Analýza ordinálních dat: úlohy, přístupy a metody

Jan Řehák, Ústav pro filozofii a sociologii ČSAV, Praha

Ordinální kategorizovanou proměnnou, neboli uspořádanou klasifikací nazveme úplnou soustavu vylučujících se tříd, které jsou uspořádané, $A = \{A_1, A_2, \ldots, A_K\}$; uspořádání je dáno indexy $k = 1, 2, \ldots, K$.

Realizací ordinálních proměnných dostáváme ordinální data, která vstupují do kontingenčních tabulek nebo tvoří jednorozměrná rozložení četností v konečném počtu uspořádaných tříd.

Ordinální proměnné (jako modely reálných vlastností) vznikají dvojím způsobem:

a) odrážejí svými třídami různé, v realitě zřetelně oddělené kvality s definovanými hranicemi; uspořádání se předpokládá jen mezi třídami, nikoliv uvnitř (např. hodnosti v armádě, funkce v řízení, hierarchie organizační, administrativní, období v životě daná zlomem, ukončená kvalifikace);

b) třídy vznikají jako výraz rozkladu nějaké spojité škály, podél jíž jsou všechny jednotky uspořádány, ale která je rozdělena na části, jejichž řazení je ovšem určující latentní škálou dáno jednoznačně; uvnitř tříd se proto od reálného (podrobného) uspořádání odhlíží (např. stupeň spokojenosti, schopnosti, kategorie věku).

V prvním případě jsou třídy určovány přirozeně, ve druhém podle situace, metodologie, cílů i metodiky zpracování. K druhému případu snad lze také přiřadit situaci, ve které kategorie vznikají jako souhrn přirozených spojení podél spojité škály.

Určení modelu bývá v praxi velmi obtížné a vede vždy k určitým nepřesnostem, či zanedbání některých reálných vztahů; proto abstrakce, kterou volbou typu proměnné provádíme, závisí na subjektu a je jen málokdy přesnou kopií skutečnosti. Důsledky, které plynou z chybného určení modelu pro statistické výsledky, nejsou prozatím dostatečně prozkoumány a proto se přístupy v praxi prolínají tak, jak jsou jednotlivé metody k dispozici.

Úlohy formulované pro ordinální data jsou obdobné k úlohám řešeným v analýze nominálních dat, vystupuje v nich ovšem skutečnost uspořádání tříd:

1. Popis rozložení četností $(n_1, n_2, \ldots, n_K)$ na hodnotách proměnné A pomocí vhodných souhrnných parametrů.

2. Komparace R rozložení ($R \geq 2$) téže proměnné u R nezávislých souborů dat, resp. komparace jejich charakteristik.

3. Korelační a asociační analýza zahrnující ordinální resp. ordinální a nominální proměnné.

4. Analýza závislostních modelů v M-rozměrných kontingenčních tabulkách ($M \geq 2$).

5. Analýza párově spřažených ordinálních dat v tabulkách K x K.

Tyto úlohy obsahují jednak testování platnosti hypotetických modelů a jednak explorační postupy (odhad parametrů, řazení, geometrická reprezentace, seskupování ap.). Vyskytují se i další úlohy, které jsou paralelní k lineárním modelům. Speciálním problémem je najít uspořádání kategorií prostých klasifikací (nominálních proměnných) podle empiricky vzniklých asociačních struktur. Tím se zabývá duální škálování, kanonická analýza, korespondenční analýza, LINDA, RC model (viz např. S.Nishisato [1980], M.G.Kendall, A. Stuart [1973], Lebart, Morineau Warwick [1984], J.Řehák, I.Loučková [1984], L.A.Goodman [1981]. Problém není v tomto příspěvku rozváděn.

V analýze ordinálních dat lze rozlišit tři od sebe značně odlišné přístupy:

a) <u>C-modely</u>: předpokládá se spojitá škála, která určuje pořadí kategorií, jež vznikají jejím rozkladem. Metody jsou většinou odvozeny z neparametrických technik založených na pořadí s využitím velkého počtu spojení. Používají se

- ridity (Wilcoxonovy skóry) resp. průměrná pořadí
- Kendallovy skóry
- vzdálenost dvojic.

b) <u>D-modely</u>: předpokládají pevné diskrétní kategorie, metody jsou založeny na distribučních funkcích resp. jejich transformacích. Patří sem

- kvadratické formy (speciální případ obecného D-modelu)
- Plackettův model stejnoměrné globální asociace
- diskrétní analogy k distančním (Kolmogorov-Smirnov, metrika absolutních hodnot) statistikám
- modely kumulačního efektu včetně logitových pro závisle proměnné.

c) <u>E-modely</u>: kategoriím se přiřazují pevné číselné hodnoty a ty určují jejich uspořádání. Jde de facto o kvantifikaci a tudíž o kardinální kategorizované proměnné. Využívají se

- log-lineární modely
- logitové modely (v kvantifikaci nezáv.prom.)
- pseudokvantifikace.

Ordinalita proměnných vyžaduje od metod splnění dvou vlastností:

1. <u>palindromická invariance</u> - výsledky se nezmění při změně pořadí na inverzní ($k \rightarrow K - k + 1$ tj. $A^* = (A^*_1, A^*_2, \ldots, A^*_K) = (A_K, A_{K-1}, \ldots, A_1)$);

2. <u>kladně monotnní invariance</u> - výsledky se nezmění pro jakoukoliv kvantifikaci kategorií, která je rostoucí funkcí indexu ($x_k = f(k)$, $x_{k+1} > x_k$).

Obě vlastnosti dohromady tvoří <u>ordinální invarianci metod</u>. Ne všechny postupy doporučované v literatuře mají tyto vlastnosti. Např. log-lineární modely jsou závislé na přijaté kvantifikaci, korelační koeficienty nejsou palindromicky invariantní (tento požadavek však pro korelační úlohu není vhodný).

V dalším budeme předpokládat, že pořadí kategorií u ordinálních proměnných je pevně dáno. Značení bude běžné pro kontingenční tabulky:

n_{rs}, n_{rst}, n_{r+}, n_{+s}, $n_{rs\ell}$ ...	jsou absolutní četnosti,
f_{rs}, f_{rst}, f_{r+}, ...	jsou relativní četnosti,
p_{rs}, p_{rst}, p_{r+}, ...	populační pravděpodobnosti,
π_{rs}, π_{rst}, π_{r+} ...	jsou hypotetické hodnoty,
F_k, P_k, Π_k	jsou postupně empirické, populační a hypotetické distribuční funkce,
$F_{k/r}$, $P_{k/r}$, $\Pi_{k/r}$	jsou podmíněné distribuční funkce v r-tém řádku,
F_{rs}, P_{rs}, Π_{rs}	jsou dvourozměrné distribuční funkce, a podobně.

Pro komparační tabulku je zvoleno značení rozměru R x K, pro asociační R x S resp. R x S x T, velikost výběru N resp. n.

1. Předpoklad existence spojité škály (C-modely, neparametrické skórování)

Pořadí $A_i \prec A_{i+1}$ je určeno spojitou škálou, tj. existuje spojitá proměnná X, ze které se odvozují (většinou neznámým způsobem) třídy A_i jako intervaly $\langle x_i, x_{i+1} \rangle$ pokrývající disjunktně její obor. Tento předpoklad vede k využití celé teorie neparametrických technik při povolení (velkého počtu) spojení v datech (za spojení jsou považována všechna pozorování ve stejné kategorii) a k analýze založené na četnostech.

