·基础理论与方法·

多水平模型在吸烟行为研究中的应用

李佳萌

【摘要】 **目的**　介绍多水平模型及其应用领域。**方法**　以中国/WHO 控烟能力建设合作项目——学校控烟子项目中天津地区中学的学生基线调查资料为例，应用多水平模型分析并与传统 logistic 回归分析的结果进行比较。**结果**　经检验，数据中存在层次结构。利用多水平模型分析显示，中学生吸烟的影响因素为性别、年龄、态度、环境及班级水平下的宣传教育。此外，在未引入班级水平下宣传教育这一变量时，利用多水平模型分析得到变量的标准误均小于相应的传统 logistic 回归分析的结果。**结论**　多水平模型适于分析具有层次结构的数据资料，在分层或整群的流行病学或社区调查中具有较高的应用价值。

【关键词】 吸烟；多水平模型

An applied multilevel model used in the study on behavior of smoking　*LI Jia-meng. Tianjin Center for Disease Control and Prevention, Tianjin* 300011, China

【**Abstract**】 **Objective**　To introduce the nature and its application of a multilevel model. **Methods**　Data was analyzed from a baseline survey of smoking behavior among middle school students sponsored by a WHO smoking controll project. Multilevel analysis was used on available data and to compare the results from logistic regression. **Results**　The outcomes of null multilevel model approved that there was hierarchical structure on data. The influencing factors of middle school students smoking appeared to be gender, age, attitude, environment and public education at schools. When the variable of public education by classes was not included, the standard errors by multilevel analysis became smaller than the corresponding standard errors through logistic regression method. **Conclusion**　Multilevel model seemed a good method for analyzing data with hierarchical or cluster structure, it could be applied in stratified or cluster sampling of epidemiological or community-based investigation.

【**Key words**】 Smoking; Multilevel model

目前在对吸烟行为的研究中多采用多元线性回归或 logistic 回归分析，但研究者研究吸烟行为时多以学校、单位、社区等群体为基础进行调查询问，由于吸烟者嵌套在班级、学校、单位中，所以这种调查方式不可避免的产生了组效应或者说背景效应。即资料产生了两层结构，第一层为个体也称水平 1 单位，第二层为班级（学校、单位）也称水平 2 单位。若研究者在关注个体效应时忽视组效应或环境效应，结果在个体这一层数据上得到的相关系数可能是错误的，因为具有相似背景的同一组内的个体之间，比该组外的个体而言，其相似性更高；另一结果是，Ⅰ类错误被放大，因为所观察到的效应既包含了个体效应，也包含了组效应[1]。多水平模型（multilevel model）是处理这类问题的统计分析方法。本文介绍多水平模型及其在吸烟行为研究中的应用。

作者单位：300011 天津市疾病预防控制中心

基本原理

多水平统计模型是英、美等发达国家教育学界 20 世纪 80 年代中后期发展起来的一门多元统计分析新技术。根据研究者的研究目的和资料的情况，多水平模型可有不同的形式，如二分类离散数据多水平模型、多分类离散数据多水平模型、重复测量数据多水平模型、多水平交叉分类模型、双变量多水平模型、非线性多水平模型、多水平时间序列模型等。但其基本的形式包括 3 个公式[1]：

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (1)$$

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j} \quad (3)$$

上式中，下标 j 代表的是第一层的个体所属的第二层的单位，如学校、班级；γ_{00} 和 γ_{10} 分别是 β_{0j} 和 β_{1j} 的平均值，并且它们在第二层的单位之间是恒定的，是 β_{0j} 和 β_{1j} 的固定成分；μ_{0j} 和 μ_{1j} 分别是 β_{0j} 和 β_{1j} 的

