·综述·

多水平模型在流行病学研究中的应用进展

邢健男　钱莎莎　王璐

【关键词】 多水平模型；流行病学研究

Application and progress of multilevel models in epidemiological research

XING Jian-nan, QIAN Sha-sha, WANG Lu. National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China

Corresponding author: *WANG Lu*, Email: wanglu64@163.com

【Key words】 Multilevel models; Epidemiological research

流行病学研究是通过目标人群收集、整理和分析数据，获得真实结果以分析和提供防治疾病及促进健康的策略。其中的关键在于数据质量、方法正确以及偏倚的大小。流病学家在研究时往往得到具有层次结构特征（hierarchical structure）的数据[1]，也称多水平数据，即若干单位聚集在不同水平的数据。其特征为同一群体内的个体由于社会背景、生活习俗等相同而具有相似性，即反应变量在个体内的分布不具有独立性，不同群体间可能差异较大，因此在特定空间范围内有聚集性（clustering）。例如调查中学生吸烟行为的研究中，学生嵌套于班级，此时学生为一水平，班级为二水平，反应变量可能由于在学生水平间不独立从而发生在班级水平上聚集的现象。一般而言，传统回归模型假设的前提条件完全不适用多水平数据。模型的误差来源应与数据的结构本身相对应，即一个结构水平对应一个误差来源[2]。当用传统回归模型错误拟合多水平数据时，由于拟合后的残差未考虑分层，影响了模型中各参数的有效性和统计特性，使统计结论出现偏倚[3]。多水平模型可将传统模型拟合后的残差按照数据再分层，减少了不能解释的残差比例，极大地改善拟合效果。因此在流行病学研究中如何正确应用多水平模型已成迫切需要。

1. 多水平模型的基本形式：以零模型和二水平随机系数模型为例，简单介绍多水平模型的基本形式[2]。即

$$y_{ij}=\beta_{0j}+e_{0ij} \qquad \beta_{0j}=\beta_0+u_{0j}$$
$$u_{0j}\sim N(0,\sigma^2_{u0}) \quad e_{0ij}\sim N(0,\sigma^2_{e0}) \qquad (1)$$

式(1)中，i为水平一单位，j为水平二单位，y_{ij}为第j层的第i个研究对象的反应变量观测值，此模型中除截距和随机误差外不含有任何解释变量，称为零模型（zero model）。该模型往往用来描述反应变量是否具有聚集性。此时数据间的相关程度可用组内相关系（intra-class correlation, ICC）ρ表示：$\rho=\sigma^2_{u0}/\sigma^2_{u0}+\sigma^2_{e0}$。

当确定存在聚集性且σ^2_{u0}的值存在统计学意义时，可将感兴趣的解释变量纳入方程，得到二水平随机系数模型：

$$y_{ij}=\beta_{0j}+\beta_{1j}x_{ij}+e_{0ij} \qquad \beta_{0j}=\beta_0+u_{0j} \quad \beta_{1j}=\beta_1+u_{1j} \qquad (2)$$

式(2)中，x_{ij}为第j层的第i个研究对象的解释变量观测值，β_{0j}为截距，β_{1j}为解释变量x的回归系数，e_{0ij}为通常的随机误差项。在β_{0j}和β_{1j}中，β_0与β_1为其固定成分，在第二层间是不变的，而u_{0j}与u_{1j}为其随机成分，代表第二层之间的差异。其期望值、方差和协方差表示为：

$$E(u_{0j})=E(u_{1j})=0 \qquad (3)$$
$$\mathrm{Var}(u_{0j})=\sigma^2_{u0},\ \mathrm{Var}(u_{1j})=\sigma^2_{u1},\ \mathrm{cov}=(u_{0j},u_{1j})=\sigma_{u01} \qquad (4)$$

多水平模型不仅可以纳入一水平单位的解释变量，也可将更高水平单位的解释变量纳入模型。同样以二水平随机系数模型为例，假定x_{ij}为一水平解释变量，w_j为二水平解释变量，则方程[4]：

$$y_{ij}=\beta_{0j}+\beta_1x_{ij}+e_{0ij} \qquad (5)$$
$$\beta_{0j}=\beta_0+\beta_{01}w_j+u_{0j},\ \beta_{1j}=\beta_1+\beta_{11}w_j+u_{1j} \qquad (6)$$