A) Ridity, průměrné skóry pořadí, skóry Wilcoxonova typu

$r_k = F_{k-1} + \frac{1}{2} f_k = (F_{k-1} + F_k) / 2$ = ridit, procentový skór,

$M_k = Nr_k + \frac{1}{2}$ = průměrný skór pořadí.

a) Spearmanovo ρ pro tabulku A x B o rozměrech R x S
- určí se ridity nebo průměrná pořadí pro obě marginální rozložení A a B a přijmou se za kvantifikaci kategorií;
- pro tuto kvantifikaci se aplikuje vzorec pro Pearsonův lineární korelační koeficient.

b) Test rovnosti mediánů - Kruskal-Wallisova analýza rozptylu

Testová statistika

$$H = \left[\frac{12}{N(N+1)} \sum_{r=1}^{R} \frac{R_r^2}{N_r} - 3\,(N+1) \right] / H_s ,$$

kde

$$H_s = 1 - \sum_{k=1}^{K} \frac{n_k^3 - n_k}{N^3 - N} , \quad R_r = \sum_{k=1}^{K} n_{rk} w_k , \quad W_k = \sum_{i=1}^{k-1} n_i + \frac{n_k + 1}{2} ,$$

má asymptoticky rozdělení chí-kvadrát s R-1 stupni volnosti. Dále lze mediány porovnávat párově pomocí R(R-1)/2 Wilcoxonových testů s využitím simultánní inference.

K Wilcoxonovým testům existuje diskrétní analog (A.Agresti [1978]). Pro řádky i,j se použije A_{ij} (odhad populační míry α_{ij}), kde

$$A_{ij} = \sum_{k=1}^{K} (f_{k/j} F_{k/i} - f_{k/i} F_{k/j}),$$

$$\mathrm{var}_0 A_{ij} = \frac{N_i + N_j}{N_i N_j} \left[\sum_{k=1}^{K} f_k^* \, (F_{k-1}^* + F_k^* - 1)^2 \right],$$

$$f_k^* = \frac{n_k^*}{N_i + N_j} , \quad F_k^* = \sum_{i=1}^{k} f_i^* , \quad n_k^* = n_{ik} + n_{jk} .$$

A_{ij} je z intervalu $\langle -1, 1 \rangle$ a $A_{ij} / \sqrt{\mathrm{var}_0 A_{ij}}$ má asymptoticky standardní normální rozložení při $\alpha = 0$. Jestliže $i > j \Rightarrow A_{ij} \geq 0$, soubory jsou konzistentně uspořádané. Jestliže $i > j \Rightarrow \alpha_{ij} > 0$ (pomocí simultánní inference), soubory jsou úplně uspořádané na hladině α. Stupeň konzistence je měřen koeficientem uspořádanosti (A.Agresti [1978]).

c) Test rovnosti mediánů pro marginální rozložení ve čtvercové tabulce - (párová data)

Wilcoxonův test pro závislé výběry (Cureton [1967], Pratt [1959]): Testová statistika W má asymptoticky standardní normální rozložení,

$$W = \frac{2R - [n\,(n+1) - d_0\,(d_0 + 1)] / 2 \pm 1}{\{ [n\,(n+1)\,(2n+1) - d_0\,(d_0+1)\,(2d_0+1) - \frac{T}{2}] / 6 \}^{\frac{1}{2}}} ,$$

[illegible] čitatel kladný, odečteme jedničku, záporný, přičteme jedničku)

$$R = \sum_{q=1}^{K-1} d_q R_q + d_0 C , \quad T = \sum_{q=1}^{K-1} (D_q^3 - D_q) , \quad C = \sum_{q=1}^{K-1} d_q , \quad R_1 = \frac{D_1 + 1}{2} ,$$

$$R_2 = \frac{2D_1 + D_2 + 1}{2}, \ldots, R_q = \frac{2D_1 + 2D_2 + \ldots + 2D_{q-1} + D_q + 1}{2},$$

$$D_q = d_q + d_{-q}, \quad q = 1, \ldots, K-1 \quad ; \quad d_q = \sum_{j-i=q} n_{ij},$$

$$q = -(K-1), -(K-2), \ldots, -1, 0, 1, \ldots, K-2, K-1.$$

Znaménkový test:

$Z = (2C - \bar{n})/\sqrt{\bar{n}}$ má asymptoticky standardní normální rozložení,

$C = \sum\sum_{i<j} n_{ij}$, $D = \sum\sum_{i>j} n_{ij}$, $\bar{n} = C + D$.

d) Lineární modely pro ridity v analýze kontingenčních tabulek byly zkoumány v pracech Williams, Grizzle [1972], Forthofer, Koch [1973]. Vycházejí z postupu vážené regresní metody zavedené v práci Grizzle, Starmer, Koch [1969].

Přednosti přístupu A:

1. předpokládá se platnost neparametrické teorie a existuje tradice, lze stále čerpat z modelů neparametrické statistiky,
2. statistiky mají asymptoticky N (0,1) nebo centrální χ^2 rozložení,
3. jednoduchá heuristika a tudíž i snadné porozumění a interpretace,
4. platí vlastnosti invariance (kromě palindromie u ρ).

Nevýhody:

1. ridity jsou závislé na datech - vynechání řádků v tabulce vede na změnu skórování; při párovém porovnávání řádků vznikají nekonzistence (paradoxy),
2. rozšíření korelace na parciální nese potíže (kvantifikace se mění na podsouborech),
3. robustnost neparametrického modelu ani platnost pro velký počet spojení není prozkoumaná.

B) Kendallovy skóry pro dvojice pozorování

Zavedeme

$C = \sum_{i<k}\sum_{j<l} n_{ij}\, n_{kl}$ (= počet shod), $D = \sum_{i<k}\sum_{j>l} n_{ij}\, n_{kl}$ (= počet neshod),

$T_A = \frac{1}{2}\sum n_{i+}(n_{i+} - 1)$ (= počet spojení u A),

$T_B = \frac{1}{2}\sum n_{+j}(n_{+j} - 1)$ (= počet spojení u B),

$T_{AB} = \frac{1}{2}\sum\sum n_{ij}(n_{ij} - 1)$ (= počet spojení u obou),

$\frac{1}{2}\, n\,(n-1) = C + D + T_A + T_B - T_{AB}$,

Π_C (Π_D) je pravděpodobnost shody (neshody) u náhodně vybrané dvojice v populaci.

Definujeme: A,B jsou ordinálně nezávislé (ve smyslu Kendallových skórů), je-li $\Pi_C - \Pi_D = 0$, jsou kladně(přímo)ordinálně závislé, je-li $\Pi_C - \Pi_D > 0$, jsou záporně (nepřímo) ordinálně závislé, je-li $\Pi_C - \Pi_D < 0$.

Poznámka: statistická nezávislost $\Rightarrow$ ordinální nezávislost,
ordinální nezávislost $\not\Rightarrow$ statistická nezávislost,
ordinální závislost $\Rightarrow$ statistická závislost,
statistická závislost $\not\Rightarrow$ ordinální závislost.

