·基础理论与方法·

纵向数据潜变量增长曲线模型及其在Mplus中的实现

宋秋月　伍亚舟
400038 重庆，第三军医大学军事预防医学系卫生统计学教研室
通信作者：伍亚舟，Email：asiawu5@sina.com
DOI：10.3760/cma.j.issn.0254-6450.2017.08.027

【摘要】 探讨纵向数据潜变量增长曲线模型及其在Mplus中的实现方法。通过实例采用Mplus软件处理某高校大学生心理健康状况纵向数据。结果表明潜变量增长曲线模型可以处理含有潜变量的纵向数据，能够比较总体发展趋势和个体发展的差异，纳入协变量可以提高模型拟合效果；采用Mplus软件实现潜变量增长曲线模型，程序简单，操作方便。纵向数据潜变量增长曲线模型及其在Mplus中的实现程序，可为实际应用尤其是流行病学队列研究提供统计方法学方面的指导和参考。

【关键词】 纵向数据；潜变量增长曲线模型；Mplus软件
基金项目：国家自然科学基金（81573254）

The latent variable growth curve model of longitudinal data and its implementation in Mplus

Song Qiuyue, Wu Yazhou
Department of Health Statistics College of Preventive Medicine, Third Military Medical University, Chongqing 400038, China
Corresponding author: Wu Yazhou, Email: asiawu5@sina.com

【**Abstract**】 To discuss the latent variable growth curve model of longitudinal data and give its implementation method in Mplus. The application of Mplus software has been used to deal with the longitudinal data of mental health status of college students in an university. Results show that the model can process the longitudinal data with latent variables, which can compare the differences of the overall development trend and individual development, also taking a covariate into the model to improve the effect of model fitting. Using Mplus software to process the longitudinal data with latent variables, the program is simple and easy to operate. This study provides the latent variable growth curve model of longitudinal data and its procedure of implementation in Mplus, and the statistical methodology guidance and reference for practical applications of epidemiological cohort study.

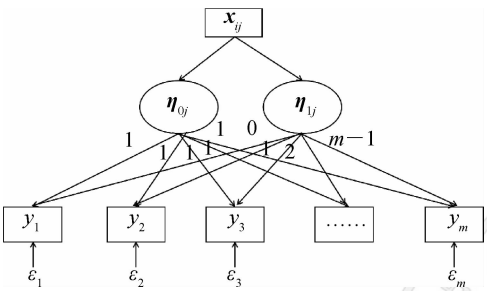
【**Key words**】 Longitudinal data; Latent variable growth curve model; Mplus
Fund program: National Natural Science Foundation of China (81573254)

随着精准医学和大数据概念的提出，流行病学大型队列研究的运用越来越广泛，同时产生了大量复杂的纵向数据（longitudinal data）。这些数据具有时间序列性、高维性、自相关性、多元性及易缺失等特性，且包含着复杂的横向和纵向关系。传统的纵向数据处理方法（如协方差分析、重复测量资料方差分析）只注重总体的发展趋势，即数据的纵向关系，难以对资料信息做出合理解释。潜变量增长曲线模型（latent variable growth curve model，LGCM）是在结构方程模型基础上发展演变而来，能够比较总体发展趋势和个体变化的差异，在含有潜变量的纵向数据中使用极为广泛[1-3]。目前Mplus软件分析处理LGCM十分普遍，操作方便，功能强大[4]。为此本文通过实例探讨纵向数据LGCM的分析方法，并提供含有潜变量的纵向数据在Mplus中的实现过程和程序。

基本原理

LGCM源于探索性因子分析及相关文献，是结构方程模型的一种变式。可以分析某一变量的变化趋势，用不可测量或难以测量的潜变量来描述总体的平均增长趋势，还可分析总体发展趋势和总体之间存在的差异，也可以分析个体之间的发展差异[5]。LGCM与含有均值的结构方程模型类似，是

将截距η_{0j}和斜率η_{1j}定义为潜在因素，以描述纵向数据的变化特征。如图1中y_1、y_2、y_3、y_m分别表示m次重复测量，ε_1、ε_2、ε_3、ε_m表示每次测量误差，LGCM中2个潜在因子是截距因子η_{0j}和斜率因子η_{1j}。截距因子表示观察对象测量的初始水平，描述当时间变量等于0时，结果变量y的水平，是常数项，不考虑协变量；斜率因子表示观察对象的增长轨迹，合适的载荷因子有利于模型参数的解释。截距η_{0j}到m次观测的载荷均定义为1，斜率η_{1j}的因子载荷称为时间分值，斜率η_{1j}到m次观测的载荷(a_t-a)定义为0，1，……，$m-1$，也可自由定义，以减少自由度。



