Thomas Rusch and Daniela Weber and Reinhold Hatzinger

Goodness-of-fit-tests for highdimensional contingency tables and pattern models based on limited information

Conference or Workshop Item

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Goodness-of-fit-tests for highdimensional contingency tables and pattern models based on limited information

Thomas Rusch, Daniela Weber, Reinhold Hatzinger

Department of Statistics and Mathematics
WU

5.Klagenfurter Statistiktag
15. May 2009

Outline

- Introduction via examples
- X^2 and G^2 and their shortcomings
- Limited Information Approach
  - Distribution of data and moments
  - Limited information tests for $H_0 : \pi = \pi_0$
  - Limited information tests for $H_0 : \pi = \pi(\theta)$
- Numerical examples
  - Bradley Terry Model
  - Rasch model

Introduction

- Example 1: Paired comparison experiment with 4 objects, N=100 and no ties
  - $2^6 = 64$ patterns aranged in a contingency table
  - Sparsity is to be expected
- Example 2: Psychological test with dichotomous scoring, 6 items, N=100
  - Can be arranged in a $2^6 = 64$ contingency table
  - Sparsity is to be expected
- General problem: The contingency table's sparsity grows exponentially with the number of items or objects (e.g. 15 item or 6 objects lead to 32 768 cells).

X^2 and G^2

- Full information tests
- Assess Goodness-of-fit of models for contingency tables
- For sparse contingency tables, the empirical type I error rates do not conform to the asymptotic null distribution
- Proposed solutions
  - Pooling cells
  - Resampling methods
  - Limited information methods

Limited information approach

- Usage of low-order marginals in goodness-of-fit assessment (two-way or three-ways marginal distributions mainly)
  - Have more accurate empirial type I error rates
  - Asymptotically more powerful than X^2
  - Include X^2 as a special case
- Allow for identification of sources for misfit
- Are computationally efficient
- Easily extend to the polytomous case

Distribution of data and moments (1)

We consider a n-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Yn)'$, with each $Y_i \sim B(1, \dot{\pi}_i)$ and $y_i \in \{0, 1\}$. The joint distribution is then

$$\pi_{\mathbf{y}} = P(Y_i = y_i, \ldots, Y_n = y_n), \tag{1}$$

for the pattern

$$\mathbf{y} = (y_1, \ldots, y_n). \tag{2}$$

We write π or $(\pi(\theta)$ in case of a parametric model with parameter vector $\theta)$ for the vector of 2^n joint probabilities.

Distribution of data and moments (2)

This distribution can alternatively be characterised by a $(2^n - 1)$-dimensional vector, $\dot{\pi}$, of joint moments $E(Y_i Y_j \dots Y_n) = P(Y_i = 1, Y_j = 1, \dots, Y_n = 1)$ for every possible combination $i, ij, ijk, \dots, i \dots n$.

$$\dot{\pi}' = (\dot{\pi_1}' \mid \cdots \mid \dot{\pi_n}') = (\dot{\pi}_1, \dots, \dot{\pi}_n \mid \dot{\pi}_{ij}, j < i \mid \dots, \mid \dot{\pi}_{1\dots n}) \quad (3)$$

Distribution of data and moments (3)

There exists a bijective relationship between $\dot{\pi}$ and π via

$$\dot{\pi} = \mathbf{T}\pi. \tag{4}$$

$\mathbf{T}$ is sort of a design matrix that assigns the marginal distributions (or moments) to the patterns.
$\mathbf{T}$ can now be partitioned to $\mathbf{T} = (\mathbf{T}'_{n1}, \ldots, \mathbf{T}'_{ni}, \ldots, \mathbf{T}'_{nn})'$ with $ni = \binom{n}{i}, i = 1 \ldots n$. This means that the vector of joint moments up to order r is

$$\dot{\pi}_r = \mathbf{T}_r\pi \tag{5}$$

with $\mathbf{T}_r = (\mathbf{T}'_{n1}, \ldots, \mathbf{T}'_{nr})'$. $\mathbf{T}_r$ is of dimensions $\sum_{i=1}^{r} \binom{n}{i} \times 2^n$.