Přístup poskytuje celou řadu koeficientů asociace a korelace a testů ordinální nezávislosti. Dále jsou koeficienty formulovány jako funkce výběrových dat, jejich populační definice jsou analogické.

a) Goodman-Kruskalovo γ

$$\gamma = (C - D) / (C + D), \quad (\in \langle -1, 1 \rangle)$$

b) Somersovo $d_{B|A}$ (asymetrická asociace $A \rightarrow B$)

$$d_{B|A} = (C - D) / \left(\frac{n(n-1)}{2} - T_A \right), \quad (\in \langle -1, 1 \rangle)$$

je analogií k regresnímu koeficientu.

c) Kendallovo τ_b

$$\tau_b = (C - D) / \left\{ \left[\frac{n(n-1)}{2} - T_A \right] \left[\frac{n(n-1)}{2} - T_B \right] \right\}^{\frac{1}{2}} = \pm \sqrt{d_{B|A} \cdot d_{A|B}} \ ,$$

sign τ_b = sign $d_{B|A}$, $\tau_b \in \langle -1, 1 \rangle$.

d) Agrestiho míry pro porovnání dvou distribucí

$$\Delta = \hat{P}(Y_2 > Y_1) - \hat{P}(Y_1 > Y_2) = \sum_{i<j}\sum f_{i/1} f_{j/2} - \sum_{i>j}\sum f_{i/1} f_{j/2} =$$

$= \Delta_{B|A}$ (= Somersovo d pro $A = \{1,2\}$ a $B = Y$), $\Delta \in \langle -1, 1 \rangle$,

$$\alpha = \left[\hat{P}(Y_2 > Y_1) \right] / \left[\hat{P}(Y_1 < Y_2) \right] = \left(\sum_{i<j}\sum f_{i/1} f_{j/2} \right) / \left(\sum_{i>j}\sum f_{i/1} f_{j/2} \right),$$

$\alpha \in (0, +\infty)$.

e) Test H_0 : A,B jsou ordinálně nezávislé proti H_1, (H_2; H_3) : A,B jsou ordinálně závislé (kladně závislé; záporně závislé)

$$Z = (C - D) / \sigma \overset{as}{\sim} N(0, 1) \text{ při } H_0 .$$

Pro $C \geq 100$ a $D \geq 100$ je možno použít přibližný odhad rozptylu

$$\hat{\sigma}^2 = (1 - \sum f_{i+}^3)(1 - \sum f_{+j}^3)\, n^3 / 9 \ .$$

(Pozn.: Nepoužívat pro intervaly spolehlivosti, jen pro test H_0 !)

f) Test parciální ordinální nezávislosti A,B při podmínkách $C = \{C_1, \ldots, C_T\}$, C nominální

$$\chi^2 = \sum Z_t^2 \overset{as}{\sim} \chi^2 (df = T) \text{ při } H_0, \quad Z_t = (C_t - D_t) / \sigma_t \ .$$

Výhody přístupu B:

1. platí teorie neparametrických technik a existuje tradice,
2. kladná monotonní invariance,
3. jednoduchá heuristika,
4. existuje rozšíření asociace na parciální,
5. invariance k rostoucímu ohodnocení kategorií.

Nevýhody:

1. není známo rozšíření na další souběžné úlohy,
2. neprozkoumaná robustnost neparametrického modelu a platnost teorie pro velký počet spojení .

C) Variabilita zavedená pomocí pořadí (Řehák [1976])

Vzájemná nepodobnost jednotek ve dvou kategoriích a_i, a_j ordinální proměnné A může být charakterizována rozdílem jejich riditů

$$d(a_i, a_j) = r_j - r_i = \frac{1}{2} f_i + \sum_{i+1}^{j-1} f_k + \frac{1}{2} f_j, \quad j > i, \quad d(a_i, a_j) = d(a_j, a_i)$$

pro $j < i$.

Celková variabilita je průměrem rozdílů: $\sum\sum f_i f_j d(a_i, a_j)$;

$$\text{corvar} = \sum f_m (F_m - \tfrac{1}{2})(F_{m-1} - \tfrac{1}{2}) + \tfrac{1}{4} \ .$$

Lze též definovat minimum průmerné vzdálenosti jednotek od pole m

$$Corvar = (F_M - \frac{1}{2})(F_{M-1} - \frac{1}{2}) + \frac{1}{4} ,$$

M je index mediánové kategorie. Podle principu redukce variability lze odvodit dvě míry pro asociaci nominální (B) a ordinální (A) proměnné. Z míry Corvar je to

$$\hat{\alpha}_{A/B} = 1 - \frac{\left[\sum_r f_{r+} \sum_k f_{k/r} (F_{k/r} - \frac{1}{2})(F_{k-1/r} - \frac{1}{2})\right] + \frac{1}{4}}{\left[\sum_k f_{+k} (F_k - \frac{1}{2})(F_{k-1} - \frac{1}{2})\right] + \frac{1}{4}} ,$$

z míry Corvar je to

$$\hat{\alpha}^{*}_{A|B} = 1 - \frac{\left[\sum f_{r+} (F_{M_r} - \frac{1}{2})(F_{M_r-1} - \frac{1}{2})\right] + \frac{1}{4}}{(F_M - \frac{1}{2})(F_{M-1} - \frac{1}{2}) + \frac{1}{4}}$$

kde M je index mediánové kategorie v celém spojeném souboru a M_r jsou indexy mediánových kategorií v řádcích tabulky.
Obe koeficienty jsou z $\langle 0, 1 \rangle$, jsou definované když existují alespoň dvě obsazené kategorie; nezávislost $\Rightarrow \alpha = 0$, $\alpha^{*} = 0$; $\alpha = 0 \Rightarrow$ nezávislost; $\alpha = 1$, $\alpha^{*} = 1 \Rightarrow$ v každém řádku je obsazena jediná kategorie.

Výhody:

1. řádky tabulky (podmíněné distribuce) jsou posuzovány podle vlastních riditů, nejde o průměrný skór vycházející z marginálních četností,
2. model odpovídá heuristicky situaci, v níž je uspořádání tříd určeno spojitou škálou a třídy samotné vznikají v procesu získávání dat (měření) jako nerozlišitelná spojení,
3. platí ordinální invariance.

Nevýhody:

1. nerozvinutá teorie pro další úlohy, např. je otevřená otázka parciální asociace, testů shody, vztah k modelům asociačních vztahů u více proměnných, vztah k modelům typu A,
2. není prozkoumána vhodnost pro jiné typy ordinálních proměnných (dané kategorie, diskrétní model) a nejsou známé teoretické vlastnosti.

II. Diskrétní uspořádané kategorizace (D-modely, distribuční funkce)

Pořadí kategorií je dáno předem a je totožné s indexováním $A = \{A_1, A_2, \dots, A_K\}$, předpoklad spojité latentní proměnné se nezavádí, naopak se předpokládá, že A_i jsou samostatné třídy, uvnitř nichž je pořadí nejen nerozlišitelné, ale také se (modelově) ani nerozlišuje. Úlohy, které se řeší pomocí tohoto přístupu, vycházejí většinou z distribuční kumulativní funkce a z příbuzných charakteristik. Uvádíme především metody vycházející z D-modelu, kde je ordinální proměnná speciálním případem. Všechny výsledky jsou invariantní ke všem kvantifikacím $g = g(A_i)$, kde g je libovolná rostoucí funkce indexu i .