注：x_{ij}为时间恒定的协变量，y_1、y_2、y_3、y_m分别表示m次重复测量，ε_1、ε_2、ε_3、ε_m表示每次测量误差，η_{0j}为截距因子，η_{1j}为斜率因子

图1 潜变量增长曲线模型

1. 简单非条件潜变量增长曲线模型：即不考虑任何协变量，其模型为

$$Y_{ij}=\eta_{0j}+t_{ij}\eta_{1j}+\varepsilon_{ij} \quad (1)$$

$$\eta_{0j}=\eta_0+\mu_{0j} \quad (2)$$

$$\eta_{1j}=\eta_1+\mu_{1j} \quad (3)$$

式中Y_{ij}表示第i个观察单位在第j个时间点的测量值；t_{ij}表示测量次数；η_{0j}表示截距即个体的初始状态；η_{1j}表示斜率即个体发展变化速度；ε_{ij}是观察单位个体内的结局变量测量的随机误差；η_0是截距的均值，表示所有个体在第一次观测时总体均值的估计；η_1表示斜率的平均值；μ_{0j}表示第一次观测时个体间的差异；μ_{1j}表示不同个体斜率变化的变异。非条件潜变量增长曲线模型与结构方程模型类似，可以用结构方程模型对其进行参数估计和评估方程拟合效果[6]。

2. 条件潜变量曲线增长模型：在非条件潜变量增长曲线模型的基础上纳入协变量，用以预测截距因子和斜率因子。协变量又分为时间恒定和时间变化两种情况。

（1）时间恒定协变量的LGCM：其模型为

$$Y_{ij}=\eta_{0j}+t_{ij}\eta_{1j}+\varepsilon_{ij} \quad (4)$$

$$\eta_{0j}=\eta_0+\gamma_{01}x_{ij}+\gamma_{02}x_{ij}+\gamma_{0m}x_{ij}+\mu_{0j} \quad (5)$$

$$\eta_{1j}=\eta_1+\gamma_{11}x_{ij}+\gamma_{12}x_{ij}+\gamma_{1m}x_{ij}+\mu_{1j} \quad (6)$$

式中x_{ij}表示时间恒定的协变量；γ_{01}表示协变量对截距的影响；γ_{11}表示协变量对斜率的影响，即协变量与时间的交互作用。

（2）时间变化的LGCM：其模型为

$$Y_{ij}=\eta_{0j}+t_{ij}\eta_{1j}+\sum_j\beta_jZ_{ij}+\varepsilon_{ij} \quad (7)$$

$$\eta_{0j}=\eta_0+\gamma_{01}x_{ij}+\gamma_{02}x_{ij}+\gamma_{0m}x_{ij}+\mu_{0j} \quad (8)$$

$$\eta_{1j}=\eta_1+\gamma_{11}x_{ij}+\gamma_{12}x_{ij}+\gamma_{1m}x_{ij}+\mu_{1j} \quad (9)$$

式中时间变化协变量Z_{ij}表示重复测量j次用于预测时间点上的结局变量Y_{ij}；β_j表示协变量对结局变量的影响大小。

3. Mplus软件实现：Mplus主要包含10个常用命令，即TITLE、DATA、VARIABLE、DEFINE、ANALYSIS、MODEL、OUTPUT、SAVEDATA、PLOT、MONTECARLO[7]。本文使用Mplus 7.0软件进行编程，非条件和条件潜变量增长曲线模型分析纵向数据的Mplus程序见表1。

