心 理 学 报 2003,35(6):835~841
Acta Psychologica Sinica

项目反应理论框架下的新等值方法——对数对比等值法*

丁树良 熊建华 毛萌萌

(江西师范大学计算机科学技术学院,南昌 330027)

摘 要 项目反应理论有一些以除法形式给出的多级评分模型,若采用 Haebara 等值法、Stocking_Lord 等值法或对称相对熵等值法进行测验等值,都因其对初值有较高要求而可能导致失败。针对这一类模型,我们给出了一种新的等值方法——对数对比等值法。这种方法收敛快,对迭代初值要求低,所得结果精度较高,可以为其他等值方法提供良好的初值。研究表明,对数对比等值法还改进和推广了 0-1 评分的两参数 Logistic 模型的 Logit 变换等值法。

关键词 多级评分模型,测验等值,对数对比等值法,初值

分类号 B841.2

1 问题的提出

研究表明,用多级评分项目(polytomous item)进行测试比用两级评分项目(dichotomous item,又称0,1记分题)所得信息更丰富,测量结果更准确可靠[1]。项目反应理论(IRT)发展了不少多级评分模型[2,3],按照 Thissen 和 Steinberg(1986)分类原则[4],减法形式模型(difference models)和除法形式模型(divide-by-total models)是其中两种重要的形式。比如著名的 Samejima 定义的等级反应模型(graded response model)及由此演化出来的 Muraki 评定量表模型(rating scale model)均属于减法形式模型;而称名反应模型(nominal response model)、分部评分模型(partial credit model, PCM)、广义分部评分模型(GPCM)、评定量表模型等均为除法形式模型[4]。另外,二项试验模型(binomial trial model)、泊松计数模型(Poisson counts model)也是除法形式的多级评分模型[2]。如果测验项目满足除法形式模型,我们发现等值工作比较困难,这是因为目前通用的几种等值方法,如 Haebara 等值法(H 方法)、Stocking_Lord 等值法(SL 方法)[3,5]都对初值要求较严,由均数-标准差法提供的初值常常使相应的迭代过程不收敛,我们给出的对称相对熵方法(SRE 方法)[6]也有相同的弱点;纵使在一些特殊情况下,迭代可以收敛,但运算时间也较长。

本文主要针对上述除法形式定义的多级评分模型设计一种新的等值方法——对数对比等值方法(Logcontrast mothod for equating test),对数对比这个名词引自文献[7],通常又称之为基准类比(baseline category contrast)[8]或 Logistic 回归[8,9]。本文第二部分中定义了对数对比等值准则及求等值系数的方法,针对几个特殊的除法形式的多级评分模型给出具体的对数对比准则形式,发现针对双参数 Logistic 模型的 Logit 变换等值法[10]是对数对比等值法的特例;第三部分介绍对数对比等值法的蒙特卡洛模拟的行为表现,得出这种等值方法有三个优点:对初值不敏感,收敛快,所求结果精度较高;第四部分是对数对比等值法的可能应用及讨论;最后是一个附录,给出了这种新的等值方法的计算公式和其特点的理论依据。

2 对数对比等值法

2.1 对数对比等值法原理与特点

对数对比,即通常所说的 Logistic 回归,常用来处理类目反应数据,特别适用于诸如得分为有序类目的反应数据,是社会研究中常用的统计方法。得分为有序类目的反应数据,或称分类反应数据,这种顺序对相同能力的被试的得分概率有影响,处理这

收稿日期:2002-10-28

* 本研究受国家自然科学基金项目(60263005)资助,并部分受教育科学规划重点课题(DBB010501)资助。

通讯作者:丁树良,E-mail: dingshuliang@jxnucie.com

种顺序的 Logistic 回归有很多方法,如基准类比方法,累积类比方法,邻接类比方法等[8,9,11,12]。如果自变量是连续型变量(如人的能力,或项目的难度),而因变量取值为 0 - 1 的数据(比如成功记为 1,失败记为 0;同意记为 1,反对记为 0 等等),亦即因变量 y 为二分变量时,设 $P(y=1|x)=p, P(y=0|x)=1-p=q$,则可定义成功与失败的概率之比为优比(odds ration) p/q,对优比取对数,得 $\ln(p/q)$,如果 p 的形式为 Logistic 曲线,即 $p=\frac{e^{-t(x)}}{1+e^{-t(x)}}$,则 $q=1-p=\frac{1}{1+e^{-t(x)}}$,于是 $\ln(p/q)=-t(x)$,$t(x)$ 是关于 x 的函数,如果 $t(x)=ax+b$,这里 p 是 x 的非线性函数,但 $\ln(p/(1-p))$ 成了 x 的线性函数。于是从 $\ln(p/q)$ 出发比从 p 出发统计处理就容易得多。如果 p 的形式更复杂一些,比如 x 是一个向量,$t(x)=\sum_{i=1}^{k}a_i x_i+a_0$,则我们仍可采用这种 Logistic 变换,化非线性问题为线性问题。