将式(6)代入式(5)可得：

$$Y_{ij}=\beta_{0j}+\beta_1x_{ij}+\beta_{01}w_j+\beta_{11}w_jx_{ij}+u_{0j}+u_1x_{ij}+e_{0ij} \qquad (7)$$

式(7)的特点在于项$\beta_{11}w_jx_{ij}$可以提示两层解释变量间的交互作用。

2. 多水平模型在流行病学研究中的应用：该模型在国外流行病学研究领域中开展的较早，近年来流行病学家将更多的目光投向宏观项目评价以及通过拟合模型对反应变量的预测。Masangwi等[5]对马拉维家庭中痢疾防治知识知晓率状况进行多水平分析，发现母亲对痢疾防治知识的知晓率在社区水平上有很强的聚集性。Yoshihisa等[6]利用日本包含45个区域的随访研究数据拟合模型，提示男性在地区范围内其社会经济的压力影响缺血性心脏病的发生率，并建议除了考虑个体自身的压力外还应考虑该地区范围内的社会经济压力对缺血心脏病发生率的影响。Clark等[7]收集了2002—2004年美国全国人院病例数据，拟合了logistic多水平随机系数模型和logistic模型，并应用模型对2004—2006年的风险调整死亡率进行预测，结果提示相对于logistic模型，logistic多水平随机系数模型预测的风险调整死亡率与实际数据差异较小。另外，许多流行病学家应用多水平模型对医疗服务系统进行了评价。如Arling等[8]抽取明尼苏达州393所老年疗养院进行系统研究和评价，将分层分析、logistic回归分析与多水平模型分析进行比较，显示应用多水平模型可得出更加准确的结果，并在多水平模型的基础上应用Bayes经验估计对12项疗养院评分的标准误及其置信区间进行更准确的计算。López-Cevallos和Chi[9]对厄瓜多尔的医疗资

DOI: 10.3760/cma.j.issn.0254-6450.2013.01.022

作者单位：102206 北京，中国疾病预防控制中心性病艾滋病预防控制中心

通信作者：王璐，Email: wanglu64@163.com

源供应水平进行分析，认为在“私人与国立”、“城市与乡村”不同水平上存在聚集性且影响显著。Grieve等[10]针对不同国家间医疗资源的使用和费用问题，运用最小二乘法回归模型和多水平模型进行比较分析，结果提示最小二乘法回归模型在评价国家间医疗资源使用及其费用时易产生错误推断，而应用多水平模型则可准确解释不同国家间医疗资源使用差异的原因。

多水平模型在国内已应用于不同的研究领域。叶小华等[11]在广东省高中生健康素养影响因素多元多水平分析中发现，高中生理念性素养、行为性素养、技能性素养的平均分存在一定相关性，且班级水平相关性明显高于个体水平相关性。曹静等[12]应用多水平模型建立了不同年龄、孕前体重指数的孕妇孕期体重指数增长模式，为妊娠期保健提供依据。李佳萌和王伟[13]发现大学生吸烟的一些影响因素在班级水平有统计学意义，可针对不同班级内学生的特点进行指导。在有关社会学等宏观领域的研究方面，周海滨等[14]运用拟合个体和城区两个水平的多水平等级logistic模型分析慢性病患者社区就诊的影响因素，得出患者就医选择存在地区聚集性，主要受医疗保障制度和社区卫生服务能力的宏观因素影响。王书梅等[15]在居家伤害的影响因素研究中，得出不同社区居家安全评分存在差异，数据存在以社区为二水平的层次结构。在有关心理健康的流行病学研究中，彭丹等[16]对汶川大地震后中、小学生创伤后应激障碍的影响因素进行分析，显示学校水平存在聚集性，性别、年龄、是否有亲人朋友遇难、是否看到遗体和残肢4个因素对症状得分的影响在学校水平上有差异。黄薇等[17]对失地农民心理健康状况及其影响因素分别使用了多水平logistic模型与传统logistic模型，并对两种模型进行分析与比较，发现多水平模型在自变量的选择与解释上更加合理。

