多层线性模型在纵向研究中的运用

盖笑松[*1,2] 张向葵[2]

([1]中国科学院心理研究所心理健康重点实验室,北京,100101)

([2]东北师范大学教育科学学院心理学系,长春,130024)

摘 要 纵向研究中传统统计技术主要是重复测量的方差分析和多元回归分析,但是这两种技术存在一些局限性,不能合理而充分地解释纵向研究资料。近年来出现的多层线性模型能够更有效地利用纵向数据。为促进多层线性模型在纵向研究中的运用,简要论述了传统统计技术的局限,介绍了多层线性模型的原理及其在纵向研究中的作用,以一组模拟数据为例对多层线性模型中各种参数在纵向研究中的含义进行了详细讨论,对于纵向研究中运用多层线性模型时须注意的问题提出了建议。

关键词:多层线性模型 纵向研究 回归分析

1 前言

纵向研究也叫追踪研究,主要用来分析事物在一段时间内或某几个时间点上的变化趋势以及不同个体之间变化趋势的差异。与横断研究相比,纵向研究的最大优点是可以描述事物的连续性变化,以及合理地推论变量之间的因果关系。在医学、生物学、环境科学、心理学等许多领域中,纵向研究都有着广泛的应用。近20年来,国外统计学领域中专门的纵向数据分析技术有了长足的发展,其中,多层线性模型在纵向研究中得到了越来越多的运用[1]。

多层线性模型又称分层线性模型或多水平模型,当数据存在于不同层级时,先以第一层级的变量建立回归方程,然后把该方程中的截距和斜率作为因变量,使用第二层数据中的变量作为自变量,再建立两个新的方程。通过这种处理,可以探索不同层面变量对因变量的影响。由于它把第一层回归方程中的截距和斜率作为第二层回归方程中的随机变量,所以这种做法也被称作"回归的回归"。

多层线性模型这一术语最早是由Lindley和Smith于1972年提出,但是由于传统的参数估计方法(OLS)不适用于该模型,它受到了计算技术的限制而没有得到推广。80年代,迭代加权广义最小二乘法(iteratively reweighted generalized least squares)等手段开始被用于多层线性模型的参数估计,并相继出现了一些计算软件如HLM,Mlwin等。由于参数估计手段的成熟和计算软件的发展,近10年来,多层线性模型得到了越来越多的运用。

虽然国内已经有关于多层线性模型的介绍[2,3],但主要是讨论其在嵌套式组织模型中的应用,而关于纵向研究方面的应用还很少见。为了国内纵向研究中的统计手段的运用情况,以"纵向研究"或"追踪研究"为关键词,在CNKI期刊数据库近十一年来(1994 - 2004)的记录中进行篇名检索,可以发现252篇涉及纵向研究设计的文章。但是,进一步的分析却发现,这些研究报告中的统计技术基本上都局限于传统的数据分析手段,例如重复测量的方差分析和多元回归分析等。又尝试使用"分层线性模型"、"多层线性模型"、"多水平模型"等关键词对上述200多篇文献进行二次检索,只得到了一篇检索结果[4]。这说明多层线性模型在国内纵向研究中基本没有得到应用。本文拟对多层线性模型的原理及其在纵向研究中的运用做简要介绍,以一组模拟数据的分析过程为例,说明该模型中各种参数在纵向研究中的实际含义,并对运用中须注意的问题进行讨论。希望未来的纵向研究能充分利用该模型的优势,提高纵向研究的质量。

2 传统统计技术在纵向研究中的局限性及多层线性模型的优势

2.1 方差齐性和随机误差独立性假设的问题

传统统计技术在处理纵向研究数据时主要采用重复测量的方差分析和多元回归分析。然而,这两种统计手段都以方差齐性和随机误差独立性为前提假设。纵向研究中,这些假设很难得到保证。来自同一个体的多次追踪数据可能存在更多的相似性,在同一时间点的多个观测数据可能存在共同方向上的系统误差,这些问题都使随机误差独立性假设受到威胁。在关于发展趋势的纵向研究中,因变量随时间的推移而发生有规律的增减变化,方差也容易发生相应的增减,使方差齐性假设受到威胁。在上述情况下,采用传统统计技术可能导致不合理的、甚至是错误的结论。而多层线性模型不需要以方差齐性和随机误差独立性为前提假设,更适合于纵向研究。