从上世纪 50 年代起,Berkson 就应用 logit 方法处理两参数 Logistic 模型中参数估计问题[13],90 年代,国外用 Logistic 回归探查项目功能差异(DIF)[14],但在 IRT 框架下未见用这种方法进行多级评分模式的测验等值研究的报导。在国内,对双参数 Logistic 模型,朱奎花等研究了 Logit 变换等值法[10],这种等值法的优点是计算简单,不用迭代。然而它只处理了双参数 Logistic 模型,对于多级记分题的等值法没有提及。我们这里给出对数对比等值方法,它适用于项目反应理论中以除法形式定义的多级评分模型的等值问题。我们的模拟研究表明,对多级评分项目,也可以给出不经过迭代的对数对比法,但通常迭代的对数对比法求出的等值系数更接近真值。与 Stocking - Lord 的测验特征曲线等值法和 Haebara 项目特征曲线等值法相比,我们给出的方法不仅仅显示出计算速度上的优势,而且通过大量蒙特卡洛模拟研究,如果采用通常的平均数 - 标准差法提供的迭代初值,Stocking - Lord 等值法、Haebara 等值法均可能不收敛,即计算不出等值系数,但对数对比等值法对迭代初值要求不严,且由于其收敛快,还可以为其它的等值法提供迭代初值。

2.2 项目反应理论框架下测验等值的基本假定

设有两个测验形式 X 和 Y,X 和 Y 含有 m 个锚题(ancher item),且这些锚题都是多级评分题,根据项目反应理论,存在两个常数 A 和 $B(A\neq 0)$,使得这 m 个在不同测验形式上的锚题的项目参数满足如下关系式:

$$a_{yj}=a_{xj}/A,\ b_{yjt}=Ab_{xjt}+B,\ j=1,2,\cdots,m\ t=0,1,2,\cdots,f_j \quad (1)$$

若同一个能力为 θ_α 的被试,在这两个测验上表现出的能力应满足关系式:

$$\theta_{ya}=A\theta_{x\alpha}+B,\ \alpha=1,2,\cdots,N \quad (2)$$

上述式子中的 A,B 称为等值系数。IRT 中等值的任务就是要通过等值设计,收集到相应数据,根据某个等值准则式(criterion),求出等值系数 A,B。

除非另有申明,本文中总设 $\alpha=1,2,\cdots,N, j=1,2,\cdots,m,\ t=0,1,2,\cdots,f_j$。

为了描述对数对比等值法,记 $P_{yajt}=P(\theta_{ya},a_{jy},b_{yjt})$,相仿地给出 P_{xajt},由(2)有

$$P(\theta_{ya},a_{yj},b_{yjt})=P(A\theta_{xa}+B,a_{yj},b_{yjt}) \quad (3)$$

对同一被试,设其参加测验 X 和 Y,则在锚题上应有 $P_{xaj0}=P_{yaj0}$,即得

$$P(\theta_{ya},\ a_{yj},\ b_{yjt})=P(A\theta_{xa}+B,\ a_{yj},\ b_{yjt}) \quad (4)$$

注意到测验 X 和 Y 的项目参数均是估计出来的,故(4)只近似成立。若 $P_{xaj0}>0, P_{yaj0}>0$,则由(4)有

$$\ln(P_{xajt}/P_{xaj0})\approx\ln(P_{xajt}/P_{xaj0}) \quad (5)$$

(5)表示在实际测验中,同一被试 α 在测验 X 和 Y 上,对同一锚题 j 同一类目 t 上得分的概率与基础类目(零类目,相当于考试中在此题得零分)上得分概率之比的对数应该近似相等,我们用 $diff(\alpha,j,t)$ 来标记这种差异:

$$diff(\alpha,j,t)=\ln(P_{xajt}/P_{xaj0})-\ln(P_{yajt}/P_{yaj0}) \quad (6)$$

显然,$diff(\alpha,j,t)$ 的值可以为正数也可以为负数,如将其累加,则可能正负相抵,为了强调这种差异,应该考虑其绝对值的大小,但绝对值运算不便于使用相关的数学工具,于是改用它的平方,我们记:

$$cdiff(\alpha,j,t)=(diff(\alpha,j,t))^2 \quad (7.1)$$

它表示类目(category)之间的差的平方,由(5)知 $\mathrm{cdiff}(\alpha,j,t)\approx 0$,将(7.1)对被试、锚题和相应的类目相加,便得到

$$LC(1)=\sum_{a=1}^{N}\sum_{j=1}^{m}\sum_{t=1}^{f_j}\mathrm{cdiff}(\alpha,j,t) \quad (7.2)$$

