doi:10.3969/j.issn.1672-5565.2013.02.13

结构方程混合模型在SNP分析中的应用

杨圆圆, 贾志杰, 李 治 ,罗艳虹 ,张岩波*
(山西医科大学卫生统计学教研室,山西 太原 030001)

摘 要:采用结构方程混合模型(SEMM)对实际SNP数据进行分析,为遗传统计学提供一种新的有效的分析方法。本研究的数据是由GAW17提供的,包含697个个体的22条常染色体的上万个SNP和根据这些SNP所模拟的697个个体的性状特点。随机挑选了1号染色体上的4个SNP和3个定量性状作为研究变量,分别进行潜在类别分析和结构方程混合模型分析。根据4个SNP数据,人群被分为3个潜在类别,概率分别为0.53,0.34,0.13。潜在类别1、2和3中的因子均值Q分别为-4.029、-2.052和0,潜在类别1、2的因子均值均低于3(<0.001)。研究表明:结构方程混合模型(SEMM)综合了结构方程模型和潜在类别模型的思想,形成了自己的优势,可用于处理同时包含分类潜变量和连续潜变量的数据。

关键词:结构方程混合模型(SEMM);单核苷酸多态性(SNP);连续潜变量;分类潜变量
中图分类号:Q343.1+5 **文献标识码**:A **文章编号**:1672-5565(2012)-02-146-04

Structural equation mixture modeling and their application in analysis of SNP

YANG Yuan-yuan, JIA Zhi-jie, LI Zhi, LUO Yan-hong, ZHANG Yan-bo*
(*Department of Health Statistics, School of Public health , Shanxi Medical University, Taiyuan* 030001, *China*)

Abstract: To analyze SNP data of GAW17 by Structural equation mixture modeling (SEMM), and to provide a new method for the study of genetic statistic . The data is provided by GAW17, it contains 697 individual, 22 autonomic tens of thousands of SNP and the SNP simulated 697 individual trait characteristics. In this study, randomly selected the four SNP from chromosome 1 and three quantitative traits as a research variable, which were analysised by latent class and mixed structural equation modeling. According to four SNP data, the crowd was divided into three potential categories, each category probability were 0.53, 0.34, 0.13. Factors mean Q of latent class 1, 2 and 3 are -4.029, -2.052 and 0. We knew that factor mean of latent class 1, 2 are lower than 3 (<0.001). So we have reasons to think that structural equation mixed modeling integrated the structural equation modeling and latent class modeling thoughts, formed its own advantage, which can be used for processing classification latent variable and continuous latent variable data.

Keywords: Structural Equation Mixture Modeling(SEMM); SNP; Continuous Latent; Categorical Latent

近年来随着基因分型技术和新一代测序技术的发展,基于群体的基因关联分析成为国际上遗传统计学领域的热点。以往分析SNP数据时,常在各种假定的遗传模式(显性模型、隐形模型、加法模型和乘法模型)下对个体基因型进行赋值量化[1],但是这种量化有许多不确定性。潜类分析对SNP既可采用某特定遗传模式,也可忽略遗传模式以保留原始的分类信息,更细致地考察基因型的概率分布特征。因此,无论心理因素,还是SNP数据 ,采用潜变量分析独具优势[2],但是当资料中同时包含临床症状、社会心理因素与SNP等不同类型的数据时,传统的潜变量分析存在缺陷,用结构方程模型分析社会心理因素时效果良好,但是用其分析SNP数据时违背结构方程模型的前提假设;潜在类别分析考察

收稿日期:2012-11-13;**修回日期**:2012-11-26.
基金项目:国家自然科学基金资助项目(31071156)。
作者简介:杨圆圆,山西医科大学,硕士研究生。
*通讯作者:张岩波,教授,博士生导师。E-mail: yanbozh@126.com.

