doi:10.3969/j.issn.1672-5565.2013.03.06

非正态验证性因子分析在基因整体效应中的应用

刘小琴,马 瑞,罗艳虹,李 治,张春森,张岩波*
(山西医科大学公共卫生学院卫生学统计教研室,山西 太原 030001)

摘 要:针对SNPs数据不服从正态分布的情况,拟采用S-B测度调整估计方法拟合验证性因子模型,进行SNPs整体效应和关联性分析。用GAW17提供的SNPs数据进行实例分析。本研究随机选取2号染色体上,分布在6个基因之中的13个SNPs作为研究对象,对选取的6个基因做潜变量得分,然后对基因和疾病感染做检验。结果显示:χ^2/df最大似然估计方法的卡方自由度比为3.59,S-B测度调整估计方法的卡方自由度比χ^2/df为2.89,最大似然估计方法的RMSEA为0.061,S-B测度调整估计方法的RMSEA为0.052。6个基因对该感染都有影响.由此得出结论,在处理SNPs数据时,使用S-B测度调整估计能得到更好的拟合模型。可以推测这6个基因下的13个SNP位点可能是感染的致病位点。

关键词:单核苷酸多态性;非正态;最大似然估计;S-B测度调整估计;验证性因子分析
中图分类号:Q75 **文献标识码**:A **文章编号**:1672-5565(2013)-03-192-04

Non-normal confirmatory factor analysis in the application of the whole gene effect

LIU Xiao-qin, MA Rui, LUO Yan-hong, LI Zhi, ZHANG Chun-sen, ZHANG Yan-bo*
(*Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan* 030001, *China*)

Abstract: This paper proposed S-B measure (scaled) estimates to fit confirmatory factor models, to analysis overall effect and correlation of SNPs which does not fit normal distribution. Example of SNPs data is provided by GAW17. The study chooses 13 SNPs located 6 gene in chromosome 2, we firstly do latent variables score in the six genes, and genes and infections do t-test. Maximum likelihood estimation Chi square degrees of freedom χ^2/df *is* 3.59, S-B scaled method χ^2/df *is* 2.89, maximum likelihood estimation RMSEA is 0.061, S-B method RMSEA is 0.052. Six genes on the infection have influence. When analysis the SNPs data, using S-B estimated can get a better fitted model. We can speculate that the 13 SNPs sites in 6 genes may be the infection pathogenic site.

Key words: Single Nucleotide Polymorphism; Nn-normal; Mximum Lielihood Estimation; Torra-bentler Saled Etimation; Cnfirmatory Fctor Aalysis

在后基因组时代,单核苷酸多态性(single nucleotide polymorphism,SNP)研究已成为统计遗传学领域研究的热点。有学者将潜在结构模型(latent structural model)或潜变量模型(latent variable model)引入单体型或高维SNP整体效应的关联分析及其相关的推断性研究。潜变量模型要求观测变量与潜变量均服从正态分布,而SNP数据无论以何种遗传模式量化,都很难符合正态假定。目前,国内已发表的大多文章,当SNP数据不满足正态分布时,仍使用ML估计拟合模型。张岩波教授在《潜变量分析》一书中通过模拟研究和实例分析得出:数据越偏离正态,得到的χ^2统计量越高;当样本含量大于250,数据偏离正态时,使用S-B方法估计的模型最理想[1]。

收稿日期:2013-01-07;**修回日期**:2013-03-13.
资助项目:国家自然科学基金资助项目(31071156)。
作者简介:刘小琴,女,山西洪洞人,硕士研究生,研究方向:生物信息统计;E-mail:shanyiliuxiaoqin@163.com.
*通讯作者:张岩波,男,山西长治人,教授,研究方向:生物信息统计,生存质量评价;E-mail: yanbozh@126.com.

1　原理与方法

1.1　最大似然估计

最大似然估计(maximum likelihood estimation，ML 估计)方法是验证性因子分析(confirmatory factor analysis，CFA)模型中最重要的参数估计方法，它也是许多结构方程模型(structural equation model,SEM)软件(如 EQS)分析时默认的参数估计方法。其目标函数形式是[2]：

$$F_{ML} = \log|\sum| - \log|S| + tr(S\sum^{-1} - p) \quad (1)$$

也可以表述为二次型(quadratic)形式：

$$F_{QD} = [s - \sigma(\theta)]' W^{-1}[s - \sigma(\theta)] \quad (2)$$

这里 s 的 $\sigma(\theta)$ 和分别是矩阵 S 和 $\sum(\theta)$ 中所有不重复元素构成的 $p(p+1)/2\times 1$ 的列向量，W 是 $p(p+1)/2\times p(p+1)/2$ 的正定权重矩阵，当 ML 估计目标函数转换为二次型形式，它的 $W = 2K_p'(\hat{\sum}\otimes\hat{\sum})K_p$，$K_p$ 是已知的转换矩阵(transition matrix)。H_0 成立前提下，即当 $\hat{\sum} = \sum(\hat{\theta})$ 时，目标函数 F_{ML} 达到最小值 $\hat{F}_{ML}$，$T_{ML} = (n-1)\hat{F}_{ML}$ 渐近服从 $df = p(p+1)/2 - q$ 的中心卡方分布，T_{ML} 即是 ML 卡方检验统计量。