我们也可以让类目之间的差异互相调整,只强调锚题项目(item)上的差异的平方(记为 idiff)

$$idiff(\alpha,j)=(\sum_{t=1}^{f_j}diff(\alpha,j,t))^2 \quad (7.3)$$

再令

$$LC(2)=\sum_{\alpha=1}^{N}\sum_{j=1}^{m}idiff(\alpha,j) \quad (7.4)$$

有时,我们甚至希望同一被试在所有锚题上的差异都可以相互调整,只强调测验(test)上的差异的平方(记为 tdiff)

$$tdiff(\alpha) = (\Sigma_{\alpha=1}^{N}\Sigma_{t=1}^{f_j} diff(\alpha,j,t)^2 \quad (7.5)$$

再令

$$LC(3) = \Sigma_{\alpha=1}^{N} tdiff(\alpha) \quad (7.6)$$

显见,(7.4)是[5]中 Haebara 的项目特征曲线等值法 Hcrit 的类似物;(7.6)是[5]中 Stocking_Lord 的测验特征曲线等值法 SLcrit 的类似物。然而,Hcrit 和 SLcrit 都是处理0-1评分项目的等值问题,而(7.2)处理了多等级评分的测验等值问题,故(7.2)在[5]中找不到相应的类似物,但是,对所有项目 j,$f_j=1$ 时,(7.2)与(7.4)相同,且都变成了[10]中所给出的等值方法,故我们也是[10]中等值方法的推广。

2.3　对数对比法的定义和意义

我们分别称(7.2),(7.4),(7.6)中定义的 LC(1),LC(2),LC(3),依次为类目对数对比准则、项目对数对比准则和测验对数对比准则。前已述及,它们适用于除法型的多级评分模型和双参数 Logistic 模型的等值。

注意到(7.2),(7.4),(7.6)中都含有未知的等值系数 A 和 B,类似于由 Hcrit 和 SLcrit 求解出等值系数 A,B,我们可以用迭代的方法,而且在多级评分模型中,迭代求解法的结果还值得推荐,然而这种迭代法对初值要求低,收敛速度快,计算结果好(结果好坏的评判见下文)。与 Hcrit 和 SLcrit 相比,对初值要求低,收敛速度快是十分突出的优点。另一点值得注意的是,除用迭代求解(7.2),(7.4),(7.6)中的等值系数外,我们也可以用类似于[10]中所给求解最小二乘估计的方法——解正规方程的方法直接求解,只不过这种求解方法在多级评分模型对应的结果有时稍显粗糙。

本文注意力放在多级评分的除法型模型的等值上,由于相应的模型的复杂性,由 LC(1),LC(2),LC(3)给出的求解公式——不论是迭代的还是非迭代的——都比较长。为了节省篇幅,又为了说明问题,我们仅对 LC(1)的迭代求解公式列在附录中,且下文的蒙特卡洛模拟也是用这些公式求解出相应结果,且将 LC(1)简记为 LC,若为强调相应的模型,也可记为 LC(model),其中 model 随所讨论的模型的变化而变化。至于非迭代计算公式,可以仿[10]进行推算,但相应的公式也比[10]更复杂。

2.4　根据(7.2)几个具体的等值准则式

例1　记 $\zeta_{ajk}=a_j(\theta_a-b_{jk})$,且规定 ζ_{aj0} 等于0,则有广义分部评分模型(GPCM)的类目反应曲线(category response curve),

$$P_{ajt} = \exp(\Sigma_{k=0}^{t}\zeta_{ajk}) / (\Sigma_{r=0}^{f_j}\exp(\Sigma_{h=0}^{r}\zeta_{ajh})) \quad (8)$$

由(8)及(7.2)可得相应于 GPCM 的对数对比等值准则式

$$LC(GPCM) = \sum_{a=1}^{N}\sum_{j=1}^{m}\sum_{t=1}^{f_j} \{\sum_{k=1}^{t}[a_{xj}(\theta_{xa}-b_{xjk}) - a_{yj}(A\theta_{xa}+B-b_{yjk})]\}^2 \quad (9)$$

若令 $a_{xj}=a_{yj}=1, j=1,2,\cdots,m$,则(9)化成分部评分模型对应的对数对比等值准则式:

$$\mathrm{LC(PCM)} = \sum_{a=1}^{N}\sum_{j=1}^{m}\sum_{t=1}^{f_j} \{\sum_{k=1}^{t}[\theta_{xa}-b_{xjk}) - (A\theta_{xa}+B-b_{yjk})]\}^2 \quad (10)$$

