models for Independence and Interaction in Three-way Tables*, www.sci.csuhayward.edu/∼jkwon/classes/stat_6841/PROJECTS/ VE_RL_project.ppt, datum stažení: 6. březen 2011

[9] Fienberg, S. E.: *The analysis of cross-clasified categorical data*, The MIT-Press, Masachusetts, 1985

[10] Fienberg, S. E., Fulp, W. J., Slavkovic, A. B. and Wrobel, T. : *„Secure“ Log-Linear and Logistic Regression Analysis of Distributed Databases*, Privacy in Statistical Databases – PSD 2006, Lecture Notes in Computer Science No.4302, str. 277-290. Berlin, Springer-Verlag.

[11] Hofmann, H.: [online] *Loglinear Models* http://www.hofroe.net/stat557/13-loglinear.pdf, datum stažení: 2. říjen 2011

[12] Cheng P. E., Liou M., Aston J. A. D.: článek *Likelihood Ratio Tests With Three-Way Tables*, Journal of the American Statistical Association, 6. červen 2010

http://www.stat.sinica.edu.tw/pcheng/wp-content/plugins/downloads-manager/upload/3-wayLRtest_JASA_2010.pdf,
datum stažení: 14. listopad 2011

[13] Christensen, R.: *Log-linear models and logistic regression*, 2nd ed., Springer Verlag, New York, 1997

[14] Irwin, M. E.: [online] *Three-way Contingency Tables*
dostupné na http://www.markirwin.net/stat149/Lecture/Lecture19.pdf, datum stažení: 22. leden 2011

[15] Jeansonne A., Univerzita San Francisco: [online] *Loglinear Models*, http://userwww.sfsu.edu/∼efc/classes/biol710/loglinear/Log%20Linear%20
Models.html, datum stažení: 20. březen 2012

[16] Katrňák, T.: [online] *Log-lineární modely* dostupné na
http://fss.muni.cz/∼katrnak/akd/log_lin_modely.pdf, datum stažení: 22. leden 2011

[17] Katrňák, T.: [online] *Analýza kategoriálních dat v sociologii*, http://www.lsvv.eu/workshop/katrnak/katrnak_prezentace_1.pdf,
datum stažení: 19.bžezen 2012

[18] Kelderman, H.: článek *Computing Maximum Likelihood Estimates Of Loglinear Models From Marginal Sums With Sspecial Attention To Loglinear Item Response Theory*, Psychometrika, Volume 57, No. 3, str. 437-450, září 1992

[19] King, W. B., Univerzita Coastal Kalifornie: [online] *Log Linear Analysis*, http://ww2.coastal.edu/kingw/statistics/R-tutorials/loglin.html, datum stažení: 12. únor 2012

[20] Kroonenberg, P. M.: [online] *Singular value decompositions of interactions in three-way contingency tables*, dostupné na https://openaccess.leidenuniv.nl/bitstream/handle/1887/11625/7_702_059.pdf?sequence=1, datum stažení: 19. únor 2012

[21] Kuha, J., Firth, D.: článek *On the index of dissimilarity for lack of fit in loglinear and log-multiplicative models*, CRISM Paper,No. 09-26, dostupné na www.warwick.ac.uk/go/crism, datum stažení 29. leden 2012

[22] Kumal A.: SAS-Procedures, [online] http://www.scribd.com/doc/6396633/SAS-Procedures, staženo 26. února 2012

[23] Mulekar, S. M., Knutson, J. C., Champanerkar, J. A.: článek *How useful are approximations to mean and variance of the index of dissimilarity?*, Computational Statistics & Data Analysis 52, str. 2098 – 2109, 2008

[24] Pecáková, I.: [online] *Testy nezávislosti v řídkých kontingenčních tabulkách* http://panda.hyperlink.cz/cestapdf/pdf07c1/pecakova.pdf datum stažení: 22. říjen 2011

[25] Pecáková, I.: *Analýza a modelování souvislostí kategoriálních proměnných*, habilitační práce, Praha, 2004