A) Ordinální proměnná jako speciální případ obecného D-modelu

Model byl zaveden v práci Řehák [1976], rozpracován a shrnut v práci Řeháková [1985].

a) Popis jednorozměrné distribuce (zavedení charakteristik):

Dorvar $= 2 \sum F_k (1 - F_k)$ (míra variability),

mediánová kategorie $= A_M \Rightarrow F_{M-1} \leq 0.5$, $F_M > 0{,}5$,

ordinální medián $= M + 0.5 - (F_M - 0.5) / f_M$,

centralita pole $A_k = \sum_i |k-i| f_i$,

d^M = Dorvar = centralita mediánové kategorie (míra variability kolem A_M) .

b) Test dobré shody $H_0 : P = \Pi$

Rozložení testové statistiky viz B. Řeháková (tento sborník).

c) Test homogenity dvou distribucí $H_0 : P_1 = P_2$, obdobně k b).

d) Test homogenity R distribucí $H_0 : P_1 = P_2 = \dots = P_R$

(též test nezávislosti A na B, kde B je nominální), obdobně k b).

e) Koeficienty asymetrické asociace nominální proměnné B k ordinální proměnné A (B → A)

$$\beta_{A/B} = 1 - \frac{\sum_r f_{r+} \sum_k F_{k/r} (1 - F_{k/r})}{\sum_k F_k (1 - F_k)} ,$$

$$\beta^*_{A/B} = 1 - \frac{\sum_r f_{r+} \sum_k f_{k/r} |k - M_r|}{\sum_k f_k |k - M|} .$$

Koeficient $\hat{\beta}_{A/B} = 1 - \nu/\xi$ má asymptoticky rozložení $N(\beta_{A/B}, \sigma^2)$,

$$\hat{\sigma}^2 = \frac{1}{n\xi^4} \sum_{b=1}^{R} \sum_{a=1}^{K} f_{ba} \left[\nu (2d^*_a - \xi) - \xi (2d^*_{a/b} - \text{dorvar } f_{(b)}) \right]^2 ,$$

$$d^*_a = \sum_{k=1}^{K} f_{+k} |a-k| , \quad d^*_{a/b} = \sum_{k=1}^{K} f_{k/b} |k - a| .$$

Rozšíření na parciální asociaci viz Řeháková [1983], [1985], Řehák, Řeháková [1984a], [1986].

f) Testy rovnosti marginálních distribucí ve čtvercové tabulce

$H_0 : P_{k+} = P_{+k}$, $k = 1, \dots, K-1$ (Řehák, Řeháková [1984b])

Analogie Stuartova testu:

$$Q_H = \bar{\Delta}' V^{-1} \bar{\Delta} \overset{as}{\sim} \chi^2_{K-1} ,$$

$\bar{\Delta}$ = vektor $(K-1) \times 1$, $\bar{\Delta}_k = n (F_{k+} - F_{+k})$, V = matice $(K-1) \times (K-1)$,

$V_{jk} = n (F_{j+} + F_{+j} - F_{jk} - F_{kj})$ pro $j \leq k$, $V_{jk} = V_{kj}$ pro $j > k$.

Analogie Bhaparova testu:

$$Q_B = \bar{\Delta}' W^{-1} \bar{\Delta} \overset{as}{\sim} \chi^2_{K-1} ,$$

W = matice $(K-1) \times (K-1)$, $W_{jk} = V_{jk} - n (F_{j+} - F_{+j}) (F_{k+} - F_{+k})$.

Simultánní inference pomocí K-1 Mc Nemarových testů:

$$Q_k = n \frac{(F_{k+} - F_{+k})^2}{F_{k+} + F_{+k} - 2F_{kk}} \overset{as}{\sim} \chi^2_1 , \quad k = 1, \dots, K-1$$

za využití postupných hladin významnosti plynoucích z Holmovy metody.

Distanční statistika: $2n \sum_{k}^{K-1} (F_{k+} - F_{+k})^2$ viz B. Řeháková (tento sborník).

g) Párové porovnávání R distribucí

- simultánní inference pro $R(R-1)/2$ dvojic pomocí c);
- seskupovací algoritmy pomocí vzdáleností, jejichž čtverce jsou dány jako

$$\sum_{k=1}^{K-1} (F_{k/r} - F_{k/r'})^2 ;$$

- škálování párových vzdáleností (viz Řehák, Loučková [1984], [1985] .

B) Plackettův model stejnosměrné globální asociace

Za předpokladu konstantního globálního součinového poměru v tabulce R x S

$$\vartheta_{rs}^{*} = \log \frac{P_{rs}\,(1 + P_{rs} - P_{r+} - P_{+s})}{(P_{r+} - P_{rs})\,(P_{+s} - P_{rs})} = \vartheta$$

a pro dané marginální distribuce P_r^X , P_s^Y definoval Plackett [1965] třídu dvourozměrných funkcí ($\vartheta \neq 1$)

$$P_{rs}^{XY} = \frac{1}{2(\vartheta-1)}\Big[\,1 + (\vartheta-1)\,(P_r^X + P_s^Y) - \Big\{\big[1 + (\vartheta-1)\,(P_r^X + P_s^Y)\big]^2 -$$

$$- 4\,\vartheta\,(\vartheta-1)\,P_r^X\,P_s^Y\Big\}^{\frac{1}{2}}\Big] .$$

Wahrendorf [1980] ukázal, jak získáme odhad ϑ metodou vážených nejmenších čtverců a jak testovat $\vartheta_{rs}^{*} = \vartheta$. Přístup byl dále zobecněn v Semenya, Koch [1980].

C) Další metody založené na distribučních funkcích

mohou vycházet např. z diskrétní formy Kolmogorov-Smirnovovy statistiky $\max_k |F_k - G_k|$ resp. $\max (F_k - G_k)$, nebo z manhatanské vzdálenosti distribučních funkcí $\sum |F_k - G_k|$ . .

D) Modely kumulačního efektu

jsou konstruovány pro situaci jedné závislé ordinální proměnné a explanačních proměnných nominálního nebo kardinálního typu. Využívají různě transformované distribuční funkce závisle proměnné.

Logitové modely: $\log P_k/(1 - P_k) = \alpha_k + \beta' x$,

probitové modely: $\Phi^{-1}(P_k) = \alpha_k + \beta' x$,

log-log modely: $\log (- \log (1 - P_k)) = \alpha_k + \beta' x$,

kde P_k je distribuční funkce závisle proměnné, α_k a β jsou neznámé parametry, x jsou známé skóry.

Z těchto tří modelů jsou nejvíce rozvíjeny logitové, které (ač svojí podstatou patří sem) jsou uvedeny ve skupině E-modelů vzhledem k jejich příbuznosti s log-lineárními modely a vzhledem ke zvyklostem v literatuře (především v monografii A. Agresti [1984]).

III. Přiřazení externích pevných čísel kategoriím (E-modely, kardinalita)

Přiřazení skórů $x_i = x\,(A_i)$ se většinou provádí pseudokvantifikací $x_i = i$ Postupy jsou však v literatuře formulovány obecně. Kvantifikace vede na kardinální proměnné (číselnou klasifikaci). Výsledky nejsou zpravidla ordinálně (ani kladně monotónně) invariantní.

A) Využití metod pro číselné proměnné - lineární model a metody

Využití metod mnohorozměrné statistické metodologie pro kategorizovaná kvantifikovaná data je v praxi běžnou věcí. Tento přístup je v literatuře doporučován i (a to dosti často) odmítán. Tam, kde očekáváme monotonní závislosti a pravidelnou strukturu datových vztahů, dává postup užitečné a přijatelné výsledky, při nepravidelných strukturách však můžeme dostat zcela mylné závěry.

Typické použití: Pearsonův lineární korelační koeficient pro kontingenční tabulku aplikovaný na kódy kategorií a jeho využití pro parcializaci a faktorovou analýzu, dále pak výpočty průměrů, analýza rozptylu, diskriminační analýza, kanonické korelace, regresní analýza. V praxi se osvědčuje pro proměnné s větším počtem kategorií a pro problémy s obsahově homogenními proměnnými.