例2　记 $\omega_{ajk}=\theta_a-(\lambda_j+\delta_k)$,且规定 $\omega_{aj0}=0$,δ_k 为类目相交参数(category intersection parameters),且 $\sum_{k=1}^{f}\delta_k=0$。则有评定量表模型(RSM)的类目反应曲线如下:

$$P_{ajt} = \exp(\Sigma_{k=0}^{t}\omega_{ajk}) / (\Sigma_{h=0}^{f_j}\exp(\Sigma_{r=0}^{h}\omega_{ajr})) \quad t=0,1,2,\cdots,f \quad (11)$$

由(7.2)和(11),又可以导出 RSM 对应的对数对比等值准则式

$$LC(RSM) = \sum_{a=1}^{N}\sum_{j=1}^{m}\sum_{t=1}^{f_j} \{\sum_{k=1}^{t}[\theta_{xa}-\lambda_{xj}) - \delta_{xk} - (A\theta_{xa}+B-\lambda_{xj}-\delta_{xk})]\}^2 \quad (12)$$

例3　对于0,1评分的两参数模型(2PLM),记 $Q_{aj}=P_{aj0}$,$P_{aj}=1-Q_{aj}=P_{aj1}$,相仿定义 P_{xaj0},P_{xaj1},P_{yaj0},P_{yaj1},则由(7.2)有

$$LC(2PLM) = \sum_{a=1}^{N}\sum_{j=1}^{m}(\ln(P_{xaj1}/P_{xaj0}) - \ln(P_{yaj1}/P_{yaj0}))^2 = \sum_{a=1}^{N}\sum_{j=1}^{m}[a_{xj}(\theta_{xa}-b_{xj}) - a_{yj}(A\theta_{xa}+B-b_{yj})]^2 \quad (13)$$

这便是[10]中定义的 Logit 变换等值方法。

3　对数对比等值法的表现——以 GPCM 为例进行蒙特卡洛模拟研究

为了比较不同等值方法的优劣,我们以 GPCM 为例进行蒙特卡洛(MC)模拟研究,并对 MC 模拟结果进行统计分析[15,16],具体做法如下:设已知 X 测验上的 m 个锚题的参数(这 m 个锚题与 GPCM 拟合良好),设定等值系数 A,B,且对参数估计中的误差加以模拟(见以下②)以得到 Y 测验上相应锚题参数。

①给定等值系数 A,B($A\neq 0$),在模拟中将其看

成真值;

②生成不同的随机误差 $\delta_j, \varepsilon_{j0}, \varepsilon_{j1}, \cdots, \varepsilon_{jf_j}$,使 $a_{yj} = a_{xj}/A + \delta_j$,且使 $a_{yj} > 0$;$b_{yjt} = Ab_{xjt} + B + \varepsilon_{jt}$;

这里,$\delta_j \sim N(0, \sigma_1^2)$,$\varepsilon_{jt} \sim N(0, \sigma_2^2)$,$\sigma_1^2 = 1/900$,$\sigma_2^2 = 1/400$,$N(\mu, \sigma^2)$ 表示均值为 μ,方差为 σ^2 的正态分布。

③根据 $\{a_{xj}, b_{xj0}, b_{xj1}, \cdots, b_{xjf_j}\}$ 和 $\{a_{yj}, b_{yj0}, b_{yj1}, \cdots, b_{yjf_j}\}$,用不同等值方法计算出等值系数 A,B,用 A_r,B_r 表示第 r 种等值方法计算出的等值系数;

④重复②、③k 次,则对每一种等值方法可以得到 k 组不同的等值系数 $A^{(h)}, B^{(h)}, h = 1, 2, \cdots, k$,计算 RMSD(root mean square deviation),

$$RMSD(r) = \sqrt{\sum_{h=1}^{k}[(A_r^{(h)} - A)^2 + (B_r^{(h)} - B)^2]/k};$$

⑤设有等值方法:H 方法,SL 方法,SRE 方法,对数对比方法(记为 LC 方法),则 r = 1,2,3,4,比较 RMSD(r),进行 Wilcoxon 符号秩检验[12];

重新预设等值系数 A,B,再重复②—⑤,以上做法是以 X 测验上这 m 个锚题为基础。

对不同的 A,B 及不同的等值法进行比较,RMSD 是估计值对真值的偏移的均方根,故比较原则是基于估计值对真值的修复(recovery)程度,若某种等值方法对应的 RMSD 越小,修复程度越好,我们认为这种等值方法越好。

本文以下部分将 SL 产生的 RMSD 的值 x 直接写成 SL = x,其他方法产生的 RMSD 也采用这种记法;各方法按 RMSD 从小到大排列;对 Wilcoxon 符号秩检验结果,我们分别用 *,* * ,* * * ,* * * * ,* * * * * 表示在 $p = 0.05, 0.01, 0.005, 0.001, 0.0001$ 水平上有显著差异,如果方法 i 与方法 j 相比,每次算出的 RMSD 都小,则称方法 i 优于方法 j。