表1 潜变量增长曲线模型Mplus程序

非条件潜变量增长曲线模型	条件潜变量增长曲线模型
TITLE: LATENT UNSPECIFIED CURVE MODEL DATA: FILE IS C:\Users\sqy\Desktop\SSSS.dat; VARIABLE: NAMES ARE gender V1 V2 V3 V4 ; USEVAR=V1–V4; MODEL: F1 BY V1–V4@1; F2 BY V1@0 V2@1 V3@2 V4@3; ［V1–V4@0 F1 F2］; V1–V4 F1 F2; F1 WITH F2; OUTPUT: SAMP STANDARDIZED MODINDICES	TITLE:LATENT UNSPECIFIED CURVE MODEL DATA:FILE IS C:\Users\sqy\Desktop\SSSS.dat; VARIABLE: NAMES ARE gender V1 V2 V3 V4; USEVAR=gender V1–V4 ; MODEL: F1 BY V1–V4@1; F2 BY V1@0 V2@1 V3@2 V4@3 ; F1 F2 ON gender ; ［V1–V4@0 F1 F2］; V1–V4 F1 F2; F1 WITH F2; OUTPUT: SAMP STANDARDIZED MODINDICES

注：TITLE为该程序的命名；DATA为指定要分析的数据集位置；VARIABLE为指定变量的名称以及分析在过程中使用的变量，本文分别列出了4次测量变量，即性别V1～V4、gender；MODEL为指令中BY语句表示潜变量因子截距和斜率有V1～V4四个可观测指标测量，且设置截距载荷因子为1，斜率载荷因子分别为0、1、2、3；［V1–V4@0 F1 F2］为估计变量的均值；V1–V4 F1 F2为估计变量的方差；F1 WITH F2表示截距和斜率的相关关系；OUTPUT即获得模型分析结果；STANDARDIZED表示提供标准化参数统计量和标准误；在纳入协变量方程中的ON语句定义2个潜变量因子对事件恒定协变量gender的线性回归

实例分析

1. 资料数据：该研究采用自制量表监测大学生心理健康状况，每隔1个月进行测量，该量表含有“感觉紧张不安”等5个条目以及“是否有自杀想法”1个特殊条目，每个条目有5个选项：0＝完全没有，1＝轻微，2＝中等程度，3＝厉害，4＝非常厉害，得分越高表示心理压力越大，需要给予适当干预。共观测了83名学生一学期4次的心理测评情况，其中男生29人，女生54人，结果见表2。

表2　本文实例大学生不同时间测量的心理评分

编号	性别	测量时间(月份)			
		9	10	11	12
1	1	6	6	5	3
2	0	7	8	8	7
3	0	6	6	10	11
4	1	7	6	9	7
5	1	5	5	5	5
⋮	⋮	⋮	⋮	⋮	⋮
80	0	1	4	9	9
81	0	9	16	7	8
82	0	0	0	0	0
83	1	3	0	4	5

注：性别：0＝男，1＝女

2. 结果分析：

（1）模型拟合情况：根据Mplus输出结果显示，无协变量情况下，χ^2＝8.743，P＞0.05，模型拟合良好，近似误差均方根RMSEA＝0.095＞0.08，拟合结果可接受，但不理想。比较拟合指数CFI＝0.974，TLI＝0.968，均＞0.95，标准化拟合残差SRMR＝0.069，表示拟合效果好。协变量的引入可以提高模型的拟合效果，模型χ^2＝11.763，P＞0.05，模型拟合效果良好，近似误差均方根RMSEA＝0.091，较前者更接近可接受界值。比较拟合指数CFI＝0.967，TLI＝0.953，也均＞0.95，标准化拟合残差SRMR＝0.069（＜0.08表示拟合效果好），结果见表3。

（2）非条件潜变量增长曲线模型结果分析：截距均值为4.475，斜率均值为0.048，学生心理评分随时间变化呈上升趋势。潜变量的方差估计结果显示，截距的方差为4.632，P＜0.05，斜率的方差为0.518，P＜0.05，差异有统计学意义，说明心理健康状况的初始水平和变化趋势存在个体差异。截距与斜率的协方差为–0.285，说明截距与斜率呈负相关关系，P＝0.461，差异无统计学意义，说明心理健康状况的初始水平与变化速度相关不显著（表4）。

决定系数是响应变量总变异中能被潜变量因子解释的比例，其值等于标准化因子负荷的平方，结果显示V3（11月）和V4（12月）测量的决定系数较大，且均P＜0.05，差异有统计学意义（表5）。