Výhody:
jednoduchá heuristika a snadné využití standardních metod.

Nevýhody:
1. nejistota o vhodnosti volby kvantifikace,
2. nejistota o platnosti statistických výsledků (robustnost metod k tomuto typu proměnných není dostatečně prozkoumána).

Doporučuje se opakovat výpočty pro několik odlišných kvantifikací a porovnat závěry. Shodný výsledek posiluje důvěru v závěr, různé výsledky negují aplikabilitu ale na druhé straně dávají možnost komparačního rozboru a interpretace.

B) <u>Logaritmicko-lineární modely</u>

Log-lineární modely se tvoří pomocí aditivních lineárních vztahů s neznámými parametry pro logaritmy očekávaných četností $m_{rs\ldots q}$ v jednotlivých polích. Analýza postupuje od odhadu parametrů zvoleného modelu k výpočtu odhadů očekávaných četností $\hat{m}$ (resp. $\log \hat{m}$) a k testování platnosti modelu pomocí statistiky

$$G^2 = 2 \sum n \log n/\hat{m},$$

kde se sčítá přes všechna pole. G^2 má rozdělení χ^2 s df = počet polí - počet nezávislých lineárních podmínek svazujících parametry.

Log-lineární model (saturovaný) pro tabulku R x S <u>třídění A x B</u> je

$$\log m_{rs} = \mu + \lambda_r^A + \lambda_s^B + \lambda_{rs}^{AB} ,$$

kde m_{rs} = očekávaná hodnota četnosti v poli (r, s),

λ_r^A, λ_s^B = marginální efekty řádků a sloupců, $\sum \lambda_r^A = \sum \lambda_s^B = 0$,

λ_{rs}^{AB} = interakční efekt, $\sum_r \lambda_{rs}^{AB} = \sum_s \lambda_{rs}^{AB} = 0$.

Model nezávislosti: $\lambda_{rs}^{AB} = 0$ pro všechna r, s.

Pro <u>tabulku A x B x C</u> o rozměru R x S x T je saturovaný model

$$\log m_{rst} = \mu + \lambda_r^A + \lambda_s^B + \lambda_t^C + \lambda_{rs}^{AB} + \lambda_{rt}^{AC} + \lambda_{st}^{BC} + \lambda_{rst}^{ABC} .$$

Vynecháním jednotlivých členů specifikujeme různé asociační modely. Např. $\lambda_{rst}^{ABC} = \lambda_{rs}^{AB} = 0$ pro všechna r,s,t, specifikuje model parciální asociace (A,B/C) nebo řetězení asociací A-C-B, který v symbolice log-lineárních modelů zapisujeme jako (AC, BC) a pro nějž je odhad očekávaných četností

$$\hat{m}_{rst} = n_{r+t}\; n_{+st} / n_{++t}$$

a testová statistika

$$G^2 = 2 \sum\sum\sum n_{rst} \log (n_{rst} / \hat{m}_{rst}) \sim \chi^2 \quad (df = (R-1)(S-1)T).$$

Úplnou nezávislost testujeme tak, že

$$G^2(I) = 2 \sum\sum\sum n_{rst} \log(n_{rst}/\hat{m}_{rst}) \sim \chi^2 \quad (df = RST-R-S-T+2),$$

$$\hat{m}_{rst} = n_{r++}\; n_{+s+}\; n_{++t} / n^2.$$

Pro log-lineární modely je vhodné zavést pojem <u>interakce</u> měřené pomocí logaritmu <u>součinového podílu</u> (poměru šancí, odds ratio). Pro tabulku 2 x 2 je součinový podíl definován jako

$$\vartheta = p_{11}\, p_{22} / p_{12}\, p_{21}$$

a je odhadován pomocí

$$\hat{\vartheta} = n_{11}\, n_{22} / n_{12}\, n_{21} \quad , \qquad \text{nebo}$$

$$\hat{\vartheta} = (n_{11} + 0.5)\,(n_{22} + 0.5) / (n_{12} + 0.5)\,(n_{21} + 0.5),$$

je-li některá četnost nulová. Intervaly spolehlivosti se určují pomocí logaritmické transformace s

$$\hat{6}\,(\log \hat{\vartheta}\,) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)^{\frac{1}{2}}$$

(nebo n_{ij} jsou nahrazeny čísly $n_{ij} + 0.5$).

Pro tabulku R x S můžeme vytvořit $\binom{R}{2}\binom{S}{2}$ možných čtyřpolních podtabulek a pro každou určit ϑ . Pro popis volíme základní množinu součinových poměrů, kterých je (R-1)(S-1) a ze kterých jsou ostatní odvoditelné. Pro ordinální proměnné používáme několik typů poměrů:

a) Lokální poměry $\quad \vartheta_{rs} = p_{rs}\, p_{r+1,s+1} / p_{r,s+1}\, p_{r+1,s}$

vzniklé ze všech (R-1)(S-1) podtabulek se sousedními sloupci a řádky.

b) Lokálně-globální poměry

$$\vartheta'_{rs} = \frac{P_{s/r}(1 - P_{s/r+1})}{P_{s/r+1}\,(1 - P_{s/r})} = \frac{\sum\limits_{i \le s} p_{ri} \quad \sum\limits_{i > s} p_{r+1,i}}{\sum\limits_{i \le s} p_{r+1,i} \quad \sum\limits_{i > s} p_{ri}} \quad .$$

Platí, že $\log \vartheta'_{rs} \geq 0$ pro $s = 1, \ldots, S-1 \Leftrightarrow P_{s/r} \geq P_{s/r+1}$ (distribuce v řádku r je stochasticky menší než distribuce v řádku r+1).

c) Globální poměry

$$\vartheta''_{rs} = \frac{P_{rs}\,(1 + P_{rs} - P_{r+} - P_{+s})}{(P_{r+} - P_{rs})\,(P_{+s} - P_{rs})} \quad .$$

Ordinální data analyzujeme pomocí log-lineárních modelů při kvantifikaci $u_r = u(A_r)$, $v_s = v(B_s)$, $w_t = w(C_t)$. Často se využívá pseudokvantifikace $u_r = r$, $v_s = s$, $w_t = t$.

a) R x S tabulka, model stejnoměrné asociace, obě proměnné ordinální

Zavedení skórů umožňuje formulovat lineární vztah pro interakční členy

$$\lambda^{AB}_{rs} = \beta\,(u_r - \bar{u})\,(v_s - \bar{v}) \ ,$$

kde jediný neznámý parametr je β ($\bar{u}$, $\bar{v}$ jsou průměrné skóry):

$$\log m_{rs} = \mu + \lambda_r^{A} + \lambda_s^{B} + \beta\,(u_r - \bar{u})\,(v_s - \bar{v}) \ .$$

Statistika

$$G^2(U) = 2 \sum\sum n_{rs} \log\,(n_{rs} / \hat{m}_{rs}) \sim \chi^2 \;(df = RS - R - S).$$

Případ $\beta = 0$ je nezávislost. Název modelu je odvozen od toho, že za platnosti jeho předpokladu jsou všechny logaritmy lokálních součinových poměrů rovny β pro jednotkové ekvidistantní skóry, jinak je hodnota úměrná rozdílům hodnot kategorií

$$\log \vartheta_{rs} = \beta\,(u_{r+1} - u_r)\,(v_{s+1} - v_s).$$

b) R x S tabulka, model řádkových efektů, A nominální, B ordinální

$$\log m_{rs} = \mu + \lambda_r^{A} + \lambda_s^{B} + \tau_r\,(v_s - \bar{v}), \qquad \sum \tau_r = 0,$$

$$G^2(R) \sim \chi^2 \left(df = (R-1)\,(S-2)\right) ,$$

$\{\tau_r\}$ jsou řádkové efekty lineárního trendu.

