检验导出的等值新方法及其表现探讨*

熊建华　丁树良　雷宁宁

(江西师范大学 计算机信息工程学院,南昌 330027)

摘要:该文受 Berkson 将检验方法用于估计未知参数的启发,根据三个拟合优度统计量导出三种新的求取等值系数的方法,即:平方根等值方法(Square Root criterion, SQRTcrit)、对称相对熵等值方法(Symmetric Relative Entropy criterion,SREcrit)、加权等值方法(Weighted criterion,Wcrit),即 Haebara 准则的加权式。虽然在被检验的两个分布列很接近时,这三个多项拟合优度检验方法是渐近等价的,然而用它们求取等值系数时,Monte – Carlo 模拟结果表明这三种新等值方法的行为表现存在差异。它们之间的差异和随机误差的大小有密切关系,即与项目参数估计的精度有关,还与等值系数 A 的范围有关。

关键词:平方根等值方法;对称相对熵等值方法;加权等值方法;Monte – Carlo 模拟

中图分类号:B841.2　　文献标识码:A　　文章编号:1003 – 5184(2007)01 – 0070 – 04

1　引言

众所周知,寻找等值系数的方法是一种估计方法[1,2],而估计与假设检验有紧密的关系[3],比如参数检验方法的构造直观上可以借用参数估计方法。在项目反应理论(IRT)中,Berkson 将检验方法用于估计未知参数,特别是对于两参数 Logistic 模型(2PLM),当能力已知时,该方法估计项目参数效果不错(可参见[4])。在 IRT 框架内讨论等值,特别是等值系数的求取过程中,我们受到 Berkson 的启发,也将三个拟合优度统计量[5]作为三种新的求取等值系数的方法。经过数学论证,在被检验的两个分布列很接近时,这三个多项拟合优度检验方法是渐近等价的[5],然而用它们求取等值系数时,这三种新等值方法的行为表现如何?本文通过 Monte Carlo 模拟进行研究。

2　IRT 测验等值的基本概念

从数学上讲,所谓等值,是要通过不同的测验 X 和 Y 得到的项目参数和能力参数的估计值,应用某个具体的等值方法,求出等值系数 A 和 B。注意等值应遵从公平性原则,即同一被试的能力估计不应受其参加的特定的考试形式的影响。因此,等值方法或等值准则中,测验形式 X 与 Y 是处于对称地位,这一点与通常统计中应用回归分析时不要求自变量与因变量地位对称是有根本不同的。IRT 认为,如果有两个不同测验形式 X 和 Y,项目 j 是含在这两个测验中的锚题,锚题参数在 X 和 Y 中的值可以不同,但它们之间存在一定的关系,比如对两参数 Logistic 模型(2PLM),存在等值系数 A 和 B(A0),使得

$$a_{yj} = a_{xj}/A, a_{yj} = Ab_{xj} + B, j = 1,2,\cdots m$$

若同一被试 α 既参加了测验 X 又参加了测验 Y,则其能力之间存在关系式

$$\theta_{y\alpha} = A\theta_{x\alpha} + B, \alpha = 1,2,\cdots N$$

除非另有申明,文中总假设 $\alpha = 1,2,\cdots N, j = 1,2,\cdots m$。

3　三种新等值方法的提出

文章主要讨论如何根据统计检验方法导出求取等值系数的新方法及新方法的一些良好表现,故为了说明问题而不致太过繁琐,讨论最简单的 2PLM,即

$$P_{x\alpha j} = \{1 + \exp[-a_{xj}(\theta_{x\alpha} - b_{xj})]\}^{-1},$$
$$Q_{x\alpha j} = 1 - P_{x\alpha j} \quad (1)$$
$$P_{y\alpha j} = \{1 + \exp[-a_{yj}(A\theta_{x\alpha} + B - b_{yj})]\}^{-1},$$
$$Q_{y\alpha j} = 1 - P_{y\alpha j} \quad (2)$$

3.1　平方根等值方法(Square root criterion, SQRTcrit)

受如下所述的 Free – man – Tukey 关于多项分布拟合检验[5]

$$K_{FT}^2(x,y) = 4\sum_i(\sqrt{x_i} - \sqrt{y_i})^2 \quad (3)$$

的启发,注意到(3)中$(x_1, x_2, \cdots x_n)$是多项分布列,而$(x, 1-x)$是一个二项分布列,将(3)式展开,则有:

* 基金项目:国家自然科学基金(60263005),全国教育考试十一五科研规划课题(2006JKS6063),江西省高校人文社会科学研究项目(JY06201),教电馆研(063120144)。

$$K_{FT}^2(x,y)=4[(\sqrt{x}-\sqrt{y})^2+(\sqrt{1-x}-\sqrt{1-y})^2] \tag{4}$$

因为常数系数对求导结果不产生影响，所以将(4)式中的系数 4 去掉，同时 x，y 用对应的 P_{xaj}，P_{yaj} 代替，并结合(1)、(2)式，便可得到一个考生作答一个项目情况，若考察 N 个考生和 m 个项目(以下都考虑 N 个考生和 m 个项目)，则可得到如下式子，并取名为平方根准则，简记为 *SQRTcrit*

$$SQRTcrit=\sum_{\alpha=1}^{N}\sum_{j=1}^{m}[(\sqrt{P_{xaj}}-\sqrt{P_{yaj}})^2+(\sqrt{Q_{xaj}}-\sqrt{Q_{yaj}})^2] \tag{5}$$

这个式子已满足了等值的对称性要求[6]。

3.2　对称相对熵等值方法(Symmetric Relative Entropy criterion，SREcrit)

同样地，由似然比检验

$$G^2(x,y)=2\sum_i x_i\log(x_i/y_i) \tag{6}$$

略去常数项，同时 x，y 用 P_{yaj}，Q_{yaj}，P_{xaj}，Q_{xaj} 代替，则可得到如下式子：

$$I(P_x\parallel P_y)=\sum_{\alpha=1}^{N}\sum_{j=1}^{m}\Big(P_{xaj}\log(P_{xaj}/P_{yaj})+Q_{xaj}\log(Q_{xaj}/Q_{yaj})\Big)$$

称 $I(P_x\parallel P_y)$ 为 Kullback－Leibler 信息，又称为 P_x 对 P_y 的相对熵或 P_x 相对于 P_y 的判别信息量，根据等值的对称性，用作求取等值系数的相应式子如下：

$$SREcrit=I(P_x\parallel P_y)+I(P_y\parallel P_x) \tag{7}$$

称(7)为对称相对熵准则式。

3.3　加权等值方法(Weighted criterion，Wcrit)

受 *Pearson* 的卡方检验

$$\chi^2(x,y)=\sum_i(x_i-y_i)^2/y_i \tag{8}$$

的启发，再根据等值的对称性，得相应的求取等值系数的加权准则(*Weighted criteria*)，并通过几步代数运算后有：

$$Wcrit=\sum_{\alpha=1}^{N}\sum_{j=1}^{m}[(P_{xaj}-P_{yaj})^2(1/P_{xai}+1/P_{yai}+1/Q_{xaj}+1/Q_{yaj})] \tag{9}$$

实际上(9)式是 *Haebara* 准则的加权式[7]。

4　新等值方法的比较——Monte－Carlo 模拟研究

研究设计了一套模拟方法，模拟不同测验中同一套锚题的不同估计值，模型中被试能力 $\theta_1,\theta_2,\cdots,\theta_N$ 为一批固定的数，且服从标准正态分布 $N(0,1)$。给定一组项目参数(a_{xj}，b_{xj})以及等值系数 A($A>0$)，B，再令 $a'_{yj}=a_{xj}/A$，$b'_{yj}=Ab_{xj}+B$，在 a'_{yj}，b'_{yj} 上分别添加一些误差，在 a'_{yj} 上添加独立同分布的 $N(0,\sigma_1^2)$ 的随机误差 $\varepsilon_{j_1}^{(k)}$，并且保证 $a'_{yj}+\varepsilon_{j1}^{(k)}$ 大于 0 且小于 2.5 在 b'_{yj} 上添加独立同分布的 $N(0,\sigma_2^2)$ 的随机误差 $\varepsilon_{j2}^{(k)}$。根据 MULTILOG 使用说明书[8]可知，区分度的估计标准误差(SE)大部分记在 0.05～0.15 之间，而难度的估计标准误差大部分在 0.1～0.25 之间，故取 σ_1 在 1/20～1/6 之间，σ_2 在 1/10～1/4 之间。令

$$a_{yj}^{(k)}=a'_{yj}+\varepsilon_{j1}^{(k)},\ b_{yj}^{(k)}=b'_{yj}+\varepsilon_{j2}^{(k)}\qquad k=1,2,\cdots K \tag{10}$$