2.2 缺失值问题和测量间隔不一致问题

纵向研究需要对同一观测对象做多次追踪观测,很容易出现样本的流失。传统的统计手段是删除存在缺失值的观测对象或对其缺失值进行拟合,前者造成了信息的浪费,而后者降低了研究的精确程度。多层线性模型允许缺失值的存在,在建立第一层回归模型时能够最大程度地利用现有样本的信息。另外,传统的统计技术还要求所有被观测对象在相同的时间间隔接受观测,这也给纵向研究增加了额外限制。多层线性模型不但允许不同时间间隔的测量,还允许不同观测对象采用不同的观测时间表,这一特性使研究者有了更大的便利和灵活性。

2.3 处理研究假设的能力。

传统统计手段的功能主要局限于比较各次观测结果间差异的显著性或前期观测结果对后期观测结果的预测程度,而多层线性模型允许研究者在不同数据层面上提出不同假设,例如是否存在着显著的增长或下降趋势?不同类型个体的变化速率是否一致?哪些因素可以预测不同类型个体在变化速率上的差异?不同类型个体在变化速率上的差异是否得到了足够的解释等等,还可以建立多个发展模型,通过拟合度检验选出最吻合观测数据的理论假设。

* 通讯作者:盖笑松,男。E - mail: gaixs669@nenu.edu.cn

3 多层线性模型的原理及其参数在纵向研究中的含义

在利用多层线性模型分析纵向数据时,以不同观测时间下的追踪结果为第一层数据,以不依时间变化的个体特征或所接受的处理为第二层数据,形成两层数据结构。

在第一层数据结构中,以追踪观测结果为因变量,以观测时间为自变量,可以建立第一层回归方程:

$$Y_{ij}=\beta_{0i}+\beta_{1i}X_{ij}+\varepsilon_{ij} \quad (1)$$

在方程(1)中,下标"0"表示截距,下标"1"表示斜率,下标"i"表示第i观测对象,下标"j"表示第j次观测时间。"β_{0i}"是方程的截距,其含义是第i个观测对象的平均数。"β_{1i}"是回归系数,其含义是第i个观测对象的变化速率。"X_{ij}"代表第i个观测对象在第j个观测时间中自变量X的取值,ε_{ij}代表残差,其含义是第i个观测对象在第j个观测时间中的测量值Y不能被自变量X所解释的部分。

方程(1)与一般的回归方程很相似,区别只在于,它的截距和斜率不是常数,而是随机变量,不同观测对象具有不同的截距和斜率,它们可能受到第二层变量的影响。

在第二层数据结构中,分别以方程(1)中的截距和斜率为因变量,以观测对象的个体特征或所接受的实验处理为自变量,可以建立两个第二层回归方程:

$$\beta_{0i}=\gamma_{00}+\gamma_{01}W_{1i}+\mu_{0i} \quad (2)$$

$$\beta_{1i}=\gamma_{10}+\gamma_{11}W_{1i}+\mu_{1i} \quad (3)$$

在上面的两个方程中,每个参数都有两个下标,其中第一个下标如果为0,则表示这是与方程(1)的截距有关的参数,第一个下标如果为1,则表示这是与方程(1)中的斜率有关的参数。第二个下标如果为0,表示该参数是第二层方程中的截距部分,第二个下标如果为1,表示该参数是第二层方程中的斜率部分。因此:

γ_{00}是方程(2)的截距,可以被理解为自变量W1为0时观测对象在因变量Y上的平均数。如果该系数的统计检验显著,意味着因变量的初始值不等于0,通常纵向研究中的研究者并不关心这一系数。

γ_{01}是方程(2)中自变量W1的回归系数,可以被理解为自变量W1对方程(1)中因变量Y初始值的影响大小。追踪研究中,如果对该系数的统计检验达到显著,意味着不同自变量水平的观测对象在观测起点上的初始值有显著差异。

μ_{0i}是方程(2)中的残差,可以被理解为方程(1)中因变量Y初始值未被自变量W1所解释的部分。如果对相应方差的统计检验达到显著,意味着模型中需要引入新的变量来解释因变量初始值上的变异。

γ_{10}是方程(3)的截距,可以被理解为自变量W1为0时观测对象的变化速率。在纵向研究中通常把它看作基础变化速率。如果对该系数的统计检验达到显著水平,则认为基础变化速率不等于0。

γ_{11}是方程(3)中自变量W1的回归系数,可以被理解为自变量W1对观测对象变化速率的影响大小。这一系数是纵向研究中最受关心的结果,可用于探索哪些因素能解释不同个体在变化速率上的差异。如果对该系数的统计检验达到显著性水平,意味着自变量W1是导致个体变化速率差异的重要原因。

μ_{1i}是方程(3)中的残差,可以被理解为方程(1)中因变量Y变化速率未被自变量W1所解释的部分。如果相应方差的统计检验达到显著,意味着模型中需要引入新的变量来解释变化速率上的变异。