Pro libovolné součinové poměry je

$$\log\,(m_{ac}\, m_{bd} / m_{ad}\, m_{bc}) = (\tau_b - \tau_a)\,(v_d - v_c)$$

a pro přilehlé sloupce v případě jednotkové ekvidistantní stupnice v_s je tento poměr roven $\tau_b - \tau_a$ pro všechna c (d = c+1) a tudíž $\tau_b - \tau_a$ je přirozený parametr rozdílnosti řádkových efektů.

Obdobně model sloupcových efektů

$$\log m_{rs} = \mu + \lambda_r^A + \lambda_s^B + \rho_s(\mu_r - \bar{\mu}),\ \Sigma \rho_s = 0,$$

$$G^2(S) \sim \chi^2 \ (df = (R-2)(S-1)).$$

Podmíněné testy nezávislosti

$$G^2(I/R) = G^2(I) - G^2(R) \sim \chi^2 \ (df = R - 1),$$

$$G^2(I/S) = G^2(I) - G^2(S) \sim \chi^2 \ (df = S - 1).$$

c) <u>R x S x T tabulka, model stejnoměrné interakce, A,B,C ordinální</u>

$$\log m_{rst} = \mu + \lambda_r^A + \lambda_s^B + \lambda_t^C + \beta^{AB}(u_r - \bar{u})(v_s - \bar{v}) + \beta^{AC}(u_r - \bar{u})(w_t - \bar{w}) + \beta^{BC}(v_s - \bar{v})(w_t - \bar{w}) + \beta^{ABC}(u_r - \bar{u})(v_s - \bar{v})(w_t - \bar{w}),$$

$$G^2(U) \sim \chi^2 \ (df = RST - R - S - T - 2).$$

Název se odvozuje z toho, že pro všechny lokální interakce v třírozměrných tabulkách 2 x 2 x 2 sousedních řádků, sloupců a vrstev je

$$\log \theta_{rst} = \log(\theta_{rs/t+1} / \theta_{rs/t}) = \beta^{ABC},$$

$$\log \theta_{rs/t} = \beta^{AB} + \beta^{ABC}\left[t - \frac{T+1}{2}\right].$$

Třírozměrné interakce jsou konstantní, interakce v podmíněných tabulkách jsou též konstantní, ale mění se postupně lineárně s indexem t.

d) <u>R x S x T tabulka, model homogenní stejnoměrné asociace, A,B,C jsou ordinální</u>

$$\log m_{rst} = \mu + \lambda_r^A + \lambda_s^B + \lambda_t^C + \beta^{AB}(u_r - \bar{u})(v_s - \bar{v}) + \beta^{AB}(u_r - \bar{u})(w_t - \bar{w}) + \beta^{BC}(v_s - \bar{v})(w_t - \bar{w}),\quad G^2 \sim \chi^2 \ (df = RST - R - S - T - 1).$$

Za předpokladu ekvidistantních skupin platí, že

$$\log \theta_{rs(t)} = \beta^{AB},\ \log \theta_{r(s)t} = \beta^{AC},\ \log \theta_{(r)st} = \beta^{BC},$$

kde index v závorce znamená podmínku, za níž jsou součinové poměry zjišťovány. Všechny podmíněné tabulky A x B za $C = C_t$ vykazují tak stejnoměrnou asociaci a její stupeň je na t nezávislý (obdobně pro každou dvojici proměnných podmíněnou třetí proměnnou).

e) <u>R x S x T tabulka, model řádkových efektů, A nominální, B,C ordinální</u>

$$\log m_{rst} = \mu + \lambda_r^A + \lambda_s^B + \lambda_t^C + \tau_r^{AB}(v_s - \bar{v}) + \tau_r^{AC}(w_t - \bar{w}) + \beta^{BC}(v_s - \bar{v})(w_t - \bar{w}),$$

$$\Sigma \lambda^A = \Sigma \lambda^B = \Sigma \lambda^C = \Sigma \tau^{AB} = \Sigma \tau^{AC} = 0.$$

V tomto modelu je

$$\log \theta_{rst} = 0,\ \log \theta_{(r)st} = \beta^{BC}$$ (stejnoměrná podm. asociace),

$$\log \theta_{r(s)t} = \tau_{r+1}^{AC} - \tau_r^{AC},\ \log \theta_{rs(t)} = \tau_{r+1}^{AB} - \tau_r^{AB},$$

τ_r^{AB} = řádkové efekty A na B v $A \longleftrightarrow B$ jsou homogenní pro různé úrovně C,

τ_r^{AC} = řádkové efekty A na C v $A \longleftrightarrow C$ jsou homogenní pro různé úrovně B,

$$G^2 \sim \chi^2 \ (df = RST - 3R - S - T + 3).$$

f) <u>R x S x T tabulka, model řádkových efektů, A,B nominální, C ordinální</u>

$$\log m_{rst} = \mu + \lambda_r^A + \lambda_s^B + \lambda_t^C + \lambda_{rs}^{AB} + \tau_r^{AC}(w_t - \bar{w}) + \tau_s^{BC}(w_t - \bar{w}),$$

$$\sum\lambda^{A} = \sum\lambda^{B} = \sum\lambda^{C} = \sum\lambda^{AB} = \sum\tau^{AC} = \sum\tau^{BC} = 0,$$

$G^2 \sim \chi^2$ (df = RST - RS - R - S - T + 3).

g) R x R tabulky: quasisymetrie

Model pro quasisymetrii ve čtvercové tabulce s nominálními daty je

$$\log m_{rs} = \mu + \lambda_r^{A} + \lambda_s^{B} + \lambda_{rs}, \quad \sum\lambda^{A} = \sum\lambda^{B} = \sum_{s}\lambda_{rs} = 0,$$

$$\lambda_{rs} = \lambda_{sr}, \quad G^2 \sim \chi^2 \text{ (df = (R-1)(R-2) / 2)}.$$

Speciální případ je symetrie pro nominální vstupy

$$\log m_{rs} = \mu + \lambda_r + \lambda_s + \lambda_{rs}, \quad G^2 \sim \chi^2 \text{ (df = R(R-1)/2)}.$$

U ordinálních dat uspořádání kategorií určuje také uspořádání vedlejších diagonál. Z toho vycházejí modely:

Model diagonálních parametrů (Goodman [1972, 1979])

$m_{rs} = m_{sr}\,\delta_{s-r}$, $r < s$, pro parametry $\delta_1, \delta_2, \dots, \delta_{R-1}$,

$G^2 \sim \chi^2$ (df = (R-1)(R-2) / 2) .

Test lze provést také tím, že sestrojíme R-2 kontingenčních tabulek T_r o rozměru 2x(R-r), r=1,...,R-2, v nichž řádky jsou vždy obě stejně dlouhé vedlejší diagonály. Za platnosti modelu jsou T_r nezávislé a společný podíl očekávaných četností ve sloupcích pro každou tabulku je $\hat{\delta}_r$. Odhady $\hat{m}_{rs}$ se zjišťují jako očekávané četnosti za nezávislosti v T_r.