同时还对 $a_{yj}b_{yj}$ 和 $a_{yj}^{(k)}$，$b_{yj}^{(k)}$ 进行检查，以保证它们符合教育与心理测量学的意义(比如区分度大于 0 的项目才有使用价值)，对两组项目参数(a_{xj}，b_{xj})和($a_{yj}^{(k)}$，$b_{yj}^{(k)}$)分别用 SQRT 方法、SRE 方法、W 方法计算等值系数 A，B，这时 A，B 与 $\varepsilon_{j_1}^{(k)}$，$\varepsilon_{j_2}^{(k)}$ 有关，记为 $A^{(k)}$，$B^{(k)}$。固定一种等值方法，可以得到一批 $A^{(k)}$，$B^{(k)}$(研究中取 $K=100$)，计算

$$\mathrm{RMSD}=\sqrt{\sum_{k=1}^{k}\big((A^{(k)}-A)^2+(B^{(k)}-B)^2\big)/k} \tag{11}$$

RMSD 便是 Monte Carlo 模拟中常用的偏离均方根(root square mean deviation)准则[6,7]，它表示 A，B 的估计值 $A^{(K)}$，$B^{(K)}$ 对真值 A，B 的恢复(recovery)程度，显然它的值越小越好。改变等值方法，可以得到相应的 RMSD。将不同方法求出的等值系数的 RMSD 进行统计分析[9](例如采用 Wilcoxon 符号秩检验[10])，看看不同等值方法在统计上是否有显著差异。

首先，讨论 $\sigma_1=1/10$，$\sigma_2=1/6$ 时，对给定真值(A，B)的 144 个组合共 14400 次模拟试验。用表格形式来描绘各种等值方法在模拟试验中的表现情况，以便从中找出一些规律，表格的第一行对应真值 B 的取值，第一列对应真值 A 的取值，A，B 取值以步长 0.1 改变，每个(A，B)的不同组合值都做了 100 次模拟试验。在每一组(A，B)取值下，用 SQRT，SRE，W 三种等值方法进行等值，求解出等值系数 $A^{(K)}$，$B^{(K)}$，并将它与真值作比较，计算 RMSD，再采用 Wilcoxon 符号秩检验，比较各种方法的差异。将表现最优的等值方法填入表 1 对应的单元格中，表现最优是指 RMSD 最小并且在统计上有显著差异，若各方法在统计上没有显著差异，对应单元格为空，得到如下的表格。对于随机误差取值为 $\sigma_1=1/20$，$\sigma_2=1/10$ 和 $\sigma_1=1/10$，$\sigma_2=1/6$ 时的结果，分别放在附录的表 3，表 4 中。

表 1 随机误差取值中等($\sigma_1 = 1/10, \sigma_2 = 1/6$)的结果

A \ B	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4
0.5	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT
0.6	SQRT	SQRT		SQRT	SQRT		SQRT	SQRT	SQRT
0.7	SQRT	W	W	SQRT	SRE	SQRT	W	W	W
0.8	W	W	W	W	W	SQRT	W	SQRT	SQRT
0.9	W	SQRT	W	W	W	W	SQRT	W	W
1.0	SQRT		SQRT	SQRT	W	W	SRE	SQRT	SQRT
1.1	SQRT	SQRT	SQRT	W		W	SQRT	SQRT	SQRT
1.2	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	W	SQRT
1.3	SQRT					SQRT	SQRT		SQRT
1.4	SQRT	SRE		SQRT	SQRT	SQRT	SQRT		SQRT
1.5	SQRT			SQRT	SQRT				SQRT
1.6						SQRT		SQRT	SQRT
1.7	SQRT	SQRT	SQRT	SQRT				SQRT	
1.8						SQRT			SQRT
1.9									
2.0		SQRT							