随机成分，它们代表第二层单位之间的变异。

方差和协方差为：

$$\mathrm{Var}(\mu_{0j}) = \tau_{00} \quad (4)$$

$$\mathrm{Var}(\mu_{1j}) = \tau_{11} \quad (5)$$

$$\mathrm{Cov}(\mu_{0j}, \mu_{1j}) = \tau_{01} \quad (6)$$

用公式(2)和(3)替换公式(1)相应的项，得到

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \mu_{0j} + \mu_{1j} X_{ij} + r_{ij} \quad (7)$$

公式(7)中，$(\mu_{0j} + \mu_{1j} X_{ij} + r_{ij})$是残差项。因为每一个第二层单位内的所有个体都有相同的 μ_{0j} 和 μ_{1j}，所以在相同的第二层单位内的个体之间的相似性就比来自不同第二层的单位的个体之间的相似性高，这就是相关残差的来源。由于 μ_{0j} 和 μ_{1j} 的值不相同，来自第二层不同单位的残差就可能有不同的方差。误差项之间就是相关的，方差不等，与 X_{ij} 有相关。如果研究资料不存在第二层单位间的差异，μ_{0j} 和 μ_{1j} 的值为 0，公式(7)就简化为简单的 OLS 回归，即 $Y_i = \beta_0 + \beta_1 X_i + r_i$。当数据存在层次结构时，$\mu_{0j}$ 和 μ_{1j} 的值不为 0，不能用简单的 OLS 回归进行分析而应该运用多水平模型。

实例分析

以中国/WHO 控烟能力建设合作项目——学校控烟子项目中天津地区中学的学生基线调查资料为例，介绍如何建立多水平模型并与传统 logistic 回归相比较。

由经统一培训的调查员，对随机整群抽样得到的 2 所学校 25 个班级的初中一年级和初中二年级的学生进行调查。调查时以班级为单位，采用统一的“中学生有关吸烟的知识、态度、行为调查问卷”进行调查。应调查 1120 人，实际调查 1086 人，有效问卷 1061 份占实际调查人数的97.70%。在 1061 人中没有试着吸过香烟者 874 人、试着吸过香烟者 187 人，分别占调查总数的82.4%和17.6%。

由于问卷是围绕着与吸烟危害健康有关的知识、态度及环境的内容展开的，所以应用 SPSS 11.5 软件对问卷中的 37 项变量进行因子分析，提取出 3 个共性因子，可以归纳为知识、态度和环境。利用 3 个共性因子及学生的性别、年龄、学习成绩建立多水平模型，统计分析采用 HLM6 进行。

1. 多水平模型的适用性：以是否尝试着吸过烟作为应变量(应变量是二分类变量)对数据进行二分类离散数据无效模型(即模型中除截距及其随机误差外，没有自变量)的拟合[2]，结果见表 1。

表1 中学生吸烟两水平无效模型

参数	估计值	$s_E(s)$	统计量值	P 值
γ_{00}	−1.647 510	0.175 928	−9.365	0.000[b]
$\sigma^2_{U_0}$	0.575 040	0.758 310[a]	109.081	0.000[b]

注：[a] 为 s；[b] $P<0.01$

$\sigma^2_{U_0}$ 是水平 2 残差方差属于模型中水平 2 随机部分，其含义是：不同班级的中学生对“是否尝试过吸烟”问题回答的变异情况，即班级差异。因模型中无解释变量，所以是指未考虑班级和中学生的特征时的变异。由于 $\sigma^2_{U_0}$ 的 $P = 0.000$，所以表明数据存在层次结构特征，即不同班级内的中学生尝试吸烟的情况存在相似性或聚集性，也就是说数据存在以班级为水平 2 单位的层次结构，适于进行多水平分析。

2. 多水平模型的建立：对与中学生吸烟有关的性别、年龄、学习成绩等因素进行分析，经逐步筛选得到多水平模型[3,4]：

水平 1 模型：$\mathrm{Prob}(Y_{ij} = 1 | \beta_j) = \phi_{ij}$

$$\eta_{ij} = \log\left[\frac{\phi_{ij}}{1-\phi_{ij}}\right] = \beta_{0j} + \beta_{1j}(\text{性别})_{ij} + \beta_{2j}(\text{年龄})_{ij} + \beta_{3j}(\text{态度})_{ij} + \beta_{4j}(\text{环境})_{ij}$$