表3　模型拟合情况

指　标	无协变量	纳入协变量
Loglikelihood		
H0 Value	–759.395	–757.215
H1 Value	–755.024	–751.334
Information Criteria		
Akaike(AIC)	1 536.790	1 536.431
Bayesian (BIC)	1 558.559	1 563.038
Sample-Size Adjusted BIC ［n* ＝(n + 2) / 24］	1 530.171	1 528.341
RMSEA (Root Mean Square Error of Approximation)		
Estimate	0.095	0.091
90 Percent C.I.	0.000 ~ 0.197	0.000 ~ 0.178
Probability RMSEA ≤0.05	0.203	0.203
CFI/TLI		
CFI	0.974	0.967
TLI	0.968	0.953
SRMR (Standardized Root Mean Square Residual)		
Value	0.069	0.069

表4　本文实例大学生心理评分潜变量增长曲线模型参数估计结果

项目		β值	$s_{\bar{x}}$	t值	P值
截距	均数	4.475	0.304	14.716	0.000
	方差	4.632	1.295	3.577	0.000
斜率	均数	0.048	0.108	0.442	0.658
	方差	0.518	0.192	2.693	0.007
截距与斜率	协方差	–0.285	0.386	–0.737	0.461

注：P＜0.05为差异有统计学意义

表5　潜变量增长曲线模型拟合的决定系数

变量	R^2值	$s_{\bar{x}}$	t值	P值
V1	0.509	0.114	4.459	0.000
V2	0.450	0.069	6.500	0.000
V3	0.716	0.055	13.127	0.000
V4	0.928	0.075	12.404	0.000

注：V1为9月，V2为10月，V3为11月，V4为12月；P＜0.05为差异有统计学意义

（3）时间恒定潜变量增长曲线模型：在非条件潜变量增长曲线模型的基础上纳入性别协变量，探讨性别对大学生心理健康状况变化的影响。结果显示，η_0＝4.005，为大学生心理评分的初始水平均值，η_1＝0.151，为大学生心理评分平均变化趋势。$\sigma^2(\mu_{0j})$＝4.383，$\sigma^2(\mu_{1j})$＝0.511，均P＜0.05，说明心理健康状况的初始水平和变化趋势存在个体差异，γ_{01}＝1.308，P＜0.05，γ_{11}＝–0.283，P＞0.05，分别是性别对截距和斜率的影响，不同性别的学生的心理评分初始水平不同，变化趋势不受性别影响（表6）。

表6 协变量对潜变量的参数估计

参数	β值	$s_{\bar{x}}$	t值	P值
η_0	4.005	0.371	10.785	0.000
γ_{01}	1.308	0.623	2.099	0.036
η_1	0.151	0.134	1.127	0.260
γ_{11}	−0.283	0.224	−1.260	0.208
$\sigma^2(\mu_{0j})$	4.383	1.240	3.534	0.000
$\sigma^2(\mu_{1j})$	0.511	0.188	2.714	0.007
$\sigma^2(\mu_{0j}, \mu_{1j})$	−0.242	0.373	−0.648	0.517
$\sigma^2(\varepsilon_1)$	4.143	1.133	3.657	0.000
$\sigma^2(\varepsilon_2)$	5.719	1.033	5.537	0.000
$\sigma^2(\varepsilon_3)$	2.244	0.487	4.608	0.000
$\sigma^2(\varepsilon_4)$	0.587	0.619	0.949	0.343

注：η_0表示截距的均值，γ_{01}表示协变量对截距的影响，η_1表示斜率的平均值，γ_{11}表示协变量对斜率的影响，μ_{0j}表示第一次观测时个体间的差异，μ_{1j}表示不同个体斜率变化的变异，ε_1、ε_2、ε_3、ε_4表示每次测量误差；$P<0.05$为差异有统计学意义

讨　论

众多的纵向数据分析处理方法各有优缺，如重复测量方法分析注重总体的发展趋势，不能分析个体发展的差异，而且无法处理数据中存在缺失值的情况。多层线性模型能够较好地处理有缺失值的纵向数据，但是对于参数的估计其方法复杂，不能测量结局变量与潜变量的关系[8]。本文采用LGCM的两种类型分析大学生心理健康状况，结果显示初始水平和变化趋势存在个体差异，不同性别学生的心理评分初始水平不同，变化趋势不受性别影响。纳入协变量的条件潜变量增长曲线模型拟合效果更好，但参数较多，建模复杂，因此在后续研究中将探讨不同模型对纵向数据的分析处理及其效果评价。

