

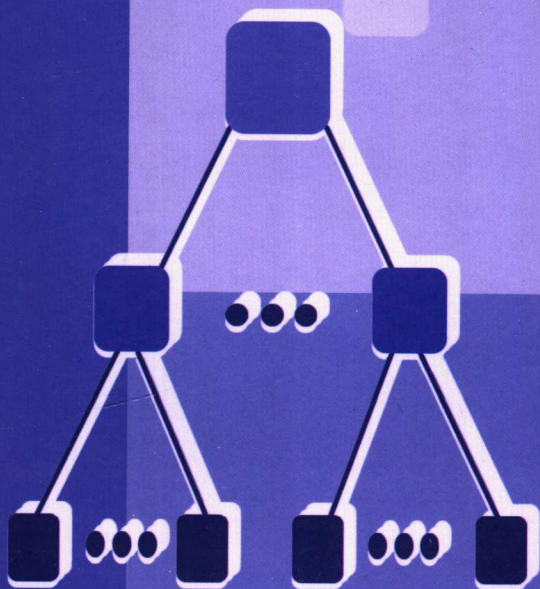


1922

China Medical Board of New York, Inc.

# 医学和公共卫生研究 常用多水平统计模型

主编 杨珉 李晓松



北京大学医学出版社

# 医学和公共卫生研究 常用多水平统计模型

主 编 杨 珉 李晓松

编 者 (按姓氏笔画排序)

刘巧兰 (四川大学华西公共卫生学院)

李晓松 (四川大学华西公共卫生学院)

朱彩蓉 (四川大学华西公共卫生学院)

杨 珉 (英国伦敦大学玛丽女王医学院)

张晓兰 (四川大学华西公共卫生学院)

张菊英 (四川大学华西公共卫生学院)

陈 峰 (南京医科大学公共卫生学院)

殷 菲 (四川大学华西公共卫生学院)

北京大学医学出版社

## 图书在版编目 (CIP) 数据

医学和公共卫生研究常用多水平统计模型/杨珉 李晓松主编.  
北京: 北京大学医学出版社, 2007.5  
ISBN 978-7-81116-239-4

I. 医... II. ①杨... ②李... III. ①医学统计—统计模型  
②公共卫生—统计模型 IV. R195.1

中国版本图书馆 CIP 数据核字 (2007) 第 034113 号

## 医学和公共卫生研究常用多水平统计模型

主 编: 杨 珉 李晓松

出版发行: 北京大学医学出版社 (电话: 010-82802230)

地 址: (100083) 北京市海淀区学院路 38 号 北京大学医学部院内

网 址: <http://www.pumpress.com.cn>

E - mail: [booksale@bjmu.edu.cn](mailto:booksale@bjmu.edu.cn)

印 刷: 莱芜市圣龙印务有限责任公司

经 销: 新华书店

责任编辑: 冯智勇 责任校对: 金彤文 责任印制: 张京生

开 本: 787mm×1092mm 1/16 印张: 13 字数: 252 千字

版 次: 2007 年 5 月第 1 版 2007 年 5 月第 1 次印刷 印数: 1—3000 册

书 号: ISBN 978-7-81116-239-4

定 价: 38.00 元

版权所有, 违者必究

(凡属质量问题请与本社发行部联系退换)

# 前 言

多水平统计模型又称随机效应模型，是将 II 型方差分析理论与多元统计分析相结合的新技术，是分析和处理具有层次结构特征数据的有力工具，亦是当前国际上统计学研究中一个新兴而重要的领域。多水平模型的研究和应用始于 20 世纪 80 年代初期英国初级中学学校效益分析及美国的补偿教育研究，其理论、方法及计算技术发展迅速，目前在心理学、社会学、经济学、政治学、地理学、实验研究、儿童生长发育、流行病学及其他医学领域应用广泛，并展现出日益广阔的应用前景。随着计算机技术的发展，现已建立起一套可以常规应用的技术与软件，MLwiN 是用于多水平分析的专用软件，此外，利用 SAS、SPSS、Stata、S-Plus、MPlus 以及 LISREL 等软件可以实现较为基本但程度不同的多水平模型分析。

