·基础理论与方法·

# 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

李丽霞 周舒冬 张敏 张岩波 郜艳晖

【导读】 比较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。文中以结直肠癌患者术后的生命质量情况为实例，比较两种方法的异同。结果表明两方法的参数估计值结果非常接近，多水平模型在模型构建时较为容易，而潜变量增长曲线模型在模型评价等方面具有优势。两方法均可很好地分析纵向观测的数据，且各有优点，研究者应根据需要选择合适的方法分析数据。

【关键词】 多水平模型；潜变量增长曲线模型；纵向数据

**Comparisons of two statistical approaches in studying the longitudinal data: the multilevel model and the latent growth curve model** *Li Lixia*[1], *Zhou Shudong*[1], *Zhang Min*[1], *Zhang Yanbo*[2], *Gao Yanhui*[1]. 1 *Department of Epidemiology and Biostatistics, School of Public Health, Guangdong Pharmaceutical University, Guangdong Key Laboratory of Molecular Epidemiology, Guangzhou* 510310, China; 2 *Department of Epidemiology and Biostatistics, School of Public Health, Shanxi Medical University*

*Corresponding author: Gao Yanhui*, Email: gao_yanhui@163.com

*This work was supported by a grant from the National Natural Science Foundation of China* (No. 30972553).

**【Introduction】** To compare two commonly used statistical approaches: the multilevel model and the latent growth curve model in analyzing longitudinal data. A longitudinal data set, obtained from the quality of life in patients with colorectal cancer after operation, was used to illustrate the similarities and differences between the two methods. Results from the study indicated that the latent growth curve modeling was equivalent to multilevel modeling with regards to longitudinal data which could yield identical results for the estimates of parameters. Multilevel model approach seemed easier for model specification. However, latent growth curve model had the advantage of providing model evaluation and was more flexible in statistical modeling by allowing the incorporation of latent variables. Both multilevel and latent growth curve models were suitable for analyzing longitudinal data with advantages on their own, they conld be chosen by researchers under different situation to be chosen accordingly by researchers under different situation.

**【Key words】** Multilevel model; Latent growth curve model; Longitudinal data

纵向数据经常应用于行为学、心理学、医学和教育学中，目前分析此类数据最常用的方法为重复测量的方差分析，但后者只能反映总体的增长趋势，不能反映个体增长轨迹的变化。而研究者往往最关心的是纵向数据增长曲线的初始状态和增长率。目前分析纵向数据既考虑到个体增长又考虑到总体增长趋势的常用方法包括多水平模型和潜变量增长曲线模型，两种方法均可分析增长曲线的初始状态和增长率的个体差异。多水平模型在国内许多研究领域已得到广泛应用[1,2]，潜变量增长曲线模型作为结构方程模型的一个新变体，逐渐引起关注[3,4]。本文拟通过实例比较两种方法在纵向数据分析中的应用。

## 基本原理

1. 多水平模型（multilevel model）：也称随机系数模型（random coefficient model）、分层线性模型（hierarchical linear model），是处理具有层次结构数据（hierarchical structure）的一种统计分析方法。研究中经常遇到层次结构数据，例如研究大学生学习倦怠的影响因素时，观察单位学生“嵌套”于班级水

DOI: 10.3760/cma.j.issn.0254-6450.2014.06.028
基金项目：国家自然科学基金（30972553）
作者单位：510310 广州，广东药学院公共卫生学院卫生统计学教研室 广东省分子流行病学重点实验室（李丽霞、周舒冬、张敏、郜艳晖）；山西医科大学公共卫生学院卫生统计学教研室（张岩波）
通信作者：郜艳晖，Email：gao_yanhui@163.com

平中，而班级又“嵌套”于学校中，这样可能具有3个水平的层次结构数据，即水平1：学生，水平2：班级，水平3：学校。相同班级的学生具有相似性，观测可能不独立，如果忽略了数据的这种层次结构可能导致有偏的参数估计值、较小的标准误、增大Ⅰ型错误，导致错误的分析结果。纵向资料中每个观测单位可被重复观测多次，因此同一观测不同时间点的测量值可视为水平1，观测个体为水平2，采用多水平模型分析纵向数据，建立线性增长模型[5]：

水平1模型：$y_{ij}=b_{0j}+t_{ij}b_{1j}+e_{ij}$　　(1)

式中$y_{ij}$为第$j$个观察单位的第$i$次测量值，$t_{ij}$为重复观测的时间，$b_{0j}$，$b_{1j}$分别表示截距（初始状态）和增长轨迹参数，$e_{ij}$为服从正态分布的测量误差，均数为0，方差为$\sigma_e^2$。