1.2　S-B 测度调整(scaled)估计(S-B 估计)

1.2.1　S-B 调整卡方统计量

S-B 卡方调整的是由于不满足正态分布所造成的卡方增加值，也就是，非正态的自由度越大，S-B 卡方能降低的卡方值越大，对 ML 估计方法得到的卡方统计量进行校正，使其更接近于 χ^2 分布，得到的卡方统计量即为 S-B 调整卡方统计量(scaled chi-square test statistic)，它的原理是：

CFA 模型检验统计量 T 的分布不是 χ^2_{df} 分布，而是一个混合分布：

$$T \xrightarrow{L} \sum_{1}^{df} a_i\tau_i \quad (3)$$

其中，a_i 为矩阵 UV_{SS} 的 df 个非零特征根中的一个。V_{SS} 是 $\sqrt{n}[s-\sigma(\theta)]$ 的渐近协方差矩阵，它是包含样本观测变量四阶矩的一个任意分布估计量。$U = W^{-1} - W^{-1}$ 为用 W 估计的残差权重矩阵。T_{ML} 的渐近分布的均数由 $tr(UV_{SS})$ 得到，$\tau_{是}$ df 个独立的 χ^2_1 变量之一。

调整校正因子(scaling correction factor)为 $\kappa = tr(\hat{U}\hat{U}_{SS})/df$，$\hat{U}$ 是 U 的一致性估计，S-B 估计的检验统计量为：

$$T_{S-B} = T_{ML}/\kappa \quad (4)$$

1.2.2　调整标准误

Browne(1982)提出一种在非正态条件时对 ML 估计的标准误进行调整的方法，Bentler 和 Dijkstra(1985)又对该调整方法进行了改进[3]。调整后的标准误可以在 EQS 软件中得到，称为稳健标准误(robust Standard error)，稳健标准误可通过对调整后的协差阵的主对角线的元素取平方根得到。

2　实例分析

本次研究数据来源于 GAW17(Genetic Analysis Workshop)的数据库，随机选取数据库中 2 号染色体上分布在 6 个基因之中的 13 个 SNPs 为研究对象。研究目是探讨用 ML 估计方法和 S-B 估计方法估计 SNP 数据时，模型的拟合优度。本次研究选择加法模型量化方法对 SNP 位点赋值(即假设等位基因 A 对疾病发生的贡献量(危险性)为 r，则基因型 Aa 的贡献量就为 r，而基因型 AA 的贡献量为 2r，基因型 aa 的贡献量为 0，其量化方法是：当个体的基因型为 AA 时赋值为 2，Aa 时赋值为 1，aa 时赋值为 0)，如对 C2S192 这个 SNP 赋值：CC 赋值为 0，CT 或者 TC 赋值为 1，TT 赋值为 2。表 1 为 6 个基因与 13 个 SNP 的关系。

表 1　各基因与 SNP 的关系

Table 1　The relationship between the gene and SNP

基因名称	SNP		基因名称	SNP	
	名称	类型		名称	类型
APOB	C2S192	CC/CT/TC/TT	LRP1B	C2S3443	CC/CT/TC/TT
	C2S200	CC/CT/TC/TT		C2S3490	AA/AG/GA/GG
ALK	C2S454	AA/AG/GA/GG	TTLL4	C2S5995	AA/AG/GA/GG
	C2S466	CC/CT/TC/TT		C2S6031	AA/AG/GA/GG
MAP4K3	C2S842	AA/AT/TA/TT	COL6A3	C2S7326	AA/AG/GA/GG
	C2S867	AA/AG/GA/GG		C2S7422	AA/AG/GA/GG
				C2S7528	AA/AG/GA/GG

表 2 列出了 EQS 软件输出的 13 个 SNP 的偏度系数和峰度系数,可知各个 SNP 均不服从正态分布。且 SNP 数据明显的高峰度特征影响 ML 估计法对 CFA 模型的估计。EQS 软件中输出的 Mardia 系数值为 91.012 4,该值大于 1.96,说明 SNP 数据不服从多元正态分布。

表 2 SNPs 偏度系数、峰度系数

Table 2 skewness coefficient and kurtosis coefficient of SNPs

基因名称	SNP 名称	偏度系数	峰度系数
APOB	C2S192	2.62	6.40
	C2S200	0.32	-0.70
ALK	C2S454	1.37	0.62
	C2S466	0.56	-1.14
MAP4K3	C2S842	0.09	-1.54
	C2S867	0.37	-1.51
LRP1B	C2S3443	0.93	-0.15
	C2S3490	1.33	0.60
TTLL4	C2S5995	0.21	-1.50
	C2S6031	0.18	-1.60
COL6A3	C2S7326	2.76	7.36
	C2S7422	2.65	6.44
C2S7528		2.58	6.21