为简化问题,上面的第二层回归方程中只包含了一个自变量W1,如果存在多个自变量,也可以继续写入W2、W3等,相应地,方程(2)和(3)的斜率部分也分别增加γ_{02}、γ_{03}和γ_{12}、γ_{13}等。

由于传统回归分析中的最小二乘法(OLS)要以方差齐性和随机误差独立性为前提假设,所以不适合多层线性模型的参数估计。在多层线性模型中采用的参数估计方法主要是迭代广义最小二乘法(IGLS)、限制性的广义最小二乘估计(RIGLS)和马尔科夫链蒙特卡罗法(MCMC)。由于Mlwin、HLM等专用统计软件的发展,参数估计过程都可以通过软件实现,这里不需要做更多介绍。

4 多层线性模型在纵向研究中的运用举例

4.1 问题说明

为了进一步理解模型中各部分参数在纵向研究中的实际意义,下面以一组模拟数据为例进行说明。为探索一种新药物对自闭症儿童的治疗作用,对120名自闭症儿童进行了研究,其中64名儿童未接受药物治疗,56名儿童接受了药物治疗。以"自闭症儿童适应功能测验"得分为观测指标对上述儿童进行了4次追踪,每次追踪间隔5个月,以反映自闭症儿童适应能力的改善情况。研究者期望了解药物治疗的作用以及治疗开始前的症状严重程度对疗效的影响。

建立两个SPSS格式的数据文件,文件1中包含儿童编号、测量时间、适应功能测验分数三个变量。文件2中包括儿童编号、治疗方法、治疗开始前的症状严重程度三个变量。利用HLM5软件读入两个SPSS格式的数据,并生成SSM数据结构。下面分别通过两个模型来分析纵向变化趋势。

4.2 随机效应模型

在随机效应模型中,第二层方程不包含任何自变量。所以该模型的作用是描述全体观测对象的变化趋势,并就是否需要进一步引入第二层解释变量做出决定。在随机效应模型中设定以下两层方程:

第一层的方程是:

$$\text{适应功能测验分数}=\beta_0+\beta_1(\text{观测时间})+\varepsilon \quad (5)$$

第二层的方程是:

$$\beta_0=\gamma_{00}+\mu_0 \quad (6)$$

$$\beta_1=\gamma_{10}+\mu_1 \quad (7)$$

在SPSS数据文件中,对于"测量时间"变量,分别用-3,-2,-1,0代表四次测量,这样做的目的是令方程(5)的截距β_0正好等于最后一次适应功能测验的平均分。

经过HLM5软件进行参数估计,结果见表1:

表1 随机效应模型的参数估计结果

固定效应部分	系数	标准误	T检验
γ_{00}	13.28	0.46	28.83***
γ_{10}	1.53	0.14	10.94***
随机效应部分	方差	自由度	卡方值
μ_0	20.09	119	700.48***
μ_1	1.08	119	232.10***

表1结果表示最近一次测量中所有儿童的适应功能平均分数是13.28。每隔5个月,自闭症儿童的适应功能分数平均增长1.53分,T检验结果说明增长趋势显著。方程(6)(7)中的残差变异都显著,说明无论是当前适应功能得分还是变化速率,都存在较大的个体间差异,需要引入第二层变量才能得

到更好的解释。

4.3 完整模型

为了更充分地解释方程(1)中当前分数和变化速率上的个体差异,引入两个第二层变量:是否接受了药物治疗以及治疗开始前的症状严重程度。这种在两层中都包含自变量的模型叫完整模型。

第一层的方程是:

$$适应功能测验分数 = \beta_0 + \beta_1(观测时间) + \varepsilon \quad (8)$$

第二层的方程是:

$$\beta_0 = \gamma_{00} + \gamma_{01}(是否治疗) + \gamma_{02}(严重程度 - 1) + \mu_0 \quad (9)$$

$$\beta_1 = \gamma_{10} + \gamma_{11}(是否治疗) + \gamma_{12}(严重程度 - 1) + \mu_1 \quad (10)$$

对于"治疗方法"变量,用0代表未接受治疗,1代表接受药物治疗。另外,考虑到"严重程度"变量的分值是1到7,不存在"严重程度"等于0的被试,为后面解释上的便利,在第二层方程中对"严重程度"变量进行了中心化。经过HLM5软件进行参数估计,结果见表2:

表2 完整模型的参数估计结果

固定效应部分	系数	*SD*	*T*
对于β_0			
截距γ_{00}	11.04	0.52	21.41***
治疗效应γ_{01}	4.65	0.81	5.77***
初始症状严重程度效应γ_{02}	-0.55	0.20	-2.76**
对于β_1			
截距γ_{10}	0.99	0.16	6.15***
治疗效应γ_{11}	1.08	0.26	4.14***
初始症状严重程度效应γ_{12}	-0.09	0.06	-1.39
随机效应部分	方差	*df*	χ^2
μ_0	14.28	117	510.48***
μ_1	0.78	117	198.51***