水平2模型：

对于截距项$b_{0j}$：$b_{0j}=\gamma_{00}+\gamma_{01}z_{1j}+\gamma_{02}z_{2j}+u_{0j}$　　(2)

对于增长轨迹$b_{1j}$：$b_{1j}=\gamma_{10}+\gamma_{11}z_{1j}+\gamma_{12}z_{2j}+u_{1j}$　　(3)

式中$\gamma_{00}$、$\gamma_{10}$分别表示$b_{0j}$、$b_{1j}$的平均值，$\gamma_{01}$、$\gamma_{11}$分别表示解释变量$z_1$对$b_0$、$b_1$的影响大小，$\gamma_{02}$、$\gamma_{12}$为解释变量$z_2$对$b_0$、$b_1$的影响大小，$u_{0j}$、$u_{1j}$的均数为0，方差分别为$\sigma^2(u_0)$、$\sigma^2(u_1)$，协方差为$\sigma(u_0,u_1)$。因此，模型中待估参数为$\gamma_{00}$、$\gamma_{10}$、$\gamma_{01}$、$\gamma_{11}$、$\gamma_{02}$、$\gamma_{12}$、$\sigma_e^2$、$\sigma^2(u_0)$、$\sigma^2(u_1)$和$\sigma(u_0,u_1)$。

如果增长曲线为非线性的，例如为二次项，水平1模型可以扩展为

$$y_{ij}=b_{0j}+t_{ij}b_{1j}+t_{ij}^2b_{2j}+e_{ij}$$

即水平2模型中增加：$b_{2j}=\gamma_{20}+\gamma_{21}z_{1j}+\gamma_{22}z_{2j}+u_{2j}$

2. 潜变量增长曲线模型（latent variable growth curve model，LGM）：是结构方程模型的变体，因此分析结构方程模型的软件，例如EQS、Lisrel等均可完成LGM的分析。Meredith和Tisak等学者提出的含2个潜在因子的增长曲线模型和多水平模型在概念上有相似之处。LGM线性增长模型一般形式为[5]

$$y_{ij}=\lambda_{0i}\eta_{0j}+\lambda_{1i}\eta_{1j}+\varepsilon_{ij}　　(4)$$

式中$y_{ij}$为第$j$个观察单位的第$i$次测量值，$\eta_{0j}$，$\eta_{1j}$分别为描述初始状态和线性增长轨迹的潜在因子，$\eta_{0j}$指当$t_i=0$时第$j$个个体的真实状态，为了模型的可识别和参数的可解释性，一般限定：$\lambda_{0i}=1$，$\lambda_{1i}=t_i$，$i$表示测量的时间点，一般选定第一个测量时点为起点，即$t_1=0$，有$t_2=1$，$t_3=2$，$t_4=3$，$t_i=i-1$，式(4)则为

$$y_{ij}=\eta_{0j}+\lambda_{1i}\eta_{1j}+\varepsilon_{ij}　　(5)$$

此时，式(5)形式上与式(1)相似，但式(5)中每次测量可以有独立的随机误差$\varepsilon_{ij}$。式(5)用矩阵的形式表示为

$$\begin{pmatrix}y_{1j}\\y_{2j}\\\vdots\\y_{ij}\end{pmatrix}=\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}+\begin{pmatrix}1&t_1\\1&t_2\\1&\vdots\\1&t_i\end{pmatrix}\begin{pmatrix}\eta_{0j}\\\eta_{1j}\end{pmatrix}+\begin{pmatrix}\varepsilon_{1j}\\\varepsilon_{2j}\\\vdots\\\varepsilon_{ij}\end{pmatrix}$$

$$Y=\tau_y+\Lambda_y\eta+\varepsilon　　(6)$$

式(6)中，$Y=\begin{pmatrix}y_{1j}\\y_{2j}\\\vdots\\y_{ij}\end{pmatrix}$，$\tau_y=\begin{pmatrix}0\\0\\0\\0\end{pmatrix}$，$\Lambda_y=\begin{pmatrix}1&t_1\\1&t_2\\1&\vdots\\1&t_i\end{pmatrix}$，$\eta=\begin{pmatrix}\eta_{0j}\\\eta_{1j}\end{pmatrix}$，$\varepsilon=\begin{pmatrix}\varepsilon_{1j}\\\varepsilon_{2j}\\\vdots\\\varepsilon_{ij}\end{pmatrix}$，式(6)为结构方程模型中的内生变量$Y$的度量模型。