由表 3 可知,ML 法的参数标准误虽然没有影响到最后的结论,但是数据偏低。由表 4 可见,ML 估计方法的 χ^2/df 为 3.59,S - B 估计方法的 χ^2/df 为 2.89,ML 估计方法的 RMSEA 为 0.061,S-B 估计方法的 RMSEA 为 0.052。由此可处理 SNPs 数据时,使用 S-B 估计方法所得的拟合模型更好。

表 3 SNPs 测量模型的结果

Table 3 The result of the SNPs measurement model

基因名称	SNP 名称	因子载荷(标准)	标准误	
			ML	S-B
APOB	C2S192	0.49	-	-
	C2S200	0.94	0.319	0.334
ALK	C2S454	0.26	-	-
	C2S466	0.80	0.753	0.742
MAP4K3	C2S842	0.92	-	-
	C2S867	0.96	0.026	0.023
LRP1B	C2S3443	0.88	-	-
	C2S3490	0.83	0.033	0.035
TTLL4	C2S5995	0.95	-	-
	C2S6031	0.99	0.020	0.023
COL6A3	C2S7326	0.97	-	-
	C2S7422	0.97	0.018	0.040
	C2S7528	0.62	0.033	0.060

表 4 ML 估计与 S-B 估计拟合指标

Table 4 The fitting index of ML estimation and S-B estimation

参数估计方法	拟合指标					
	χ^2/df	CFI	NNF	IRMSEA	IFI	MFI
ML 估计	3.59	0.982	0.971	0.061	0.982	0.911
S - B 估计	2.89	0.982	0.974	0.052	0.983	0.934

对选取的 6 个基因做潜变量得分,然后对基因和疾病患病情况做检验。由表 5 可得 6 个基因对该疾病的患病情况都有影响。我们可假设这些基因所对应的 SNP 位点就是造成患病的致病位点,结合基因学的相关知识,就可以明确这些位点是不是真正的致病位点。

表 5 基因与疾病间关系

Table 5 The relationship between genes and disease

基因	分组	n	$\bar{x} \pm s$	t	P
APOB	正常组	488	-0.076 ±1.009	-3.135	0.002
	病例组	209	0.177 ±0.957		
ALK	正常组	488	-0.911 ±0.984	-3.710	<0.001
	病例组	209	0.213 ±0.973		
MAP4K3	正常组	488	-0.071 ±0.984	-2.874	0.004
	病例组	209	0.166 ±1.021		
LRP1B	正常组	488	-0.079 ±0.980	-3.211	0.001
	病例组	209	0.1856 ±1.02		
TTLL4	正常组	488	-0.085 ±0.982	-3.445	0.001
	病例组	209	0.198 ±1.015		
COL6A3	正常组	488	-0.978 ±0.875	-3.502	0.001
	病例组	209	0.228 ±1.216		

3 讨论

(1)LISREL 和 AMOS 软件默认的验证性因子模型分析的参数估计方法为 ML 估计,在使用 ML 估计方法时,必须先考察观测变量的正态性。当用 ML 估计非正态数据时,由于卡方统计量增大进行拟合优度卡方检验时产生 I 型错误的概率增加,同时降低了参数的标准误。West 等综合了各种研究结果得出,当数据成非正态或者数据为正态但观测变量值很小时,ML 法高估了,而低估了参数的标准误。造成偏高、参数标准误偏低的原因可能是由于修正模型时生成了多余相关、多余参数的复杂模型[4]。

(2)若观测变量值明显违反正态性假设且样本量大时,用 Browne's 渐近任意分布方法;对于类似 SNP 数据,不论如何赋值都不满足正态分布,同时也

没有足够大的样本支持任意分布估计方法(ADF 方法要求样本量至少是 2 500,得到的结果才趋于稳定[5])时,用 S-B 测度调整估计方法。

(3)本文针对 SNP 数据不满足正态分布的特点,将 S-B 调整估计方法和 ML 估计方法进行对比研究,使用 S-B 调整方法,所得的模型拟合指标如 RMSEA 等,都优于用 ML 估计所得的拟合指标,同时可减少传统方法所导致的多重检验和检验功效降低。因此,本文提出在进行 SNPs 整体效应分析时,S-B 调整估计方法是首选的参数估计方法,为研究 SNPs 与疾病的关联性提供了一种新方法。这种方法应用到实际的科研当中,可找到可疑的致病位点,再经过相关的分子生物学检测手段,可确认此可疑致病位点是否真正为致病位点,为疾病的诊断和治疗提供帮助。

参考文献(References)

[1] 张岩波. 潜变量分析[M]. 北京:高等教育出版社, 2009,138-153.

[2] 郝彦斌. 小样本非正态数据结构方程模型估计方法研究与医学应用[D]. 山西:山西医科大学,2007.

[3] Satorra, A, P. M. Bentler. Latent. Variables Analysis: Application to Developmental Research[M]. Thousand Oaks, CA: SAGE Publications, 1994. 399-419.

[4] 邱皓政,林碧芳. 结构方程模型的原理与应用[M]. 北京:中国轻工业出版社,2009. 43-194.

[5] Rick H. Hoyle. Structural Equation Modeling: Concepts, Issues and Applications[M]. Newbury Park CA: SAGE Publications, 1995. 56-75.