SNP 的分布特征或评价其整体效应时,是一种强而有力的分析工具,但是无法分析连续变量。而结构方程混合模型(SEMM)综合了结构方程模型和潜在类别模型的思想,描述的是既包含分类潜变量又包含连续潜变量的数学模型,它的构建拓展了潜变量模型的应用范围。

1　资料与方法

1.1　研究资料

数据是 GAW17(Gennetic Analysis Workshop)工作组提供的,包含家系数据和独立个体的数据两部分。两部分数据都包含关于 22 条常染色体的 697 个个体的 24 487 个 SNPs 信息,包括 SNP 的名称、SNP 所在的染色体编号、基因名称和 SNP 的最小等位基因频率等。数据中还有根据这些 SNP 所模拟的 697 个个体的性状特点,包括 3 个定量性状(Q1,Q2,Q3)和 1 个定性性状。本文选取 1 号染色体上的 4 个 SNP 进行研究,表 1 是 4 个 SNP 的具体信息,研究目的是分析它们的关联情况以及与性状之间的关系。每个 SNP 位点被分为 3 类,第一个类别为该 SNP 位点最常见的纯合碱基对(比如 CC),第二个类别为该 SNP 位点的杂合碱基对(比如 CT 或 TC),第三个类别为该 SNP 位点变异的纯合碱基对(比如 TT)。

表 1　基因与 SNP 指标

Table 1　Gene and SNP index

Gene	SNP	
	dbSNPbID	Genetype
ANKRD38	C1S3592	GG/GA/AG/TT
	C1S3593	CC/CT/TC/TT
	C1S3594	CC/CT/TC/TT
	C1S3595	GG/GA/AG/AA

1.2　方法

1.2.1　结构方程混合模型的构建

传统的因子分析和结构方程模型都假设数据来自一个单一的同质总体,该假设中假定的平均分配参数往往是不现实的。如果异质性被忽略,统计分析的结果可能会出现严重的偏差。结构方程混合模型将类别潜变量和连续潜变量融合在一起,人口的异质性表示人群可能由两个或两个以上的不同潜在类别组成,不同的潜在组别之间存在不同的分布特点。这种人口异质性分析一直是混合结构模型的传统领域[3-4]。在结构方程混合模型中,潜在类别分析是用来确定潜在的分类。在结构部分的模型中,所有参数来自于各不相同的潜在类别。

1.2.2　参数估计

结构方程混合模型常用的估计方法有最大似然估计法(Maximum Likelihood Estimation)和 EM(Expected Maximum)算法。

1.2.3　模型评价

在结构方程混合模型中,仍然是通过常见的信息准则的比较,来确定潜在类别的最佳数目,使估计的模型最简洁、而且拟合优度较高。常用的措施包括赤池信息标准(AIC)、贝叶斯信息准则(BIC)和一致性赤池信息量准则(CAIC)。每个指标,最低的值被认为是最佳选择。

2　结果分析

2.1　潜在类别模型

潜在类别分析采用 Mplus 软件完成,结果如下:

表 2　不同类别数模型的拟合指标

Table 2　Model fitting index of different category numbers

模型	LL	BIC	AIC	χ2	G2	df	p
2 – cluster	–1 760.356 0	3 632.007 3	3 554.711 9	2 659.823 5	627.747 7	63	0.000
3 – cluster	–1 505.489 0	3 181.194 4	3 062.978 0	140.752 7	118.013 8	54	0.000
4 – cluster	–1 479.170 0	3 187.477 4	3 028.339 9	92.913 6	65.375 7	45	0.025
5 – cluster	–1 462.350 2	3 212.759 0	3 012.700 5	44.901 9	31.736 2	36	0.670
6 – cluster	–1 457.046 8	3 261.073 2	3 020.093 6	25.510 9	21.129 4	27	0.780
7 – cluster	–1 454.447 8	3 314.796 3	3 032.895 6	13.474 6	15.931 4	18	0.600
8 – cluster	–1 452.984 3	3 370.145 4	3 047.323 7	8.684 8	12.359 5	9	0.190

从表 2 所列 7 个模型的结果可知,当潜在类别数逐渐增大的时候,模型的对数似然值(LL)下降,卡方值也逐渐减少。7 个模型中,模型 3 的 BIC(3 181.194 4)值最低,模型 5 的 AIC(3 012.700 5)指标值最低。一般来说,当样本量很大时,建议以 BIC 指标作为模型适配性决策的标准[5],所以模型 3 为最佳模型。模型参数分析结果可得出各外显变量在各潜在类别上的条件概率值(见表 3)。

表 3 SNP 在三个潜在类别上的条件概率与潜在类别概率

Table 3 The conditional probability and latent class probability of SNP which in three latent class