Distribution of data and moments (4)

Let $\mathbf{p}_r$ denote the vector of sample moments up to order r. with dimension $\sum_{i=1}^{r} \binom{n}{i}$, then the asymptotic distribution of the marginal residuals is normal, i.e.

$$\sqrt{N}(\mathbf{p}_r - \dot{\pi}_r) \longrightarrow N(0, \Xi_r), \quad \Xi_r = \mathbf{T}_r \Gamma \mathbf{T}_r'. \tag{6}$$

with

$$\Gamma = diag(\pi) - \pi\pi' \tag{7}$$

With these results we can propose test statistics for the goodness-of-fit based on marginal residuals up to order r.

Limited information test statistics for $H_0 : \pi = \pi_0$

Maydeu-Olivares & Joe (2005) proposed the class of limited information test statistics,

$$L_r = N(\mathbf{p}_r - \dot{\pi}_r)' \Xi_r^{-1} (\mathbf{p}_r - \dot{\pi}_r), \quad r = 1 \ldots n. \tag{8}$$

$$L_r \sim \chi^2 \text{ with } df = \sum_{i=1}^{r} \binom{n}{i} \tag{9}$$

- X^2 is of form (8) with $r = n$
- The choice of r depends on the size n relative to N. The second or third order statistics are commonly used.

Limited information test statistics for $H_0 : \pi = \pi(\theta)$ (1)

Let $\hat{\theta}$ denote the MLE. Maydeu-Olivares & Joe (2005) propose the class of test statistics up to order r

$$M_r = N(\mathbf{p}_r - \dot{\pi}_r(\hat{\theta}))' C_r(\hat{\theta})(\mathbf{p}_r - \dot{\pi}_r(\hat{\theta})), \quad r = 1 \dots n, \tag{10}$$

with

$$C_r(\hat{\theta}) \approx C_r(\theta) = \delta_r^{(c)} (\delta_r^{(c)'} \Xi_r \delta_r^{(c)})^{-1} \delta_r^{(c)'}, \tag{11}$$

and

$$\delta_r = \frac{\partial \dot{\pi}_r(\theta)}{\partial \theta'} = \mathbf{T}_r \frac{\partial \pi(\theta)}{\partial \theta'}. \tag{12}$$

$\delta_r^{(c)}$ is the orthogonal complement of δ_r, i.e. $\delta_r^{(c)'} \delta_r = 0$.

Limited information test statistics for $H_0 : \pi = \pi(\theta)$ (2)

It holds that

$$M_r \longrightarrow \chi^2 \text{ with } df = -q + \sum_{i=1}^{r} \binom{n}{i} \tag{13}$$

Please note that the use of C_r is an auxiliary construct that allows to find the inverse of

$$\Sigma_r = \Xi_r - \delta_r I^{-1} \delta_r' \tag{14}$$

which is the asymptotic variance-covariance matrix of the MLE with I denoting the Fisher information matrix. One can show that C_r has Σ_r as a general inverse.

Example

- 5 items (from simulated Rasch model)
- 50 persons
- Leads to a $2^5 = 32$ contingency table
- Loglinear model fitted

Test results

- $M_2 = 6.716$ with $df = 9 (p = 0.67)$
- $M_3 = 10.162$ with $df = 19 (p = 0.95)$
- $M_5 = 12.338$ with $df = 25 (p = 0.98)$

```
> summary(mod1)

Call:
glm(formula = patts ~ des[, 1] + des[, 2] + des[, 3] + des[,
    4] + des[, 5], family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6172  -0.9719  -0.4220   0.3940   1.9812

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5940     0.2956   5.392 6.98e-08 ***
des[, 1]      0.4895     0.2914   1.680    0.093 .
des[, 2]     -0.4895     0.2914  -1.680    0.093 .
des[, 3]     -0.1603     0.2838  -0.565    0.572
des[, 4]     -2.1972     0.4714  -4.661 3.15e-06 ***
des[, 5]     -1.8153     0.4076  -4.454 8.43e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.0
```

Bibliography

- Maydeu-Olivares, A. & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *JASA, 471*, 1009-1020.
- Cai, L., Maydeu-Olivares, A., Coffman, D.L. & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *The British Journal of Mathematical and Statistical Psychology, 59*, 172-194.