[26] Prášková, Z.: *Kontingenční tabulky*, Univerzita Karlova, Praha, 1985

[27] Shuhua, H., Univerzita Severní Karolína: [online] http://www4.ncsu.edu/∼shu3/Presentation/AIC.pdf, datum stažení: 22. únor 2012

[28] Simpson, L., Tranmer, M.: článek *Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software*, The Professional Geographer Volume 57, Issue 2, str. 222–234, květen 2005

[29] Smith, N. A., Univerzita Maryland: [online] *Log-Linear models*, http://www.cs.cmu.edu/∼nasmith/papers/smith.tut04.pdf, datum stažení: 17. října 2011

[30] Stokes, M. E., Davis, Ch. S., Koch, G. G.: *Categorical data analysis using the SAS system*, SAS Institute Inc. and John Wiley and Sons, Inc., NY, 2003

[31] Theus, M., Lauer, S. R. W.: [online] *Visualizing Loglinear Models* dostupné na http://home.vrweb.de/∼martin.theus/theus.pdf, datum stažení: 8. únor 2011

[32] Thomson, L.: [online] *R (and S-PLUS) Manual to Accompany Agresti's Categorical Data Analysis (2002) 2nd edition*, https://home.comcast.net/∼lthompson221/Splusdiscrete2.pdf, datum stažení: 2. říjen 2011

[33] Vokey, J. R.: článek *Collapsing multiway contingency tables: Simpson's paradox and homogenization*, Behavior Research Methods, Volume 29, Number 2, str. 210-215, [online] http://www.springerlink.com/content/r9887u55842x1j21/fulltext.pdf, datum stažení 2. prosinec 2011

[34] Univerzita Idaho: [online] studijní text, http://www.uiweb.uidaho.edu/ag/statprog/sas/workshops/catmod/handout5_cat.pdf, datum stažení: 11. únor 2012

[35] Univerzita Illinois: [online] *Examples of Loglinear Models in R*, https://netfiles.uiuc.edu/jimarden/www/Classes/gems/loglin_ex.pdf, datum stažení: 20. únor 2012

[36] Univerzita Karlova: [online] *Loglineární analýza*, http://samba.fsv.cuni.cz/∼soukup/ANALYZAKVANTITDAT/lekce9_10.ppt, datum stažení: 12. březen 2012

[37] Univerzita Minnesota: [online] studijní text, http://www1.umn.edu/statsoft/doc/statnotes/stat01.txt, datum stažení: 6. březen 2012

[38] Univerzita Oregon: [online] studijní text, http://rfd.uoregon.edu/files/rfd/StatisticalResources/gnmd09_loglin.txt, datum stažení: 12. březen 2012

[39] Univezita Pennsylvánie: [online] studijní text, http://sites.stat.psu.edu/∼jls/stat544/lectures/lec7.pdf, datum stažení: 14. listopad 2011

[40] Univerzita Pennsylvánie: [online] studijní text, https://onlinecourses.science.psu.edu/ stat504/node/122, datum stažení: 14. listopad 2011

[41] Univerzita Princeton: [online]*Log-Linear Models for Contingency Tables*, http://data.princeton.edu/wws509/notes/c5.pdf, datum stažení: 8. únor 2011

[42] Univerzita Texas: [online] *Loglinear Models for Contingency Tables* http://faculty.business.utsa.edu/rtripath/7853/Chapter8/Lecture8_1u.pdf, datum stažení: 8. únor 2011

[43] Univerzita Washinghton: [online] studijní text, http://www.stat.washington.edu/quinn/classes/536/S/loglinexample.html, datum stažení: 11. únor 2012

[44] EuroMise: [online] studijní text, http://ucebnice.euromise.cz, datum stažení: 2. říjen 2011

[45] Support SAS: [online] http://support.sas.com/kb/24/447.html#ex3.e.1, datum stažení: 23. březen 2012

[46] Semináře statistiky, nápověda a příklady v R: [online] https://stat.ethz.ch/, datum stažení: 3. březen 2012

[47] wikipedia: [online] http://cs.wikipedia.org, datum: 23. duben 2011