SNPS		Genetype	类别 1	类别 2	类别 3
Gene	dbSNPb ID				
		0(GG)	0.003 4	0.018 2	0.644 0
	C1S3592	1(GA/AG)	0.017 7	0.936 5	0.189 0
		2(AA)	0.978 9	0.045 3	0.167 0
		0(CC)	0.009 1	0.008 9	0.788 5
ANKRD38	C1S3593	1(CT/TC)	0.048 3	0.879 8	0.046 6
		2(TT)	0.942 6	0.111 3	0.164 9
		0(CC)	0.002 9	0.004 6	0.836 3
	C1S3594	1(CT/TC)	0.104 8	0.965 0	0.046 3
		2(TT)	0.892 4	0.030 5	0.117 5
		0(AA)	0.005 1	0.005 5	0.383 7
	C1S3595	1(GA/AG)	0.059 8	0.962 8	0.521 9
		2(GG)	0.935 1	0.031 7	0.094 4
潜在类别概率			0.530 8	0.339 0	0.130 2

由表 3 知类别 1 在四个显变量中取值为 2 的概率较大，分别为 0.978 9、0.945 6、0.892 4 和 0.935 1，定义为变异纯合子。类别 2 在四个显变量中取值为 1 的概率较大，分别为 0.936 5、0.879 8、0.965 0 和 0.962 6，可定义为杂合子。类别 3 在四个显变量中取值为 0 的概率较大，分别为 0.644 0、0.788 5、0.836 3 和 0.383 7，定义为纯合子。通过潜在类别分析将 4 个潜在类别分析的 34 个组合用一个 3 分类的潜变量代替，各潜在类别概率分别为 0.530 8，0.339 0，0.130 2，总和为 1，表示三个潜在类别所占比例，类别 1 的比例最高。

表 4 各外显变量与潜在变量的关联强度

Table 4 Strength of a relationship of the observed variables and latent variables

Loadings	Clusters	R^2
C1S3592	0.875 0	0.765 7
C1S3593	0.830 4	0.689 6
C1S3594	0.850 5	0.723 4
C1S3595	0.836 8	0.700 3

表 4 类似因子分析的因子载荷，各负荷量的平方就是各 SNP 被解释的比率，数据显示，潜在变量与 C1S3592 的关系最强，因子载荷达到 0.875 0，R^2 为 0.765 7，表示潜在类别变量可以解释该 SNP 变异量的 76.57%。

2.2 结构方程混合模型

由上述的潜在类别分析可知四个 SNP 数据可以由一个 3 分类的潜变量解释，我们构建一个包含七个外显变量（包含四个 SNP 分类变量和三个连续变量 Q1、Q2 和 Q3）和 2 个潜变量的混合结构方程模型。结果（见表 5）。

表 5 潜在类别分析与混合结构方程在不同类别数模型的拟合指标比较

Table 5 The comparison of model fitting index between latent class analysis and mixed structure equation in different category number

	潜在类别分析					混合结构方程模型				
模型	χ^2	BIC	AIC	df	p	χ^2	BIC	AIC	df	p
1 – cluster	21 098.305 8	5 166.287 3	5 129.913 0	72	0.000	21 096.839	14 000.298	13 923.003	72	0.000
2 – cluster	2 659.823 5	3 632.007 3	3 554.711 9	63	0.000	2 522.637	12 156.212	12 033.449	63	0.000
3 – cluster	140.752 7	3 181.194 4	3 062.978 0	54	0.000	143.608	11 557.886	11 389.655	54	0.000
4 – cluster	92.913 6	3 187.477 4	3 028.339 9	45	0.025	106.470	11 585.521	11 371.822	45	0.000
5 – cluster	44.901 9	3 212.759 0	3 012.700 5	36	0.670	77.437	11 630.021	11 370.854	36	0.001
6 – cluster	25.510 9	3 261.073 2	3 020.093 6	27	0.780	97.437	11 696.860	11 392.225	27	0.000
7 – cluster	13.474 6	3 314.796 3	3 032.895 6	18	0.600	72.745	11 743.968	11 393.866	18	0.000
8 – cluster	8.684 8	3 370.145 4	3 047.323 7	9	0.190	70.385	11 798.102	11 402.531	9	0.000