3. 多水平模型在流行病研究应用中的优缺点：该模型在不同水平上将变异分解从而最大程度的利用信息并分析不同水平间的交互作用，除此之外还具有其他优势。例如在一项分析中，当每个二水平下样本数量极端缺乏（二水平下样本量<2）时，多水平模型与传统回归模型均无法得到准确真实的参数估计，但是当二水平下样本量>5时，不论反应变量为离散变量还是连续变量，多水平模型对总体参数的估计较传统模型更为准确[18]。多水平模型在个体生长变化的研究中也具有一些优势，如对个体不必在相同时间或地点进行观察、随时间变化的变量可以拟合人方程、不同个体可以有不同的测量次数等[19]。多水平模型不仅可以用于单变量的分析，还可用于多个应变量的分析。国外研究资料表明，多变量多水平模型主要应用于变量的相关性研究，根据模型的原理和方法，利用HLM软件或SAS软件，不仅可以得出固定效应和随机效应的估计值，还可以得到多个反应变量间随时间变化的关系。另外，多水平模型在宏观流行病研究中，可在二水平或更高水平上进行变量分析，得出更加有效和针对性的宏观决策。

但多水平模型也存在缺陷。如有可能发生一水平解释变量与二水平解释变量间的关联，从而低估或高估某些因素对反应变量的影响[4]。此外，二水平下包含的样本量太小易导致模型结果不稳定、计算复杂，影响了其使用。特别值得注意的是，由于对多水平模型的盲目过度且错误使用，从而产生更大的偏倚。

4. 多水平模型可能出现的偏倚及校正：该模型在拟合流行病学数据时不可避免出现偏倚，但可在某些程度上减少偏倚带来的不良影响。例如在研究某种肝病的影响因素时，个体一水平的影响因素酒精消耗量与二水平该地区的酒吧数量往往存在关联，很难分辨出这两个因素的真实效果，一水平或二水平间往往也存在关联，由此带来混杂偏倚，此时可以分析其交互作用或者利用分层分析更好模拟方程。再如，若分层资料计数取值中含有大量的零，即零过多现象时，多水平零膨胀Poisson和多水平零膨胀负二项是最佳模型。计数数据的额外零产生于两个不同的过程，第一个过程由生成零计数的二项分布支配，第二过程由常规计数分布支配。对应于这两个过程，多水平零膨胀模型由两个相应部分组成，可很好处理和解释由于零过多而导致的参数有偏估计问题。

所以在流行病学研究中使用多水平模型时应注意所使用的抽样方法、数据类型以及变量选择等，并尽可能避免不必要的偏倚，而正确有效地使用多水平模型，则可得到真实可靠的结果。

参 考 文 献

[1] Goldstein H. Multilevel Statistical Model. Version 2. (多水平统计模型). 李晓松译. 2版. 成都：四川科学技术出版社，1999：2-48.

[2] Yang M, Li XS. Multilevel statistical model in medical and public health research. Beijing: Peking University Medical Press, 2007: 6-9. (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. 北京：北京大学医学出版社，2007：6-9.

[3] Rasbash J, Browne W, Goldstein H, et al. A user's guide to MLwiN. Centre for Multilevel Modelling, Institute of Education, University of London, 2000: 7.

[4] Bingenheimer JB, Raudenbush SW. Statistical and substantive inferences in public health: issues in the application of multilevel models. Annu Rev Public Health, 2004, 25: 53-77.

[5] Masangwi SJ, Grimason AM, Morse TD, et al. Pattern of maternal knowledge and its implications for diarrhoea control in Southern Malawi: multilevel thresholds of change analysis. Int J Environ Res Public Health, 2012, 9(3): 955-969.

[6] Yoshihisa F, Naohito T, Kaori H, et al. A prospective cohort study of neighborhood stress and ischemic heart disease in Japan: a multilevel analysis using the JACC study data. BMC Public Health, 2011, 11(1): 398.

[7] Clark DE, Hannan EL, Wu C, et al. Predicting risk-adjusted mortality for trauma patients: logistic versus multilevel logistic models. J Am Coll Surg, 2010, 211(2): 224-231.

[8] Arling G, Lewis T, Kane RL, et al. Improving quality assessment through multilevel modeling: the case of nursing home compare.

Health Serv Res, 2007, 42(3 Pt 1): 1177-1199.

[9] López-Cevallos DF, Chi C. Assessing the context of health care utilization in Ecuador: a spatial and multilevel analysis. BMC Health Serv Res, 2010, 10(1): 64.

[10] Grieve R, Nixon R, Thompson SG, et al. Using multilevel models for assessing the variability of multinational resource use and cost data. Health Eco, 2005, 14(2): 185-196.