为了使 MC 模拟结果与统计分析结果表现得更简洁、更直观,我们给出了一种图示方法如下:用小圆圈表示不同的等值方法,如果方法 i 和方法 j 在显著性水平 p 上有差异,且方法 i 比方法 j 的 RMSD 小,则将代表方法 i 的小圆圈画在代表方法 j 的小圆圈之上,且从圆圈 i 出发到圆圈 j 连线,在该连线旁边标注上显著性水平 p。由此方法给出的图称为 RMSD 比较图。

例 4 0,1 评分题 12 道,符合 2PLM,对四种等值方法(H 方法,SL 方法,SRE 方法,LC 方法)进行比较。每种等值系数的组合生成 50 批 Y 测验上的项目参数(即 K = 50)

对于 $A < 1, B < 0$($A = 0.886, B = -0.225$)有:

$SRE = 0.02116$,$H = 0.022926$,$SL = 0.023479$,$LC = 0.029758$,

表 1 (2PLM)Wilcoxon 符号秩检验结果($A < 1, B < 0$)

比较对象	符号秩	显著性
H:SRE	$W^- = 429, W^+ = 846$	*
H:LC	$W^- = 952, W^+ = 323$	* *
H:SL	$W^- = 734, W^+ = 541$	
LC:SRE	$W^- = 252, W^+ = 1023$	* * * *
SL:SRE	$W^- = 360, W^+ = 915$	* *
SL:LC	$W^- = 909, W^+ = 366$	* *

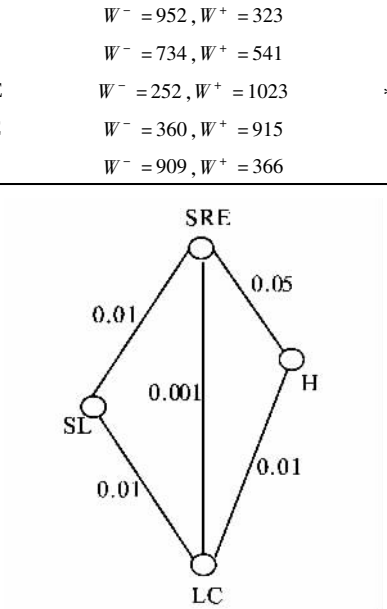

图 1 RMSD 比较图 $A < 1, B < 0$

按照作图原则,以表 1 为例,用四个小圆圈分别表示四种方法,并在小圆圈旁边标记方法名称,其中 SRE 方法的 RMSD 比 H 方法、SL 方法、LC 方法的都小,则将 SRE 画在 H、SL、LC 之上,且在 SRE 与 H 之间、SRE 与 SL 之间、SRE 与 LC 之间各画一条连线,并在连线边上分别注明显著性水平 0.05、0.01、0.001;同理 H 方法的 RMSD 比 LC 方法的小,则也将 H 画在 LC 之上,在 H 与 LC 之间画一条连线,并在连线边上注明著性水平 0.01;SL 方法的 RMSD 比 LC 方法的小,则也将 SL 画在 LC 之上,在 SL 与 LC 之间画一条连线,并在连线边上注明著性水平 0.01;于是可以作出相应于表 1 的 RMSD 比较图 1。

对于 $A \leqslant 1, B \leqslant 0$($A = 1.032, B = -0.072$),$A < 1, B > 0$($A = 0.857, B = 0.245$),$A > 1, B < 0$($A = 1.345, B = -0.234$),$A > 1, B > 0$($A = 1.356, B = 0.278$)我们也分别作了如例 4 所示的 MC 模拟研究及统计检验,相仿,可以得到另外四个与表 1 类似的表以及类似于图 1 的图,限于篇幅,我们只列出了相应的 RMSD 比较图。

MC 模拟结果显示,当 $A \cong 1, B \cong 0$ 和 $A > 1, B >$

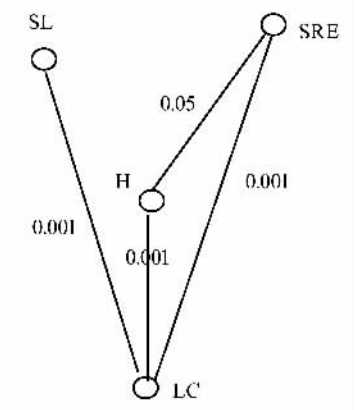

图2　RMSD 比较图　$A<1, B>0$

0 时各方法在给定显著水平($p \leqslant 0.1$)上无差异。而对于 $A<1, B>0, A>1, B<0$,我们将 MC 模拟及假设检验的结果所对应的 RMSD 比较图(图 2,图 3)列于图 1 下面。