水平 2 模型：$\beta_{0j} = \gamma_{00} + u_{0j}$；$\beta_{1j} = \gamma_{10}$；$\beta_{2j} = \gamma_{20}$；

$$\beta_{3j} = \gamma_{30}；\ \beta_{4j} = \gamma_{40} + \gamma_{41}(\mathrm{M1})_j + u_{4j}$$

总的模型可写成：

$$\eta_{ij} = \gamma_{00} + \gamma_{10}(\text{性别})_{ij} + \gamma_{20}(\text{年龄})_{ij} + \gamma_{30}(\text{态度})_{ij} + \gamma_{40}(\text{环境})_{ij} + \gamma_{41}(\text{环境})_{ij}(\mathrm{M1})_j + u_{4j}(\text{环境})_{ij} + u_{0j}$$

模型中年龄是以各自班级的年龄均数为中心进行转换，M1 是以班级水平衡量学校或班级组织吸烟危害活动宣传教育的情况。模型中各个参数的具体数值见表 2 和表 3。

表 2 和表 3 为模型拟合的最后结果。在模型拟合过程中，个人水平上引入的有统计学意义的变量只有“环境”一个变量水平 2 的随机效应有统计学意义；在有统计学意义的低水平变量上引入高水平的协变量——“宣传教育”，除环境因素外，其余变量引入协变量后均无统计学意义，即班级水平上的宣传教育对个人水平上的性别与吸烟的关系、年龄与吸烟的关系以及态度与吸烟的关系无影响，个人水平上的上述特征对吸烟行为的影响不受班级水平上的

宣传教育情况的不同而改变。

表2　中学生吸烟两水平模型参数估计

参数	固定效应		参数	随机效应	
	估计值	统计量		估计值	统计量
截距 β_{0j}					
截距 γ_{00}	−1.299 428	−7.101[a]	$\sigma^2_{U_0}$	0.537	104.934[a]
性别 β_{1j}					
截距 γ_{10}	−1.125 882	−6.320[a]	$\sigma^2_{U_1}$	0.116	29.651[b]
年龄 β_{2j}					
截距 γ_{20}	0.351 587	2.320[a]	$\sigma^2_{U_2}$	0.010	10.976[b]
态度 β_{3j}					
截距 γ_{30}	0.193 374	2.466[a]	$\sigma^2_{U_3}$	0.029	24.141[b]
环境 β_{4j}					
截距 γ_{40}	1.497 832	2.734[a]	$\sigma^2_{U_4}$	0.149	41.661[b]
M1 γ_{41}	0.106 674	2.335[a]			

注：[a] $P<0.01$；[b] $P<0.05$

表3　中学生吸烟两水平模型 *OR* 值

参数	*OR* 值(95% *CI*)
截距 β_{0j}	
截距 γ_{00}	0.273(0.187～0.398)
性别 β_{1j}	
截距 γ_{10}	0.324(0.229～0.460)
年龄 β_{2j}	
截距 γ_{20}	1.421(1.056～1.913)
态度 β_{3j}	
截距 γ_{30}	1.213(1.040～1.415)
环境 β_{4j}	
截距 γ_{40}	4.472(1.442～13.870)
M1 γ_{41}	1.113(1.012～1.223)