本文实例中针对大学生心理健康状况提供了纵向数据LGCM的Mplus实现程序，其编程语言简单易学、操作方便，且软件更新速度快，计算方法丰富，能分析处理含有潜变量的纵向数据，为流行病学队列研究中统计学方法的应用提供参考。

利益冲突 无

参 考 文 献

[1] 刘红云，孟庆茂. 纵向数据分析方法[J]. 心理科学进展，2003，11(5)：586–592. DOI：10.3969/j.issn.1671–3710.2003.05.019.
Liu HY, Meng QM. A review on longitudinal data analysis method and it's development[J]. Adv Psychol Sci, 2003, 11(5): 586–592. DOI: 10.3969/j.issn.1671–3710.2003.05.019.

[2] 李丽霞，周舒冬，张敏，等. 多水平模型和潜变量增长曲线模型在纵向数据分析中的应用及比较[J]. 中华流行病学杂志，2014，34(6)：741–744. DOI：10.3760/cma.j.issn.0254–6450.2014.06.028.
Li LX, Zhou SD, Zhang M, et al. Comparisons of two statistical approaches in studying the longitudinal data: the multilevel model and the latent growth curve model[J]. Chin J Epidemiol, 2014, 34(6): 741–744. DOI: 10.3760/cma.j.issn.0254–6450.2014.06.028.

[3] Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: a review[J]. Stat Methods Med Res, 2014, 23(1): 42–59. DOI: 10.1177/0962280212445834.

[4] Gunzler DD, Morris N. A tutorial on structural equation modeling for analysis of overlapping symptoms in co-occurring conditions using Mplus[J]. Stat Med, 2015, 34(24): 3246–3280. DOI: 10.1002/sim.6541.

[5] 李丽霞，郜艳晖，张敏，等. 潜变量增长曲线模型及其应用[J]. 中国卫生统计，2012，29(5)：713–716.
Li LX, Gao YH, Zhang M, et al. Latent growth curves model and its application[J]. Chin J Health Stat, 2012, 29(5): 713–716.

[6] 王济川，王小倩，姜宝法. 结构方程模型：方法与应用[M]. 北京：高等教育出版社，2011.
Wang JC, Wang XQ, Jiang BF. Structural equation models: methods and applications[M]. Beijing: Higher Education Press, 2011.

[7] 裴磊磊，任琳，张岩波，等. Mplus软件简介[J]. 中国卫生统计，2013，30(4)：614–616.
Pei LL, Ren L, Zhang YB, et al. Mplus software profile[J]. Chin J Health Stat, 2013, 30(4): 614–616.

[8] 高彩虹. 基于广义估计方程和潜变量增长曲线模型的阿尔茨海默病健康相关生命质量动态变化研究[D]. 太原：山西医科大学，2012.
Gao CH. Dynamic study on health-related quality of life in the progression of Alzheimer's disease based on generalized estimating equations and latent growth curve model [D]. Taiyuan: Shanxi Medical University, 2012.

（收稿日期：2016–12–21）
（本文编辑：张林东）

中华流行病学杂志第七届编辑委员会通讯编委名单

（按姓氏汉语拼音排序）

陈　曦（湖南）	党少农（陕西）	窦丰满（四川）	高　婷（北京）	高立冬（湖南）	还锡萍（江苏）	贾曼红（云南）
金连梅（北京）	荆春霞（广东）	李　琦（河北）	李十月（湖北）	李秀央（浙江）	林　玫（广西）	林　鹏（广东）
刘　莉（四川）	刘　玮（北京）	刘爱忠（湖南）	马家奇（北京）	倪明健（新疆）	欧剑鸣（福建）	潘晓红（浙江）
彭晓旻（北京）	彭志行（江苏）	任泽舫（广东）	施国庆（北京）	汤奋扬（江苏）	田庆宝（河北）	王　丽（北京）
王　璐（北京）	王金桃（山西）	王丽敏（北京）	王志萍（山东）	武　鸣（江苏）	谢　娟（天津）	解恒革（海南）
严卫丽（上海）	阎丽静（北京）	么鸿雁（北京）	余运贤（浙江）	张宏伟（上海）	张茂俊（北京）	张卫东（河南）
郑　莹（上海）	郑素华（北京）	周脉耕（北京）	朱益民（浙江）	祖荣强（江苏）		