[11] Ye XH, Xu Y, Zhou SD, et al. Multivariate and multilevel model analysis on factors that influencing the literacy of health among high school students in Guangdong province. Chin J Epidemiol, 2011, 32(9): 873-876. (in Chinese)
叶小华, 许雅, 周舒冬, 等. 广东省高中生健康素养影响因素多元多水平分析. 中华流行病学杂志, 2011, 32(9): 873-876.

[12] Cao J, Hu XY, Liu XH, et al. The application of multilevel model in the increasing pattern of body mass index during pregnancy. Chin J Health Stat, 2011, 28(5): 485-487. (in Chinese)
曹静, 胡晓吟, 刘兴会, 等. 多水平模型在妊娠期体重指数增长模式中的应用. 中国卫生统计, 2011, 28(5): 485-487.

[13] Li JM, Wang W. Multilevel model analysis on the influence of smoking among university students in Tianjin. Chin J Epidemiol, 2006, 27(6): 494-498. (in Chinese)
李佳萌, 王伟. 天津市大学生吸烟影响因素多水平模型分析. 中华流行病学杂志, 2006, 27(6): 494-498.

[14] Zhou HB, Ye CG, Zhu B, et al. Multilevel model analysis on health seeking behavior of patients with chronic Diseases in Shenzhen city. Chin J Soc Med, 2011, 28(4): 249-251. (in Chinese)
周海滨, 叶承刚, 朱斌, 等. 深圳市慢性病患者就医选择影响因素的多水平模型分析. 中国社会医学杂志, 2011, 28(4): 249-251.

[15] Wang SM, Zou JL, Xu WY, et al. Elative factors in home safety evaluated by multilevel statistical models. Chin J Epidemiol, 2010, 31(9): 975-978. (in Chinese)
王书梅, 邹金良, 徐文燕, 等. 居家伤害相关因素的多水平模型分析. 中华流行病学杂志, 2010, 31(9): 975-978.

[16] Peng D, Deng H, Zhang Q, et al. A multilevel model analysis on the influencing factors of post-traumatic stress disorder in primary and secondary school students exposed to the Wenchuan earthquake. Mod Prev Med, 2011, 38(17): 3411-3414. (in Chinese)
彭丹, 邓红, 张强, 等. 汶川大地震中、小学生创伤后应激障碍影响因素的多水平模型分析. 现代预防医学, 2011, 38(17): 3411-3414.

[17] Huang W, Huang SP, Zhuo L, et al. Evaluation of mental health status in farmers losing farmland with different models. Chin J Public Health, 2010, 26(5): 563-565. (in Chinese)
黄薇, 黄水平, 卓朗, 等. 失地农民心理健康状况不同模型评价比较. 中国公共卫生, 2010, 26(5): 563-565.

[18] Clarke P. When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. J Epidemiol Community Health, 2008, 62(8): 752-758.

[19] Jin F, Ni ZZ, Li XS, et al. Multivariate response model with multilevel and its application in the influencing factors of blood pressure. Chin J Health Stat, 2004, 21(4): 204-206. (in Chinese)
金芳, 倪宗瓒, 李晓松, 等. 多元多水平模型及其在儿童生长发育研究中的应用. 中国卫生统计, 2004, 21(4): 204-206.

(收稿日期: 2012-08-17)
(本文编辑: 张林东)

读者·作者·编者

关于中华医学会系列杂志投稿网址的声明

为维护广大读者和作者的权益以及中华医学会系列杂志的声誉，防止非法网站假冒我方网站诱导作者投稿、并通过骗取相关费用非法获利，现将中华医学系列杂志稿件管理系统网址公布如下，请广大作者加以甄别。

1. “稿件远程管理系统”网址：中华医学会网站(http://www.cma.org.cn)首页的“业务中心”栏目、中华医学会杂志社网站(http://www.medline.org.cn)首页的“稿件远程管理系统”以及各中华医学会系列杂志官方网站接受投稿。作者可随时查阅到稿件处理情况。

2. 编辑部信息获取：登录中华医学会杂志社网站(http://www.medline.org.cn)首页，在《中华医学会系列杂志一览表》中可查阅系列杂志名称、编辑部地址、联系电话等信息。

3. 费用支付：中华医学会系列杂志视杂志具体情况，按照有关规定，酌情收取稿件处理费和版面费。稿件处理费作者在投稿时支付；版面费为该稿件通过专家审稿并决定刊用后才收取。

欢迎投稿，并与编辑部联系。特此声明。

中华医学会杂志社