由表 2 所示，截距 $\gamma_{00}=-1.299$ 表示当性别为男性(性别=0)，年龄为相应班级的平均年龄，态度和环境取值均为 0 时，所调查班级的中学生尝试吸烟发生的平均概率为 $1/(1+\exp\{1.299\})=0.214$。$\gamma_{10}$ 表示在个人水平上中学生的性别对其吸烟发生概率的影响，$\sigma^2_{U_1}$ 表示不同班级中学生性别的随机效应，其无统计学意义，说明性别对吸烟行为的影响不存在班级间的差异。γ_{20} 表示在个人水平上中学生的年龄对其吸烟发生概率的影响，同样 $\sigma^2_{U_2}$ 无统计学意义，表明年龄对吸烟行为的影响也不存在班级间的差异。γ_{30} 表示在个人水平上中学生对吸烟行为的态度对其吸烟发生概率的影响，$\sigma^2_{U_3}$ 仍无统计学意义，表明态度这一因素对吸烟行为的影响也不存在班级间的差异，不同班级的中学生对待吸烟的态度是相似的。γ_{40} 表示在个人水平上中学生周围的环境对其吸烟发生概率的影响，$\sigma^2_{U_4}$ 具有统计学意义，不同班级之间是有差别的，即不同班级的环境情况是不同的，中学生吸烟情况在不同班级的环境下存在聚集性，而且以环境的系数最高，说明不同班级环境给中学生吸烟行为造成的影响是最显著的。γ_{41} 表示班级水平上的宣传教育对个人水平上的环境与吸烟的关系，其系数为正，表明宣传教育越强环境与吸烟的关系就越强，从本次研究内容上看由于结局变量是“是否尝试过吸烟”，所以这可能是一种伴随关系，可能是由于环境中的吸烟行为多，使得班级上的宣传教育增多。

从表 3 中可知，在其他条件相同的情况下女性吸烟的危险性是男性的0.324倍，即女性是一个保护因素；年龄、态度、环境均为危险因素，其中以环境的 *OR* 值为最高是4.472(95% *CI*：1.442～13.870)。

3. 与一般 logistic 回归模型分析结果的比较：本次研究的资料在不引入班级水平变量的情况下，利用两水平模型和一般 logistic 回归模型分别进行分析，比较结果见表 4。从表 4 中可以看出，利用一般 logistic 回归模型分析获得的各个估计值的标准误均比用两水平模型分析获得的标准误大，说明多水平模型处理具有层次结构的数据时，把低水平上的随机误差分解到高水平上，因而提高了模型估计的准确度。

表4　两种方法分析中学生吸烟情况的结果比较

变量	一般 logistic 回归模型		两水平模型	
	估计值	$s_{\bar{x}}$	估计值	$s_{\bar{x}}$
常数项	−5.196	1.663	−1.124	0.143
性别	−1.067	0.179	−0.905	0.113
年龄	0.251	0.113	0.300	0.070
态度	0.176	0.076	0.178	0.047
环境	0.268	0.083	0.159	0.075

讨　论

由于本次分析的资料是采用以班级为单位对每一名中学生进行调查的，且一般情况下吸烟行为除了受到其本身个体特征的影响，也受到生活环境的影响，即周围环境中吸烟的行为多，受学校、家长教育少的中学生吸烟危险性就增加，所以考虑资料可能具有层次结构。通过上述分析可以看出资料具有层次结构，利用传统的回归方程即忽视组效应得到的结果与利用多水平模型分析的结果有所出入，经比较可以认为多水平模型分析上述资料更加适宜，可信度更高。

多层数据结构比较普遍，在教育研究中，学生嵌

套于学校；家庭研究中，儿童嵌套于家庭；医学研究中患者嵌套于医生或医院等。层次结构数据也可出现在特殊的研究设计中，例如流行病学调查或社区调查中，按照地区、个人进行分层随机抽样，所得数据具有地区和个人两个层次结构。许多实验设计也产生层次数据，例如新药的多中心试验等[5]。多水平模型主要有以下特点：①可以同时检验群组水平和个体水平上因变量的效应；②考虑到群组内个体的非独立性；③没有把群组或环境当作无关联的，而是看作来自于大的群组总体的一部分；④可以检验个体间和群组间的变异。因此，研究者利用多水平模型同时处理细微的层次（个体）和大的层次（群组或环境）[6-10]。

