ANALÝZA KATEGORIÁLNÍCH DAT – PROBLÉM VÍCENÁSOBNÉ VOLBY V ODPOVĚDI

Julie Rendlová

Katedra matematické analýzy a aplikací matematiky,
Přírodovědecká fakulta, Univerzita Palackého v Olomouci

Robust, Jeseníky, 15. 9. 2016

Motivace

- Způsob hodnocení kategoriálních veličin – jeden objekt ve více kategoriích (data z ČSFD)
- Dotazníky – možnost více odpovědí zároveň
- Převládající nevhodné přístupy:
  - Logistická regrese – zanedbání korelací
  - Analýza sumarizační tabulky – pouze jedna z marginálií

Kategoriální proměnné s vícenásobnou volbou v odpovědi

- MRCV – závislost mezi items => každé pozorování je set korelovaných binárních odpovědí
- Item-response tabulka, sub-tabulky

Items otázky Y		Y_1		…	Y_J	
Items otázky W		1	0		1	0
W_1	1	$n_{11(11)}$	$n_{10(11)}$	…	$n_{11(1J)}$	$n_{10(1J)}$
	0	$n_{01(11)}$	$n_{00(11)}$		$n_{01(1J)}$	$n_{00(1J)}$
…		…		…	…	
W_I	1	$n_{11(I1)}$	$n_{10(I1)}$	…	$n_{11(IJ)}$	$n_{10(IJ)}$
	0	$n_{01(I1)}$	$n_{00(I1)}$		$n_{01(IJ)}$	$n_{00(IJ)}$

Log-lineární modely pro MRCV:
model za platnosti SPMI

- Upravená Pearsonova statistika pro χ^2- test o SPMI:

$$\chi^2_S = \sum_{i=1}^{I}\sum_{j=1}^{J}\chi^2_{S,i,j}$$

- Rozdělení? => Rao-Scottovy korekce, bootsrap
- Log-lineární model: $\log\lambda_i = \sum_{j=1}^{p} X_{ij}\beta_j$, i = 1, …, n

  pro y = 0, 1, 2, …: $f(y,\lambda) = \frac{1}{y!}e^{-\lambda}e^{\log\lambda^y}, \quad \lambda > 0$

- Log-lineární model za platnosti SPMI:

$$\log(\lambda_{ab(ij)}) = \gamma_{ij} + \eta^W_{a(ij)} + \eta^Y_{b(ij)}$$, i = 1, …, I, j = 1, …, J, a, b = 0, 1

Log-lineární modely pro MRCV:
význam parametrů v modelu za platnosti SPMI

- Sub-tabulka k item response tabulce s odhadnutými četnostmi

W \ Y	0	1	Σ
0	$e^{\hat{\gamma}_{ij}}$	$e^{\hat{\gamma}_{ij}} e^{\hat{\eta}^{Y}_{1(ij)}}$	$m_{0\cdot}$
1	$e^{\hat{\gamma}_{ij}} e^{\hat{\eta}^{W}_{1(ij)}}$	$e^{\hat{\gamma}_{ij}} e^{\hat{\eta}^{W}_{1(ij)}} e^{\hat{\eta}^{Y}_{1(ij)}}$	$m_{1\cdot}$
Σ	$m_{\cdot 0}$	$m_{\cdot 1}$	$m_{\cdot\cdot}$

1.) změna $\eta^{Y}_{1(ij)}$

2.) změna $\eta^{W}_{1(ij)}$

3.) změna γ_{ij}

1.)	0	1	Σ
0	$m_{00(ij)}$	$\boldsymbol{m_{01(ij)}}$	$m_{0\cdot}$
1	$m_{10(ij)}$	$\boldsymbol{m_{11(ij)}}$	$m_{1\cdot}$
Σ	$m_{\cdot 0}$	$m_{\cdot 1}$	$m_{\cdot\cdot}$

2.)	0	1	Σ
0	$m_{00(ij)}$	$m_{01(ij)}$	$m_{0\cdot}$
1	$\boldsymbol{m_{10(ij)}}$	$\boldsymbol{m_{11(ij)}}$	$m_{1\cdot}$
Σ	$m_{\cdot 0}$	$m_{\cdot 1}$	$m_{\cdot\cdot}$

3.)	0	1	Σ
0	$\boldsymbol{m_{00(ij)}}$	$\boldsymbol{m_{01(ij)}}$	$m_{0\cdot}$
1	$\boldsymbol{m_{10(ij)}}$	$\boldsymbol{m_{11(ij)}}$	$m_{1\cdot}$
Σ	$m_{\cdot 0}$	$m_{\cdot 1}$	$m_{\cdot\cdot}$