Model podmíněné symetrie (Mc Cullagh [1978], Bishop [1975])

$m_{rs} = \delta\, m_{sr}$, $r < s$ (speciální případ Goodmanova modelu s $\delta_r = \delta$).

Název souvisí s tím, že pro $r < s$

$P(A=A_r, B=B_s / A < B) = P(A=A_s, B=B_r / A > B)$.

$G^2 \sim \chi^2$ (df = (R+1)(R-2) /2) ; $\hat{\delta} = \left(\sum\sum_{r<s} n_{rs}\right) / \left(\sum_{r>s}\sum n_{rs}\right)$,

$\hat{m}_{rs} = (n_{rs} + n_{sr}) / (\hat{\delta} + 1)$ pro $r > s$, $\hat{m}_{rs} = \hat{\delta}\,\hat{m}_{sr}$ pro $r < s$.

Model dvourozměrného normálního rozložení (Agresti [1983])

$m_{rs} = m_{sr}\,\delta^{r-s}$, $r \geq s$ (speciální případ Goodmanova modelu s $\delta_k = \delta^k$),

za předpokladu normálního rozložení latentních proměnných pro párová data. Log-lineární model je

$$\log m_{rs} = \mu + \lambda_r + \lambda_s + \beta(r-s) + \lambda_{rs}, \quad \lambda_{rs} = \lambda_{sr}, \ \sum\lambda_r = \sum_{r}\lambda_{rs} = 0,$$

$$2\beta = \log\delta, \quad G^2 \sim \chi^2 \text{ (df = (R+1) (R-2) / 2)} .$$

Odhady lze získat pomocí Newton-Raphsonovy metody, β lze odhadnout také z logitového vztahu

$$\log (m_{rs} / m_{sr}) = 2\beta(r-s).$$

Odhad parametrů log-lineárního modelu

Maximálně věrohodné rovnice, které poskytují odhady $\hat{m}$, se řeší

1. pro nominální proměnné metodou IPF (iterative proportinal fitting) (Deming, Stephan [1940]); pro hierarchické modely dává metoda, má-li řešení, ML-odhady (Birch [1963]).
2. pro ordinální proměnné jsou uváděny

- jednorozměrná Newton-Raphsonova metoda (Goodman [1979]),
- obecná Newton-Raphsonova metoda (Nelder, Wedderburn [1973]), Haberman [1974],
- metoda iterativního škálování (Is-iterative scaling method) (Darroch, Rattcliffe [1972])

Přednosti log-lineárního modelu:

1. vybudována standardní teorie,
2. možnost formulace různých modelů podle situace,
3. existují programy (SPSSX, MULTIQUAL, ANOAS, GLIM),
4. testové statistiky mají asymptoticky rozložení chí-kvadrát a mají aditivní vlastnosti, residua mají normální rozložení,
5. možnost inference o jednotlivých parametrech, možnost seskupování parametrů a budování modelů.

Nevýhody:

1. závislost na kvantifikaci (nejde de facto o ordinální ale kardinální metodologii, včetně možností studia i jiných než lineárních vazeb),
2. složité odhady parametrů.

C) <u>Logitové modely</u>

jsou vhodné tam, kde zkoumáme vliv jedné nebo více kategorizovaných proměnných na ordinální proměnnou. Ve skupině III metod jsou uváděny vzhledem ke zvyklostem a k příbuznosti s log-lineárními modely. Bylo by přirozenější uvést je ve skupině II, neboť závislá ordinální proměnná není v modelech kvantifikována. Kvantifikovány jsou však vždy nezávislé ordinální proměnné.

<u>Logit četnosti</u> p je určen jako logit $p = \log(p/(1-p))$. Vztah četnosti p k nějaké proměnné x (jednorozměrné nebo vektorové) můžeme vyjádřit pomocí <u>logistické regresní křivky</u>

$$\text{logit } p(\underset{\sim}{x}) = f(\underset{\sim}{x}/\underset{\sim}{b}),$$

kde f je daná funkce s neznámými parametry.

Pro dichotomickou proměnnou $C = (0,1)$ a kardinální proměnnou A se skóry $v_s = v(A_s)$ můžeme napsat rovnici :

$$\text{logit } P_s = \alpha + \beta v_s, \quad s = 1,2,\dots,S, \quad \text{kde } P_s = P(C=1/A = A_s).$$

Pro vazbu <u>A,B → C, A,B nominální</u>, můžeme formulovat model

$$\text{logit } P_{2(rs)} = \log(m_{rs2}/m_{rs1}) = \alpha + \tau_r^A + \tau_s^B, \quad \sum\tau^A = \sum\tau^B = 0,$$

$$G^2 \sim \chi^2 \quad (df = (R-1)(S-1)) .$$

Pro <u>A,B → C, A nominální, B ordinální</u>, můžeme zavést vztah

$$\text{logit } P_{2(rs)} = \log(m_{rs2}/m_{rs1}) = \alpha + \tau_r^A + \beta^B(v_j-\bar{v}), \qquad \sum\tau^A = 0,$$

$$G^2 \sim \chi^2 \quad (df = R(S-1)-1).$$

Pro vícehodnotovou proměnnou $A = \{A_1,\dots,A_K\}$ je možné vytvořit $\binom{K}{2}$ logitů typu

$$\log \frac{p_j / (p_j+p_k)}{p_k / (p_j+p_k)} = \log \frac{p_j}{p_k} .$$

K popisu však stačí <u>báze</u> o počtu K-1 základních logitů

$$L_j = \log p_j/p_K , \quad j = 1,\dots,K-1.$$

Pro ordinální proměnné tvoříme jiné typy logitů, které odrážejí uspořádanost kategorií:

<u>kumulovaný logit</u> $L_j = \log\left[\dfrac{P_{j+1}+\dots+P_K}{P_1+\dots+P_j}\right] = -\log\left(\dfrac{P_j}{1-P_j}\right)$,

<u>přírůstkový logit</u> $L_j' = \log\left[\dfrac{P_{j+1}}{P_1+\dots+P_j}\right] = \log\left(\dfrac{P_{j+1}}{P_j}\right)$

sousedský logit $L_j'' = \log (p_{j+1} / p_j)$ (tvoří bázi)

Používá se především kumulovaný logit.

Logitové modely pro ordinální proměnné lze formulovat obdobně jako log-lineární.

a) Model stejnoměrné asociace, tabulka AxB s rozměry RxK, AiB ordinální

$$L_{k(r)} = \log \frac{m_{r,k+1}+\dots+m_{r,K}}{m_{r,1}+\dots+m_{r,k}} = \alpha_k + \beta(u_r-\bar{u}), \quad r = 1,\dots,R$$

$G^2 \sim \chi^2$ (df = RK - R - K).

Protože $L_{1(r)} \geq L_{2(r)} \geq \dots \geq L_{K-1(r)} \Rightarrow \{\alpha_j\}$ je nerostoucí. $\{-\alpha_j\}$ jsou často interpretovány jako hranice kategorií A.

$L_{k(r+1)} - L_{k(r)} = \beta$ (= lokálně-globální součinový poměr).

Za předpokladu platnosti modelu lze podmíněně testovat $H_0: \beta = 0$ obdobně jako u log-lineárního modelu pomocí

$G^2(I/U) = G^2(I) - G^2(U)$, df = 1.

b) Model řádkových efektů, tabulka A x B o rozměrech RxK, A nominální (řádky), B ordinální (sloupce)

$L_{k(r)} = \alpha_k + \tau_r$, df = (R-1)(K-2), rozdíl logitů: $L_{k(b)} - L_{k(a)} = \tau_b - \tau_a$.