描述初始状态和线性增长轨迹的潜在因子$\eta_{0j}$，$\eta_{1j}$可表达为

$$\eta_{0j}=\mu_0+\gamma_{01}\xi_{1j}+\gamma_{02}\xi_{2j}+\zeta_{0j}　　(7)$$

$$\eta_{1j}=\mu_1+\gamma_{11}\xi_{1j}+\gamma_{12}\xi_{2j}+\zeta_{1j}　　(8)$$

$$\eta=\alpha+B\eta+\Gamma\xi+\zeta　　(9)$$

其中，$\eta=\begin{pmatrix}\eta_{0j}\\\eta_{1j}\end{pmatrix}$，$\alpha=\begin{pmatrix}\mu_0\\\mu_1\end{pmatrix}$，$B=\begin{pmatrix}0&0\\0&0\end{pmatrix}$，$\Gamma=\begin{pmatrix}\gamma_{01}&\gamma_{02}\\\gamma_{11}&\gamma_{12}\end{pmatrix}$，$\zeta=\begin{pmatrix}\zeta_0\\\zeta_1\end{pmatrix}$，$\mu_0$，$\mu_1$为$\eta_{0j}$，$\eta_{1j}$的总体均数，$\xi_{1j}$和$\xi_{2j}$为解释变量，$\gamma_{01}$，$\gamma_{02}$和$\gamma_{11}$，$\gamma_{12}$，分别表示解释变量对两个潜在因子影响大小，$\zeta_{0j}$、$\zeta_{1j}$为各自的残差。式(9)为结构方程模型中的结构模型。

从式(1)～(3)和(5)、(7)、(8)的表达式可见，潜变量增长曲线模型和多水平模型有相同的表达式，式(1)中的$y_{ij}$，$b_{0j}$，$b_{1j}$，$t_{ij}$，$e_{ij}$与式(5)中的$y_{ij}$，$\eta_{0j}$，$\eta_{1j}$，$\lambda_{1i}$，$\varepsilon_{ij}$对应，多水平模型中式(2)、(3)的$\gamma_{00}$，$\gamma_{01}$，$\gamma_{02}$，$\gamma_{10}$，$\gamma_{11}$，$\gamma_{12}$，$z_{1j}$，$z_{2j}$，$u_{0j}$，$u_{1j}$与潜变量增长曲线模型式(7)、(8)中的$\mu_{00}$，$\gamma_{01}$，$\gamma_{02}$，$\mu_{10}$，$\gamma_{11}$，$\gamma_{12}$，$\xi_{1j}$，$\xi_{2j}$，$\zeta_{0j}$，$\zeta_{1j}$对应，不同的是在多水平模型中，假设每次测量有相同的测量误差，而在潜变量增长曲线模型中，可对每次测量误差分别估计，即对每次测量误差指定不同的测量误差方差。

与多水平模型相同，潜变量增长曲线模型也可以指定增长曲线为非线性，例如增长曲线为时间的二次项，此时式(4)可扩展为

$$y_{ij}=\lambda_{0i}\eta_{0j}+\lambda_{1i}\eta_{1j}+\lambda_{2i}\eta_{2j}+\varepsilon_{ij}$$

$$\eta_{2j}=\mu_2+\gamma_{21}\xi_{1j}+\gamma_{22}\xi_{2j}+\zeta_{2j}$$

式中$\eta_{2j}$是描述模型中非线性增长轨迹的潜在因子，一般固定参数$\lambda_{2i}$为一个已知的常数，$\lambda_{2i}=t_i^2$，$i$为测量

的时间点。

## 实例分析

某研究者采用自行设计的生命质量简表评价结直肠癌患者术后生命质量情况，每隔3个月进行测量，该量表包括“近1个月内身体状况”等6个条目，每个条目的选项为非常差=1，差=2，一般=3，好=4，非常好=5，得分越高表明生命质量越好，收集到病情相似患者92例，其中男性43例、女性49例，共进行4次随访，各次生命质量的平均得分分别为3.05±1.08、4.08±1.18、5.11±1.24、6.12±1.36。采用潜变量增长曲线模型、多水平模型分别分析该纵向数据，探讨性别对增长曲线的影响，比较两种方法参数估计的结果。

用Mlwin 2.0软件拟合两水平重复测量模型，其中患者为水平2单位，同一患者不同时间点的观测值为水平1单位，性别为解释变量，建模时考虑性别和时间的交互作用，采用迭代广义最小二乘法（iterative generalized least squares，IGLS）估计参数，参数估计值结果见表1。