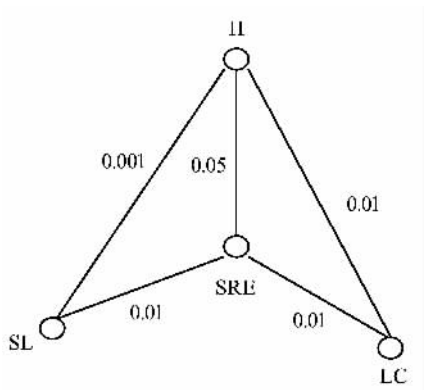

图3　RMSD 比较图　$A>1, B<0$

例 5　取锚题数 $m=12$,每题至少有三等级,至多有四等级,对四种等值方法(H 方法,SL 方法,SRE 方法,LC 方法)进行比较,每种等值系数的组合生成 30 批 Y 测验上的项目参数(即 k = 30)。

对于 $A \cong 1, B \cong 0$ ($A=1.032, B=-0.072$)有:

$SL=0.031539, LC=0.032502, H=0.0042583, SRE=0.079246$。相应的符号秩检验结果列于表 2,对应比较图为图 4。

表 2　Wilcoxon 符号秩检验结果($A \cong 1, B \cong 0$)

比较对象	符号秩	显著性
H:SRE	$W^-=455, W^+=10$	* * * * *
H:LC	$W^-=112, W^+=353$	* *
H:SL	$W^-=30, W^+=435$	* * * * *
LC:SRE	$W^-=460, W^+=5$	* * * * *
SL:SRE	$W^-=460, W^+=5$	* * * * *
SL:LC	$W^-=240, W^+=225$	

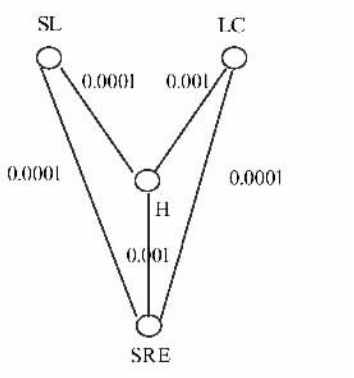

图4　RMSD 比较图　$A \cong 1, B \cong 0$

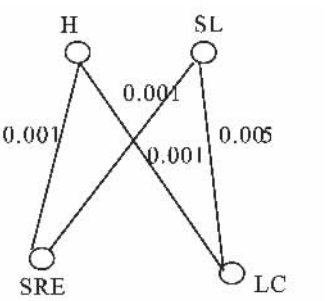

图5　RMSD 比较图　$A<1, B>0$

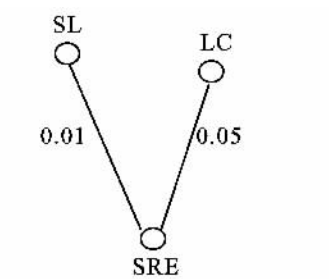

图6　RMSD 比较图　$A<1, B>0$

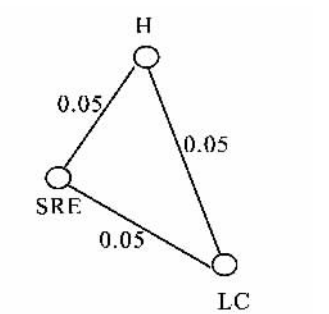

图7　RMSD 比较图　$A>1, B<0$

同理可得其他情形下的比较图 5 ~ 图 8。

4　结论和讨论

对于 GPCM 来讲,(ⅰ)当 $A \cong 1, B \cong 0$ 时,对数对比方法表现较为突出,和 H 方法相比有显著差

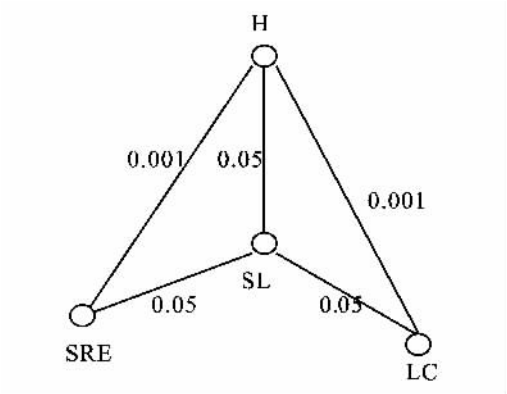

图 8 RMSD 比较图 $A>1, B>0$

异,特别地比 SRE 要好(即重复做 30 次试验,几乎每次试验计算出的等值系数的偏移均方根 RMSD 都小于 SRE 对应的 RMSD),且对数对比方法和 SL 方法在 RMSD 方面不相上下,但对数对比方法有以下两个优点:第一,对迭代初值要求不严;第二,收敛速度快,这是 SL 方法不能够与之相比的。(ⅱ)对于其他四种情形,对数对比方法的 RMSD 均较大或最大,但是对这四种情形,如果用均值 – 标准差方法提供的值作为迭代初值,SL 方法、H 方法、SRE 方法几乎都不能收敛,纵使某种情况下可收敛,也需要进行很多次迭代,这时用对数对比方法计算结果作为初值进行迭代,则迭代较少的次数便可收敛,在大多数情况下迭代结果能改善对数对比方法提供的初值,得到较好的结果,也有时不改变初值或得到更差的结果。