Podmíněný test nezávislosti:

$G^2(I/R) = G^2(I) - G^2(R) \sim \chi^2$ (df = R-1).

c) Tabulka [B x C] x A se dvěma explanačními proměnnými B,C o rozměrech R x S x K a závislou proměnnou A

Podmíněné kumulační logity

$$L_{k(rs)} = \log \frac{m_{rs,k+1}+\dots+m_{rsK}}{m_{rs1}+\dots+m_{rsk}}, \quad k = 1, \dots, K-1.$$

Modely nezahrnující vliv interakcí (B,C) mají tvar

$L_{k(rs)} = \alpha_k + (B) + (C)$, kde (B), (C) se určí podle typu proměnné z tabulky .

B	C	(B)	(C)	df pro testování platnosti modelu
ord.	ord.	$\beta^B(u_r-\bar{u})$	$\beta^C(w_s-\bar{w})$	RSK - RS - K - 1
nom.	ord.	τ_r^B	$\beta^C(w_s-\bar{w})$	RSK - RS - K - S - 1
nom.	nom.	τ_r^B	τ_s^C	RSK - RS - R - S - K + 3

Pozn.: $\{u_r\}$ je kvantifikace B, $\{w_s\}$ je kvantifikace C.

Logitové modely se odhadují
- metodou vážených nejmenších čtverců (GENCAT; Williams, Grizzle [1972])
- Newton-Raphsonovou metodou (MULTIQUAL; McCullagh [1980])
- pro spojitou explanační proměnnou X lze použít ML-odhady (Cox [1970]).

Výhody:
1. pro závislou ordinální proměnnou je možno vytvářet vhodné modely podle potřeby,
2. testové statistiky mají chí-kvadrát rozložení,
3. pro aplikace lze využít jak standardní programové systémy, tak řadu speciálních programů.

Nevýhody:
1. explanační proměnné nejsou chápány ordinálně ale kardinálně (problém skórování a absence ordinality metod),

2. složitý postup odhadu parametrů.

Z á v ě r

Přehled metod odpovídající současnému stavu ukazuje jednak na vývoj v oblasti metod analýzy dat, jednak na šíři úloh, které jsou v praxi řešeny. Zároveň je též vidět, že existují velmi rozmanité přístupy. Vývoj metodologie se zdá být teprve v první etapě. V následující tabulce je shrnutí současných možností modelů.

	C - modely	D - modely	R - modely
Kategorie určeny	pevně nebo podle výskytu spojení	pevně	pevně, jsou kvantifikovány nezávisle na datech
Invariance	palindromie (kromě korelací), monotonie	palindromie monotonie	palindromie a monotonie jen u logit
Popis distribuce	symetrie, shoda s Me	ano	číselné charakteristiky
Testy dobré shody	-	ano	ano
Homogenita	mediány	ano	ano
Marginální shoda závislých distribucí	mediány	ano	ano
Strukturní symetrie ve čtvercové tabulce	-	-	ano
Koeficienty asociace	korelace (i parciální)	asociace (i parciální)	korelační poměr korelační koeficient
Komplexní asociační modely	parciální korelace	ordinální závislé proměnné	ano
Regrese, ANOVA	-	ordinální závislé proměnné	převedením na klasický model
Seskupování distribucí	párové porovnávání a řazení mediánů	ano	=
Škálování	-	ano	=

LITERATURA (výběr):

Agresti A. [1979]. "Describing Differences on Ordered Categorical Response", Technical Report No 137, University of Florida

Agresti A. [1984]: Analysis of Ordinal Categorical Data, John Wiley & Sons, New York

Birch M.V. [1963]: "Maximum Likelihood in Three-Way Contingency Tables", JRSSB 25, 220-233

Cox D.R. [1970]: The Analysis of Binary Data, London, Chapman and Hall

Cureton E. [1967]: "The Normal Approximation to the Signed-Rank, Sampling Distribution When Zero Differences Are Present", JASA 62, 1068-1069

Darroch J.N., Ratcliff D. [1972]: "Generalized Iterative Scaling for Log-Linear Models", AMS 43, 1470-1480

Deming W.E., Stephan F.F. [1940]: "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known", AMS 11, 427-444

Forthofer R.N., Koch G.G. [1973]: "An Analysis for Compounded Functions of Categorical Data", Biometrics 29, 143-157

Goodman L.A. [1972]: "Some Multiplicative Models for the Analysis of Cross-Classified Data", Proc. 6-th Berkeley Symp. Vol.1, Berkeley, University of California Press, 649-696

Goodman L.A. [1979]: "Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories", JASA 74, 537-552
Goodman L.A. [1981]: "Association Models and Canonical Correlation in the Analysis of Cross-Classifications Having Ordered Categories", JASA 76, 320-334
Grizzle J.E., Starmer C.F., Koch G.G. [1969]: " Analysis of Categorical Data by Linear Models", Biometrics 25, 489-504
Haberman S.J. [1974]: "Loglinear Models for Frequency Tables with Ordered Classifications", Biometrics 30, 589-600
Kendall M.G., Stuart A. [1973]: Statističeskije vyvody i svjazy, Moskva, Nauka
Lebart L., Morineau G., Warwick K.M. [1984]: Multivariate Descriptive Statistical Analysis, New York, John Wiley & Sons
Mc Cullagh P. [1977]: "A Class of Parametric Models for the Analysis of Square Contingency Tables with Ordered Categories", Biometrika 65, 413-418
Nelder J.A., Wedderburn W.M. [1972]: "Generalized Linear Models", JRSSA 135, 370-384
Nishisato S. [1980]: Analysis of Categorical Data: Dual Scaling and Its Applications, Toronto, University of Toronto Press
Placket R.L. [1965]: "A Class of Bivariate Distributions", JASA 60, 516-522
Pratt J.W. [1959]: "Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures, JASA 54, 655-667
Řehák J. [1976]: "Základní deskriptivní míry pro rozložení ordinálních dat", Sociologický časopis XII, 416-431
Řehák J., Řeháková B. [1984a]: "Parciální asociační koeficienty v kontingenčních tabulkách", In: Antoch J., Jurečková J. (red.), Robust 84, s.105-108.
Řehák J., Řeháková B. [1984b]: "Porovnání distribucí u párově závislých dat ve čtvercových kontingenčních tabulkách". Sociologický časopis XX, 516-541.
Řehák J., Řeháková B. [1985]: Analýza kategorizovaných dat v sociologii, Praha. Academia
Řehák J., Řeháková B. [1986]: " Vícenásobná a parciální asociace v kontingenčních tabulkách", Sociologický časopis (v tisku)
Řeháková B.: [1985]: Model a metoda pro analýzu kategorizovaných ... kandidátská disertační práce, MFF UK Praha
Semenya K., Koch G.G. [1980]: Compound Functions and Linear Model Methods for the Multivariate Analysis of Ordinal Categorical Data, Mimeo Series No 1323, Chapel Hill, Univer.of North Carolina Institute of Statistics
Wahrendorf J. [1980]: "Inference in Contingency Tables with Ordered Categories Using Plackett's Coefficient of Association for Bivariate Distribution", Biometrika 67, 15-21
Williams O.D., Grizzle J.E. [1972]: "Analysis of Contingency Tables Having Ordered Response Categories", JASA 67, 55-63.