**表1** 多水平模型与潜变量增长曲线模型的参数估计值结果比较

| 多水平模型 | | | 潜变量增长曲线模型 | | | | |
|---|---|---|---|---|---|---|---|
| 参数 | 估计值 | $s_{\bar{x}}$ | 参数 | 估计值 | $s_{\bar{x}}$ | $t$值 | $P$值 |
| $\gamma_{00}$ | 3.100 | 0.046 | $\mu_0$ | 3.101 | 0.046 | 67.535 | 0 |
| $\gamma_{01}$ | −0.088 | 0.066 | $\gamma_{01}$ | −0.086 | 0.066 | −1.309 | 0.191 |
| $\gamma_{10}$ | 1.023 | 0.014 | $\mu_1$ | 1.021 | 0.014 | 72.375 | 0 |
| $\gamma_{11}$ | −0.001 | 0.020 | $\gamma_{11}$ | −0.001 | 0.020 | −0.028 | 0.977 |
| $\sigma^2(u_0)$ | 0.894 | 0.048 | $\sigma^2(\zeta_0)$ | 0.860 | 0.049 | 17.471 | 0 |
| $\sigma^2(u_1)$ | 0.052 | 0.005 | $\sigma^2(\zeta_1)$ | 0.054 | 0.006 | 9.560 | 0 |
| $\sigma(u_0,u_1)$ | 0.047 | 0.011 | $\sigma(\zeta_0,\zeta_1)$ | 0.058 | 0.012 | 4.880 | 0 |
| $\sigma^2(e)$ | 0.257 | 0.008 | $\sigma^2(\varepsilon_1)$ | 0.315 | 0.026 | | |
| | | | $\sigma^2(\varepsilon_2)$ | 0.333 | 0.018 | | |
| | | | $\sigma^2(\varepsilon_3)$ | 0.196 | 0.014 | | |
| | | | $\sigma^2(\varepsilon_4)$ | 0.167 | 0.022 | | |

用Mplus 5.0软件拟合潜变量线性增长模型，图1为构建的潜变量线性增长模型，其中$Y_0$、$Y_1$、$Y_2$、$Y_3$是4次重复测量值，$\eta_0$、$\eta_1$分别是增长轨迹截距和斜率潜在因子，$\eta_0$的因子载荷固定为1，$\eta_1$的因子载荷固定为0、1、2和3，性别为协变量。采用最大似然估计进行参数估计：$\chi^2=15.468$，$P=0.030\,5$，拟合指数CFI=0.998，TLI=0.997，RMSEA=0.035，SRMR=0.015，提示模型拟合较好。参数估计结果见表1。

从表1可知两种分析方法的参数估计值非常接近。$\gamma_{00}=3.100$（$\mu_0=3.101$）为患者术后生命质量平均得分，$\sigma^2(u_0)=0.894$[$\sigma^2(\zeta_0)=0.860$]，$P<0.05$，表明患者术后初期的生命质量得分存在个体差异；$\gamma_{10}=1.023$（$\mu_1=1.021$）为随访时间内患者生命质量得分的平均变化率，表明随着时间的增长生命质量有提高的趋势，因为$\sigma^2(u_1)=0.052$[$\sigma^2(\zeta_1)=0.054$]，$P<0.05$，提示不同个体提高趋势存在个体差异；$\gamma_{01}=-0.086$为性别对截距因子（初始状态）的效应，$P=0.191$，$\gamma_{11}=-0.001$为性别对斜率因子的影响（即性别与时间的交互效应），$P=0.977$，提示性别对患者的术后生命质量的初期得分和变化率均无影响。

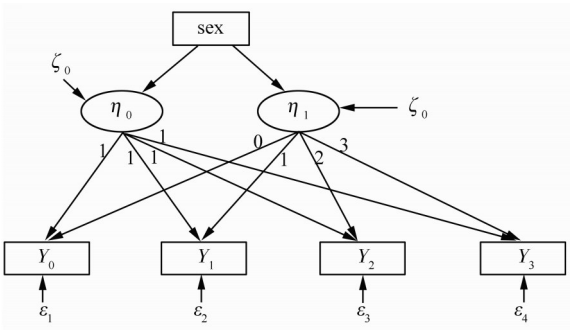


**图1** 结直肠癌患者治疗后生命质量情况的线性潜变量增长模型

## 讨　论

目前对纵向数据多采用重复测量的方差分析，但该方法不能很好地处理缺失值的问题，无法分析个体发展轨迹的差异。在纵向研究中，将同一个体不同时间点的观测值视为水平1单位，个体为水平2单位，对具有层次结构的数据可采用多水平模型分析。本研究通过实例比较多水平模型和潜变量增长曲线模型对分析纵向数据的参数估计值，结果显示两种分析方法的结果除测量误差外，其余参数估计值非常接近。多水平模型、潜变量增长曲线模型有各自不同的统计假设，起源于不同的理论，但在分析纵向数据时，基于结构方程模型的潜变量增长曲线模型与多水平模型相似。