与β_0相关的固定效应提供了以下信息:未接受药物治疗的最低严重程度的自闭症儿童当前适应功能得分平均是11.04分;药物治疗对当前得分有显著影响,治疗组得分比未治疗组高出4.65分;初始症状严重程度对被试当前得分有显著影响,初始症状严重程度每升高1级会导致当前得分下降0.55分。

与β_1相关的固定效应提供了以下信息:未接受药物治疗的最低严重程度的自闭症儿童适应功能变化趋势是每5个月升高0.99分;药物治疗对适应功能增长速度有显著的积极影响,可以使增速提高0.16分;初始症状严重程度对适应功能增长速度无显著影响。

随机效应的分析表明,由于完整模型中增加了两个自变量,截距的方差从随机效应模型中的20.09下降到现在的14.28,说明两个新增自变量解释掉随机效应模型截距29%的方差,斜率的方差从随机效应模型中的1.08下降到现在的0.78,说明两个新增自变量解释掉随机效应模型斜率28%的方差。从随机效应的显著性检验结果来看,剩余的未被解释的方差仍然显著,还需要考虑引入新的自变量以增加模型的解释力度。

5 纵向研究中运用多层线性模型时需要注意的一些问题

运用多层线性模型分析纵向研究资料时,需要注意以下三方面的问题:

5.1 至少需要三波次以上的追踪数据。有些纵向研究中只包括了两波次追踪数据,在这种情况下,有些水平1的问题无法得到确定,例如所有的变化都在两次测量之间发生了吗?变化趋势是稳定的还是波动的?两次测量之间的差异是源自变化还是源自测量误差?因此,无法根据两波次数据做出不同观测对象的发展曲线,而这正是水平1分析中的核心部分。一般认为[1],测量的波次越多,多层线性分析的结果也就越可靠,而最低要求是有三波次以上的追踪数据。

5.2 应根据研究需要合理选择时间变量。时间变量是水平1分析中的主要自变量,它既影响到研究成本和研究设计,也影响到结果的解释。时间变量的选择没有固定的要求,例如,在疗效研究中既可以选取周数也可以选取治疗次数为时间变量;在教育研究中既可以选取年级也可以选取年龄为时间变量。研究者应根据研究需要,设定最合适的时间变量。各次观测所间隔的时间可以是不相等的(unequally spaced waves),在急速变化期可以有较多的观测次数,而缓慢变化期则可以减少观测次数。如果不同被试采用相同的观测时间表,这种情况被称为结构化时间数据(time - structured),如果不同被试采用不同的观测时间表,这种情况被称为非结构化时间数据(time - unstructured),两种情形都可适用于多层线性模型。

5.3 水平1中的结果变量必须在各次测量中具有等价性。结果变量是指水平1上的因变量,它在追踪过程中得到了多次测量。如果各次测量都使用了相同的测量工具,结果变量的等价性基本可以得到保证,即各次测量结果是可比的。但是,有时候在前次测量中的工具无法适合后次测量的需要,例如考察儿童在一学期内阅读能力的变化,每次都需要采用不同的阅读测验。这种情况下,不可以简单地根据标准分进行转化,而需要进行测验等价处理。

6 参考文献

1 Singer J D, Willett J B. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence, New York: Oxford University Press, 2003

2 张雷,雷雳,郭伯良. 多层线性模型应用,北京:教育科学出版社,2003

3 刘红云,孟庆茂. 教育和心理研究中的多层线性模型. 心理科学进展,2002,10(2):213 - 219

4 吴康敏,杨速飞,罗红裔等. 足月小样儿生长发育纵向研究. 华西医科大学学报,1998,29(3):310 - 314

The Application Of Multilevel Model In Longitudinal Research

Gai Xiaosong[1,2], *Zhang Xiangkui*[2]

([1] Key Lab of Mental Health, Chinese Academic of Sciences, Beijing, 100101)

([2] Department of Psychology, Northeast China Normal University, Changchun, 130024)

Abstract Some limits exist in traditional longitudinal analysis technique such as repeated ANOVA and multivariate regression. The emergence of multilevel analysis improved the power in longitudinal data analysis. The limits of traditional methods and the advantage of multilevel analysis were introduced. Meanings of parameters of multilevel model in longitudinal research were discussed and an example data were used to illuminate these meanings. Some suggestions about future application of multilevel model in longitudinal research were put forward.

Key words: multilevel model, longitudinal research, regression analysis