由表 5 的潜在类别分析和混合结构方程模型的分析结果可知，在对 SNP 进行归类分析的时候，3 分类模型为最佳模型。分类结果与潜在类别分析的结果一致。

表 6 结构方程混合模型参数估计表和分类比例

Table 6 Parameter estimation table and classification ratio of Structural equation mixture modeling

	类别 1				类别 2				类别 3			
	因子载荷	标准因子载荷	标准误	t	因子载荷	标准因子载荷	标准误	t	因子载荷	标准因子载荷	标准误	t
Q1	1.000	0.854	0.028	30.924	1.000	0.854	0.028	30.924	1.000	0.854	0.028	30.924
Q2	0.304	0.648	0.030	17.114	0.304	0.648	0.030	17.114	0.304	0.648	0.030	17.114
Q3	0.130	0.466	0.040	11.802	0.130	0.466	0.040	11.802	0.130	0.466	0.040	11.802
潜在类别比例		0.53				0.34				0.13		

表 6 为结构方程混合模型的因子载荷和潜在类别比例表,由表 6 可以得出潜在类别 1、2 和 3 所占比例分别为 0.53、0.34 和 0.13,潜在类别 1 所占比例最高。最后的分类结果潜在类别 1、2 和 3 分别为 374 人、233 人和 90 人,与潜在类别分析的分类结果相似,结合潜在类别分析可将类别 1 定位为变异纯合子,类别 2 为杂合子,类别 3 为纯合子。

表 7　三种潜在类别的因子均值 Q 比较

Table 7　Factors mean θ of three Latent Classes

	因子均值	标准误	*t*
潜在类别 1	-4.029	0.223	-18.903
潜在类别 2	-2.052	0.168	-12.238

由表 7 可知,潜在类别 1、2 和 3 中的因子均值 Q 分别为 -4.029、-2.052 和 0,潜在类别 1、2 因子均值均低于潜在类别 3($p < 0.001$)。说明变异纯合子和杂合子的连续性状潜变量的因子均值 Q 小于纯合子的。

3　讨论

本文通过分析结果看出探索性潜在类别模型可以确定潜变量类别数,选出最佳模型。验证性混合模型是基于确定的潜在的类别数,检验特定的类别群体的结构关系及进行潜变量均值的比较。结构方程混合模型分析方法相对于以往的潜变量分析方法,具有不可比拟的优越性。它可以使潜变量的"降维简化"技术中众多复杂的变量综合明朗化,克服变量描述的单一性。更重要的一点是在结构方程混合模型中不仅通过构造分类潜变量,发掘出了潜在的类别,同时也对不同的潜在分类之间的连续潜变量进行比较研究。

混合模型的提出为解决人口中的同质性和异质性提供了有效的处理方法,当模型的类别数为 1 时,混合分布为单一分布,因而数据是同质的。当类别数大于 1 时,模型存在多个群体,数据存在异质性。本文提出的结构方程混合模型在解决传统潜变量模型中包含未知的潜在类别群体时独具优势,它成功的将类别潜变量引入到常规的潜变量模型中,为潜变量的发展提供新的思路。同时,结构方程混合模型在基因关联分析中的应用,为多基因、多位点的 SNP 数据提供了有效的分析工具,为群体遗传学研究提供适用有效的分析方法,也为复杂性状疾病的遗传以及基因定位等方面的研究提供方法支持。

在 SEMM 的模型评价中,研究者不可以太过于依赖某一指标,最好的方式是对估计的多种指标做综合性的判断。目前处理混合潜变量模型最主要的软件有 SAS 和 Mplus,本文使用 Mplus 软件进行编程实现,相比其他软件,Mplus 在处理结构方程混合模型时具有综合性强、程序简单的特点和优势。

参考文献(References)

[1] 李照海,覃红,张洪. 遗传学中的统计方法[M]. 北京:科学出版社,2006.

[2] 裴磊磊,张岩波,张克让. 抑郁症单核苷酸多态性(SNPS)分布特征的潜在类别分析[J]. 中国卫生统计,2010,27(2):7-10.

[3] 李锡钦. 结构方程模型:贝叶斯方法[M]. 北京:高等教育出版社,2011.

[4] McLachlan, G., Peel, D. Finite mixture models[M]. New York: Wiley,2000.

[5] 邱皓政. 潜在类别模型的原理与技术[M]. 北京:教育科学出版社,2008:57.