Log-lineární modely pro MRCV:
poměry šancí v modelu za platnosti SPMI

- Za platnosti SPMI – poměr šancí roven 1
- Pro jednu sub-tabulku:

$$\log(\lambda_{00}) = \gamma + \eta_0^W + \eta_0^Y \qquad \log(\lambda_{01}) = \gamma + \eta_0^W + \eta_1^Y$$
$$\log(\lambda_{10}) = \gamma + \eta_1^W + \eta_0^Y \qquad \log(\lambda_{11}) = \gamma + \eta_1^W + \eta_1^Y$$

$$\log OR = \log \frac{\lambda_{00}\lambda_{11}}{\lambda_{01}\lambda_{10}} = \gamma + \eta_0^W + \eta_0^Y + \gamma + \eta_1^W + \eta_1^Y - \gamma - \eta_0^W - \eta_1^Y - \gamma - \eta_1^W - \eta_0^Y = 0$$

Log-lineární modely pro MRCV:
další asociační stupně

- $\log(\lambda_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \psi_{ab}$

  $\log(\lambda_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \psi_{ab} + \psi^{Y}_{ab(j)}$

  $\log(\lambda_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \psi_{ab} + \psi^{W}_{ab(i)}$

  $\log(\lambda_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \psi_{ab} + \psi^{W}_{ab(i)} + \psi^{Y}_{ab(j)}$

  $\log(\lambda_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \psi_{ab} + \psi^{W}_{ab(i)} + \psi^{Y}_{ab(j)} + \psi^{WY}_{ab(ij)}$

- i = 1, …, I, j = 1, …, J, a, b = 0, 1

- Další parametry v modelu:

  ψ_{ab} : homogenita – konstantní poměry šancí

  $\psi^{Y}_{ab(j)}, \psi^{W}_{ab(i)}, \psi^{W}_{ab(i)} + \psi^{Y}_{ab(j)}$: částečná homogenita – konstantní v rámci Y_j / W_i / mezi dvěma items proměnné Y při změně items proměnné W

Log-lineární modely pro MRCV:
testování podmodelů

- Upravená Pearsonova statistika pro testování podmodelů:

$$\chi^2_M = \sum_{a,b,i,j} \frac{(\hat{\lambda}^{(1)}_{ab(ij)} - \hat{\lambda}^{(0)}_{ab(ij)})^2}{\hat{\lambda}^{(0)}_{ab(ij)}}, \quad \hat{\lambda}^{(0)}_{ab(ij)}, \hat{\lambda}^{(1)}_{ab(ij)}$$ očekávané četnosti v M_0 a M_1

- Rozdělení? => Rao-Scottovy korekce, Gange bootsrap

- Standardizovaná Pearsonova rezidua v každé buňce:

$$e_{ab(ij)} = \frac{\left(n_{ab(ij)} - \frac{n_{ab(i\bullet)} n_{ab(\bullet j)}}{n} \right)}{\sqrt{\frac{n_{ab(i\bullet)} n_{ab(\bullet j)}}{n} \left(1 - \frac{n_{ab(i\bullet)}}{n} \right) \cdot \left(1 - \frac{n_{ab(\bullet j)}}{n} \right)}}$$, i = 1, …, I, j = 1, …, J

Dotazník o studijních návycích

- Jaké materiály obvykle využíváte ke studiu na zkoušky?
- Kdo podporuje Vaše studia finančně?
- Abyste si rozšířili znalosti ve Vašem oboru studia, z kterých dalších oborů si aktivně vyhledáváte informace?
- Hypotézy o SPMI – srovnání zemí a pohlaví, srovnání 1. a 2. otázky

Literatura

- Bilder, C. R., Loughin, T. M., *Modeling Association Between Two or More Categorical Variables that Allow for Multiple Category Choices*, Communications in Statistics – Theory and Methods, Vol. 36, 2007, 433–451
- Agresti, A., Liu, I.-M., *Modeling a Categorical Variable Allowing Arbitrarily Many Category Choices, Biometrics*, Vol. 55, 1999, 936–943
- Rendlová, J., *Analýza kategoriálních dat – problém vícenásobné volby v odpovědi*, diplomová práce, Univerzita Palackého v Olomouci, 2015
- Agresti, A., *Categorical Data Analysis*, second edition, John Wiley & Sons, Inc., 2002
- McCullagh, P., Nelder, J. A., *Generalized Linear Models*, second edition, London; New York: Chapman and Hall, 1989