为了能够给出一个清晰的轮廓,将研究结果用表 2 统计出来。

前已述及,在数据与分布相拟合时,三种等值方法对应的检验统计量的渐近分布是相同的。而从表 2 的统计中可以得知,对于 Monte Carlo 模拟数据(它们与理论分布是拟合的)这三种等值方法是有差异的,并且它们之间的差异根据随机误差取值的大小而发生变化。我们认为虽然估计与假设检验有紧密的关系,但估计和检验毕竟有差异的,检验是定性地回答问题(拒绝还是接受),而估计通常是定量地回答问题。如果样本 x 点落在拒绝域之中(为简单起见,设 $T(x) \leq C$),同时如果有另一个样本 x',且检验统计量的值 $T(x') \leq T(x)$,则 $T(x') \leq C$,这时不论 $T(x')$与 $T(x)$相差多大,同样是拒绝原假设,即 x'与 x 都落在拒绝域之类。但是作为估计量来讲,$T(x')$与 $T(x)$的大小却可能是关注的焦点。所以从这一角度来看,估计和检验尽管有很多联系,但它们毕竟有差异。

表 2 各种等值方法表现最优的个数统计表

随机误差 \ 等值方法	*SQRT* 方法表现最优的个数	*SRE* 方法表现最优的个数	*W* 方法表现最优的个数	无显著差异个数	合计
$\sigma_1 = 1/20, \sigma_2 = 1/10$	86	0	20	38	144
$\sigma_1 = 1/15, \sigma_2 = 1/8$	83	1	14	46	144
$\sigma_1 = 1/10, \sigma_2 = 1/6$	68	3	23	50	144
$\sigma_1 = 1/6, \sigma_2 = 1/4$	42	12	27	63	144
$\sigma_1 = 1/5, \sigma_2 = 1/2$	18	30	77	19	144

另外,从表 1,附录中的表 3 和表 4 的详细描述中,可以发现这三种等值方法的差异不是杂乱无章的,能够找到它们之间的规律,根据这种规律来合理的选用等值方法。并且我们也作了大量的 Monte - Carlo 模拟,发现它们在一定条件下比目前流行的等值方法有更优的表现。

参考文献

1 漆书青,戴海崎,丁树良.现代教育与心理测量学原理.南昌:江西教育出版社,1998.229 - 236.

2 Kolen M J, Brennan R L. Test Equating: Methods and Practices. New York: Springer - Verlag, New York. Inc, 1995.169 - 173.

3 Mood A M, Graybill F A. 统计学导论.史定华译.北京:科学出版社,1978.312.

4 Baker F B. Item Response Theory: Parameter Estimation Techniques. Marcel Dekker, Inc, 1992.57 - 62.

5 Bishop Y M M, Fienberg S E, Holland P W. 离散多元分析理论与实践.张尧庭译.北京:中国统计出版社,1998.619

– 620,626 – 630.

6 丁树良,熊建华,罗芬,等.一种新的等值准则及其适用范围的探讨.心理学报,2005,37(5):674 – 680.

7 熊建华,丁树良.Haebara 等值方法及其加权准则.江西师范大学学报,2005,29(5):434 – 437.

8 Thissen D. MULTILOG User's Guide. Scientific Software, Inc.,1991. Examples3 – 7—3 – 9.

9 Harwell M R. Analyzing the results of Monte Carlo studies in item response theory. Educational and Psychological Measurement,1997,57(2):266 – 279.

10 吴喜之,王兆平.非参数统计方法.北京:高等教育出版社,1996.35 – 41.

附录

表 3　随机误差取值($\sigma_1 = 1/20, \sigma_2 = 1/10$)的结果

A \ B	– 0.4	– 0.3	– 0.2	– 0.1	0	0.1	0.2	0.3	0.4
0.5	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT
0.6	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT
0.7	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT
0.8	SQRT	W	SQRT	SQRT	W	W	SQRT	W	W
0.9	SQRT	W	W	W	SQRT	SQRT	SQRT	W	W
1.0	W	W	SQRT	W	W	SQRT	SQRT	W	W
1.1	SQRT	SQRT	SQRT	W	SQRT	SQRT	SQRT	W	SQRT
1.2	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT	SQRT
1.3		SQRT	SQRT	SQRT	SQRT	SQRT		SQRT	SQRT
1.4	W	SQRT		SQRT	SQRT	W	SQRT		SQRT
1.5	SQRT	SQRT		SQRT	SQRT	SQRT		SQRT	SQRT
1.6	SQRT	SQRT		SQRT	SQRT	SQRT			
1.7				SQRT		SQRT			
1.8			SQRT						
1.9				SQRT					
2.0						SQRT		SQRT	SQRT