对于 0,1 评分项目,针对 2PLM,除 $A\cong 1, B\cong 0$ 及 $A>1, B>0$ 时,四种方法在 $p\leqslant 0.1$ 时无显著差异外,对数对比方法在计算精度方面没有显示出其优越性,然而令人吃惊的是,SL 方法有时也和对数对比方法一样不太令人满意,但是对数对比仍然保留对初值要求不严且收敛速度快的特点。

综上所述,对于以除法形式定义的多等级评分模型,采用对数对比等值法,不仅收敛速度快,而且对迭代初值要求不严,所得结果也较好,可以作为其他等值方法迭代计算的良好初值。

由(5)和(7.2),对数对比也可以不将 P_{aj0} 作分母,而采用某个其他的固定项,比如 P_{ajt} 作分母,只要在特定模型下使运算变得简单即可。

参 考 文 献

1 Embretson S E, Reise S P, Item Response Theory for Psychologists. Lawrence Erlbaum Associates Publishers, 2000

2 Ou D M. Constructing Metric System of item pool for Qualification examinations(in Chinese). Unpublished doctoral dissertation, Beijing Normal University, 2002. 30 ~ 34
(欧东明,资格考试题库度量系统的构建,北京师范大学博士学位论文,2002)

3 Qi S Q, Dai H Q, Ding S L. Modern Educational and Psychological Measument(in Chinese). Nan chang: Jiangxi Education Press, 1998. 229 ~ 236
(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)

4 Dodd B G, De Ayala R J, Koch W R. Computerized Adaptive Testing with Polytomous Items. Applied Psychological Measurement, 1995, 19(1): 5 ~ 22

5 Kolen M J, Brennan R L. Test Equating: Methods and Practices. New York: Springer – Verlag, New York. Inc. 1995. 169 ~ 173

6 Ding S L, Xiong J H. Treat some common equating methods in a unified form(in Chinese). In: Huang R H, Chen M L ed. Proceedings of the 6th Global Chinese Conference on Computers in Education/National Education Information Forum 2002. Bejing: Center of Broadcasting and Televising University press, 2002. 457 ~ 460
(丁树良,熊建华. 几种常见等值方法的统一处理,第六届全球华人计算机教育应用大会暨 2002 年全国教育信息化论坛论文集. 北京:中央广播电视大学出版社,2002. 457 ~ 460)

7 Aitchison J. The Statistical Analysis of Compositional Data. Chapman & Hall Ltd, 1986
(J. 艾奇逊著. 周蒂等译. 成分数据的统计分析. 武汉:中国地质大学出版社,1990. 49 ~ 50)

8 Li P L. Statistical Applications for Social Research(in Chinese). Bejing: Social Science Documents Publishing, 2001. 272, 306 ~ 309, 316 ~ 319
(李沛良. 社会研究的统计应用. 北京:社会科学文献出版社,2001. 272, 306 ~ 309, 316 ~ 319)

9 Zhang Y T, et al. Statistical Analysis for Quantitative Data(in Chinese). Guilin: Guangxi Normal University Press, 1991. 111 ~ 164
(张尧庭等. 定性资料的统计分析. 桂林:广西师范大学出版社,1991. 111 ~ 164)

10 Zhu K H, Zhou J X. Test Equating on Bichotomous Items(in Chinese). Chinese Journal of Applied Probability and Statistics, 2001, 17(3): 260 ~ 266
(朱奎花,周纪芗. 二级评分题目测验的等值. 应用概率统计,2001,17(3): 260 ~ 266)

11 Agresti A. Categorical Data Analysis. New York: Wiley, Inc. 1990. 306 ~ 318

12 Hosmer, D. W., and S. Lemeshow., Applied Logistic Regression (2nd). New York: Wiley, Inc, 2000. 288 ~ 290

13 Baker F B. Item Response Theory: Parameter Estimation Techniques. New York: Marcel Dekker, Inc, 1992. 57 ~ 62

14 Miller T R, Spray J A. Logistic regression for DIF of polytomously scored items. Journal of Educational Measurement, 1993, 30(2): 107 ~ 122

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

16 Wu X Z, Wang Z P. Nonparametric Statistical Methods(in Chinese). Bejing: Higher Education Press, 1996. 35 ~ 41
(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)