不同软件对分析数据的格式要求不同，在采用Mplus软件构建潜变量增长曲线模型时，要求数据整理为“宽”的格式，即每个个体对应一条记录，不同的时间点用不同的变量表示，协方差矩阵可作为分析的原始数据；使用Mlwin软件进行多水平模型分析时，要求录入的数据整理为“长”的格式，即每个个体有多条记录，每个时间点对应一条记录，只能分析原始数据。在多水平模型中，时间作为一个取值已知的变量纳入模型，而在LGM中，可将因子载荷指

定为时间取值，达到与多水平模型相同的效果，也可以自由估计不同时间点的因子载荷，例如未设定因子载荷的潜变量增长曲线模型。

从表1可见，参数估计值略有差异，可能与参数估计方法有关，Mlwin软件拟合多水平模型时，采用IGLS，本文实例中增长曲线模型采用了最大似然估计法，当数据满足正态分布时，这两种估计方法是等价的。

多水平模型在模型构建方面较为直接、简单，分析2个水平以上的多水平嵌套数据时，比较容易实现，这是该模型的优势所在。潜变量增长曲线模型在模型构建方面较复杂，在参数估计时，由于待估参数个数增多，需考虑模型识别的问题，且有时收敛困难，但可以提供整个模型拟合优度情况，评价模型的适用性，对模型拟合效果进行评价时可参照$\chi^2$统计量、CFI、TLI、RMSEA等拟合指数，并根据软件提供的修正指数对模型进行修订。潜变量增长曲线模型运用灵活，可放松对测量误差相等的限制，自由估计每次测量的误差，在参数估计精准度方面较多水平模型更具有优势。

另外，作为结构方程模型的变体，LGM模型更大的优势是也可以分析变量间直接、间接的复杂因果关系。图2中潜在因子$\eta_{0j}$，$\eta_{1j}$可以作为自变量预测结果变量$Y$，而多水平模型在处理此类问题时则存在困难[6]。

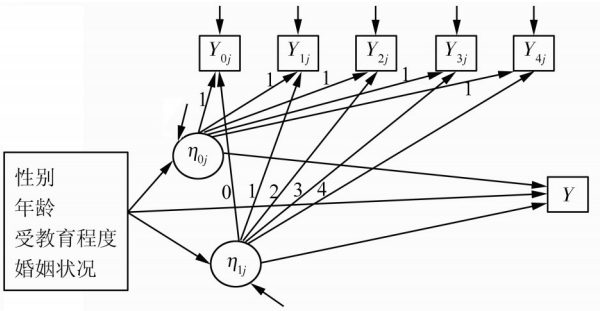


**图2**　含解释变量和结果变量的潜变量增长曲线模型

## 参 考 文 献


[1] Wang JC, Xie HY, Jiang BF, et al. Multilevel models: methods and applications [M]. Beijing: Higher Education Press, 2008: 81–110. (in Chinese)
王济川，谢海义，姜宝法，等. 多层统计分析模型—方法与应用[M]. 北京：高等教育出版社，2008：81–110.

[2] Yang M, Li XS. Multilevel models for medicine and public health research [M]. Beijing: Peking University Medical Press, 2007: 49–61. (in Chinese)
杨珉，李晓松. 医学和公共卫生研究常用多水平统计模型[M]. 北京：北京大学医学出版社，2007：49–61.

[3] Xu BY, Chen BW, Chen QG. The application of structural equation modeling in repeated measured data[J]. Mod Prev Med, 2007, 34(20): 3805–3807. (in Chinese)
许碧云，陈炳为，陈启光. 结构方程模型在重复测量数据中的应用[J]，现代预防医学，2007，34(20)：3805–3807.

[4] Liu HY, Meng QM. A review on longitudinal data analysis method and it's development[J]. Adv Psychol Sci, 2003, 11(5): 586–592. (in Chinese)
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.

[5] Chou CP, Bentler PM, Pentz MA. Comparisons of two statistical approaches to study growth curves: the multilevel model and the latent curve analysis[J]. Struct Equat Model, 1998, 5(3): 247–266.

[6] Wu XG. Latent growth curve modeling[M]. Shanghai: Ge Zhi Press, 2012: 109–124. (in Chinese)
吴晓刚. 潜变量增长曲线模型[M]. 上海：格致出版社，2012：109–124.


（收稿日期：2013–10–30）
（本文编辑：张林东）