表 4　随机误差取值($\sigma_1 = 1/5, \sigma_2 = 1/2$)的结果

A \ B	– 0.4	– 0.3	– 0.2	– 0.1	0	0.1	0.2	0.3	0.4
0.5	W	W	W	W	SRE	W	W	W	W
0.6	W	W	W	W	W	W	W	W	W
0.7	W	W	W	W	W	W	W	W	W
0.8	W	W	W	W	W	W	W	W	W
0.9	W	W	W	W	W	W	W	W	W
1.0	SRE	W	W	W	W	W	W	W	W
1.1	W	W	W	W	W	W	W	W	W
1.2	W	W	W	SRE	W	W	W	SRE	SRE
1.3	SRE	W	SRE	SRE	W	SRE	W	SRE	SRE
1.4	W	W	SRE	SRE	SRE	SRE	W	W	SRE
1.5	SRE	SRE	SRE	SRE	W	SRE	W	SRE	SRE
1.6	SQRT	SQRT	SQRT	W	SQRT	SQRT	SQRT	SRE	SQRT
1.7	SQRT	SQRT	SQRT	SRE	SRE		SQRT		
1.8		SQRT	SRE	SQRT	SRE	SQRT		SRE	
1.9		SQRT			SQRT	SQRT			
2.0							SQRT		

The New Equating Methods Derived from Test Statistic and Their Performances

Xiong Jianhua Ding Shuliang Lei Ningning
(Computer Information Engineering College of Jiangxi Normal University, NanChang 330027)

Abstract: This paper is inspired by applying Test Statistic to estimate unknown parameter, three new solving the equating coefficients methods (for short, equating method) are proposed in the light of goodness-of-fit test statistic, that are Square Root method, Symmetric Relative method, Weighted method which is the Weighted Haebara method. When two distributions are approximate, the three goodness-of-fit test statistic are near equivalent. But what is the result about the three equating methods derived from test statistic? Monte – Carlo study shows that there are differences among three equating methods. The difference has closely relationship with estimation random error and the domain of equating coefficient A.

Key words: square root method; weighted method; symmetric relative entropy method; Monte – Carlo simulation

---

(上接第 60 页)

8 Kavanaugh R D, Goodrich T, Harris P L. Counterfactual reasoning in two – year – olds. Paper Presented at the VIIth European Conference on Developmental Psychology. Kraków, Poland, 1995.

9 Hadwin J, Bruins J. Imagining alternative outcomes: Counterfactual reasoning in children with autism. Unpublished manuscript, University of Essex, 1997.

10 Robinson E J, Beck. What is difficult about counterfactual reasoning. In: P. Mitchell, K. J. Riggs, Eds. Children's reasoning and the mind. Hove, UK: Psychology Press, 2000. 101 – 119.

11 卿素兰,方富熹.儿童反事实思维研究综述.湖北大学学报(哲学社会科学版),2004,31(4):470 – 473.

12 Bloom A H. The linguistic shaping of thought: A study in the impact of language on thinking in China and the West. Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

13 Turley – Ames K J, Whitfield M M. Working memory and controlled processing of counterfactuals. Paper presented at the 1st Annual Meeting of the Society for Personality and Social Psychology, Nashville, TN, 2000.

14 Peterson D, Riggs K J. Adaptive modelling and mindreading. Mind & Language, 1999, 14: 80 – 112.

The Development of 3 – to 5 – year – old Children's Counterfactual Thinking

Zhang Kun
(Sociology Department, East China University of Politics and Law, Shanghai 201620)

Abstract: Data were collected to a sample of 58 3 – to 5 – year – old children to investigate the development of their counterfactual thinking using consequent counterfactual tasks and antecedent counterfactual tasks. The results indicated: (1) The scores of 3 – year – old children's consequent counterfactual reasoning were lower than that of 4 and 5 – year – old children. But there was no significant difference between 4 and 5 – year – old children's consequent counterfactual reasoning. (2) Significant difference existed in antecedent counterfactual reasoning in terms of direction and structure. (3) Young children who were able to generate counterfactual statement can generate both upward and downward counterfactuals equally well. Results also indicated that young children, similar to adults, generate fewer subtractive than additive counterfactuals.

Key words: 3 and 5 – year – old children; counterfactual thinking; development