多水平模型在国外应用比较广泛。在我国，由于研究者逐渐认识到研究资料中存在层次结构，因而多水平模型的应用范围也在逐步扩大。金芳等[11]在对儿童生长发育研究中考虑到调查资料来自不同的家庭且同一儿童的身高、体重之间存在一定的相关性，所以利用多元多水平模型进行分析。郭伯良等[12]发现在不同的研究中，儿童行为问题（如攻击）与学校适应间的关系会有两种截然不同的表现（排斥或拥戴），其原因即为忽略了班级水平上环境变量（如老师因素）的影响。同样，李新华等[13]在对学龄前集体儿童行为问题的研究中发现，班级内的儿童之间在行为特征上趋于一致，不同班级之间趋向不一致。多水平模型也可用于分析疾病的地理分布，如李德云等[14]利用四川省 3 年碘缺乏病监测项目资料分析儿童尿碘的变异，发现在地区和个体水平上尿碘的含量均不相同。

基于可操作性的原因和成本效益的考虑，大多数流行病学调查均采用分层或整群的抽样方法，这就不可避免的产生了层次结构，在某种意义上，多水平模型提供了一种途径，它把传统意义上截然不同的生态学和个体水平上的研究结合起来，克服了只聚焦于一个单一水平的局限性。它可以使得研究人员特别是公共卫生研究者更能清楚的解释不同层次上的影响因素。但当数据具有很少的结构性时，则几乎不需要多水平方法，用传统的单水平模型分析和交流就足够了[6,10,15]。

参 考 文 献

[1] 张雷，雷雳，郭伯良. 多层线性模型应用. 北京：教育科学出版社，2003：4-20.

[2] Hox JJ. Applied multilevel analysis. Amsterdam: TT-Publikaties, 1995: 31-47.

[3] Stephen WR, Anghony SB, Yuk FC, et al. HLM5: Hierarchical Linear and Nonlinear Modeling, Chicago: Scientific Software International, 2001: 10-25.

[4] Stephen WR, Anghony SB, Yuk FC, et al. HLM6: Hierarchical Linear and Nonlinear Modeling, Chicago: Scientific Software International, 2004: 113-134.

[5] Hox JJ. Multilevel modeling: when and why//Balderjahn I, Mathar R, Schader M. Classification, data analysis, and data highways. New York: Springer Verlag, 1998: 147-154.

[6] Goldstein H. Multilevel Statistical Models(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.

[7] 张岩波，何大卫，刘桂芬，等. 离散型二分类数据的广义混合线性模型分析——混合线性模型及其 SAS 软件实现（三）. 中国卫生统计，2001，18：328-330.

[8] Rasbash J, Browne W, Goldstein H, et al. A user's guide to MlwiN(Version 2.1d). Centre for Multilevel Modelling Institute of Education University of London, 2002.

[9] Duncan C, Jones K, Moon G. Context, composition, and heterogeneity: using multilevel models in health research. Soc Sci Med, 1998, 46: 97-117.

[10] Ana V Diez-Roux. Multilevel analysis in public health research. Ann Rev Public Health, 2000, 21: 171-192.

[11] 金芳，倪宗赞，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，2004，21：204-206.

[12] 郭伯良，王燕，张雷. 班级环境变量对儿童社会行为与学校适应间关系的影响. 心理学报，2005，37：233-239.

[13] 李新华，张伟，夏梓红. 学龄前集体儿童行为问题影响因素的多水平模型. 贵阳医学院学报，2006，31：38-42.

[14] 李德云，高亚礼，李晓松，等. 儿童尿碘地理分布的多水平统计模型. 中国卫生统计，2006，23：31-33.

[15] Schwartz S, Susser E, Susser M. A future for epidemiology? Ann Rev Public Health, 1999, 20: 1-19.

（收稿日期：2006-11-01）

（本文编辑：张林东）