17 Shi M G, Gu L Z. The Foundation of Scientific and Engineering Computation(in Chinese). Bejing: Tsing hua University Press, 1999. 298 ~ 300
(施妙根,顾丽珍. 科学和工程计算基础. 北京:清华大学出版社 1999. 298 ~ 300)

附录(GPCM 的对数对比等值计算公式及迭代收敛性质)

计算公式

记 $F = LC(1)$, $LC(1)$ 如(7.2)所示,根据(4)、(5),记

$$f_1 = \frac{\partial F}{\partial A}, f_2 = \frac{\partial F}{\partial B}, g_{11} = \frac{\partial^2 F}{\partial A^2}, g_{12} = g_{21}\frac{\partial^2 F}{\partial A \partial B},$$

$$g_{11} = \frac{\partial^2 F}{\partial B^2}$$

用 A_t, B_t 表示等值系数 A, B 的第 t 次迭代值,则

$$\begin{pmatrix} A_{t+1} \\ B_{t+1} \end{pmatrix} = \begin{pmatrix} A_t \\ B_t \end{pmatrix} - \begin{pmatrix} g_{11} & g_{12} \\ g_{11} & g_{12} \end{pmatrix}^{-1} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}_{A=A_t, B=B_t} \quad (14)$$

以上迭代一直进行到相邻两次迭代值充分接近为止,(14)给出了由(7.2)式求解等值系数 A, B 的迭代式,并称由(14)导出的 A, B 为由对数对比等值法导出的等值系数。以上文中(9)式为例,我们推导出具体计算公式如下:

对(9)求导并令导数为 0,

记 $\xi_a = \sum_{j=1}^{m}\sum_{l=1}^{f_j}\{\sum_{k=1}^{t}[a_{yj}(A\theta_{xa} + B - b_{yjk})$

$\{-a_{xj}(\theta_{xa} + B - b_{yjk}) - a_{xj}(\theta_{xa} - b_{xjk})]\} \cdot t \cdot a_{yj}$,

则 $f_1 = \sum_{a=1}^{N}\theta_{xa}\xi_a$, $f_2 = \sum_{a=1}^{N}\xi_a$,再令 $\omega = \sum_{a=1}^{N} a_{yj}^2(\sum_{t=1}^{f_j} t^2)$,

则, $g_{11} = \omega\sum_{a=1}^{N}\theta_{xa}^2$, $g_{12} = g = {}_{21} = \omega\sum_{a=1}^{N}\theta_{xa}$,

$g_{22} = \omega\sum_{a=1}^{N}\omega = N\omega$

根据迭代公式(7),则给定初值 A_0, B_0 以后,可求解 A, B。

对于 2PLM,其准则式由(13)给出,仿 GPCM,结合具体的类目特征曲线 P_{xaj}, P_{yaj},则也可以给出相应的求取等值常数 A, B 的公式,在此不再赘述。

迭代收敛性质

由于迭代计算公式 $\begin{pmatrix} A \\ B \end{pmatrix}_{t+1} = \begin{pmatrix} A \\ B \end{pmatrix}_t - (Df)^{-1} f$ 中 Df 不含未知参数 A, B,故对任何 $x = (A, B)$,设 $x_* = (A_*, B_*)$ 为方程的根,则 $\| Df(x) - Df(x_*) \| = 0$,从而对任意 $a > 0$,有 $\| Df(x) - Df(x_*) \| \leqslant a \| x - x_* \|$,知该迭代方法至少二次收敛于 x_*[17],而通常 N - R 迭代具有二次局部收敛。

另外,MC 模拟表明,对于 $A \cong 1, B \cong 0$,任意给出初值,比如令初值为 $A = 100, B = 30$,该迭代也能很快收敛,其他情形也有类似结果。故这一迭代方法对初值是稳健的(robust),这与 H 方法, SL 方法, SRE 方法大不相同。

LOGCONTRAST METHOD FOR EQUATING TEST BASED ON IRT

Ding Shuliang, Xiong Jianhua, Mao Mengmeng
(*College of Computer Science and Technology, Jiangxi Normal University, Nanchang* 330027, *China*)

Abstract

There are a lot of polytomous item response theory models known as devide - by - total. The equating test for these models is very difficult, because hardly can one to find out quite accurately initial values of the equating coefficients. Lack of the accuracy of the initial values leads to the Haebara approach, Stocking-Lord approach, Symmetric Relative Entropy approach to fail for equating tests under these models. A new equating method, Logcontrast approach, is introduced in this paper. This approach has some advantages for these models, such as the robustness for choicing the initial values of the iteration, fast convergent, and accurate result. And Logcontrast equating approach could supply quite accurate initial values for the rest of the equating methods mentioned above. Moreover, Logcontrast approach generalized and impoved the logit method for equating the dichotomous items.

Key words polytomous score models, test equating, Logcontrast method for equating test, initial value for the interation.