在我国，多水平模型的研究和应用才刚刚起步，有关研究和应用的文献尚不多见。许多研究人员已经意识到多水平模型在处理层次结构数据上的重要性，但目前尚缺乏一本较为系统介绍多水平模型及其应用的专著。

本书的目的旨在向医疗卫生领域里有定量研究经验的学术界人士介绍基本的多水平统计模型的定义和模型的估计值的解释以指导正确应用。书中针对医学和公共卫生研究中常见的数据结构，以医学和公共卫生领域的实际数据为例，介绍了各类多水平模型的基本原理、模型结构、建模方法、软件实现以及结果解释，尤其是介绍了多水平模型与常规分析方法的联系，以及各类软件的具体操作方法，使读者能通过自学实践并重复书中的模型拟合过程，在重复实践中加深对相关模型的理解并掌握操作技术。

本书第一章“概述”介绍了医学和公共卫生研究中常见的多水平数据以及忽略数据多水平结构的后果，介绍了多水平模型和计算程序的最新进展以及如何使用本书；第二章“方差成分模型”从方差分析和简单回归模型入手，介绍了简单多水平模型的基本原理；第三章“随机系数模型”则涉及常规分析方法难以处理的情形；第四章“重复测量数据的多水平模型”则针对重复测量数据时间上的强相关性特点，介绍了一类相对独立的多水平模型尤其是生长模型；第五章“二分类资料的多水平模型”从常规的 logistic 模型入手，根据数据结构特征将其推广到两水平和三水平的 logistic 模型，此外还重点介绍了两个相关应变量的三水平的 logistic 模型分析；第六章“Poisson 分布资料的多水平模型”则覆盖了医学和公共卫生研究中常见的另一大类数据；第七章“多水平 Meta 分析模型”较为系

统地介绍了随机效应模型与多水平模型的联系，介绍了不同设计和不同类型资料的多水平 Meta 分析，还介绍了聚集数据和个体数据组合的多水平 Meta 分析；第八章“多水平生存时间的统计模型”针对医学和公共卫生领域中十分常见的研究类型和数据类型，介绍了多水平对数时间模型和比例风险模型。

本书的出版将有助于我国应用统计界对多水平模型的研究及推广应用，同时也有助于带动相关学科领域的研究和应用，有助于提高应用统计研究生的教学水平。

本书适合高等学校和科研机构具有基础统计学理论知识和统计软件应用技能的教师、科研人员和研究生做专业参考工具书，亦适合于卫生统计学、流行病学、公共卫生学和社会医学的研究生做专业教材。

尽管本书所用例题采用的是医学和公共卫生研究数据，但基本原理、模型结构、分析方法以及结果解释可以推广到其他领域例如教育学、心理学、社会学、经济学等研究领域。非医学专业的读者可以触类旁通、举一反三地加以应用。本书所附光盘含有书中例题的数据以及多水平统计模型专用软件 MLwiN（试用版），读者可应用该数据以及书中所给出的程序进行自学和练习。此外，MLwiN 的其他资料可在 <http://www.cmm.bris.ac.uk/MLwiN> 下载。

感谢美国纽约中华医学基金会（China Medical Board of New York, Inc.）为本书的出版所提供的资助。四川大学华西公共卫生学院卫生统计学教研室陈勇、陈晓娟和沈卓之三位硕士研究生先后承担了本书的学术秘书工作，为本书的例题计算结果复核、书稿编排及校对等工作付出了辛勤劳动，在此一并致谢。

杨 珉 李晓松

# 目 录

<b>1 概 述</b> .....	(1)
1.1 医疗和卫生研究中的多水平数据 .....	(1)
1.2 忽略数据多水平结构的后果 .....	(2)
1.3 多水平模型和计算程序的进展 .....	(4)
<b>2 方差成分模型</b> .....	(6)
2.1 固定效应和随机效应方差分析模型 .....	(6)
2.2 带协变量的随机效应方差分析模型 .....	(9)
2.3 两水平方差成分模型 .....	(10)
2.4 模型参数估计值的假设检验和区间估计 .....	(13)
2.5 残差分析及其应用 .....	(14)
2.5.1 模型假设的诊断 .....	(15)
2.5.2 高水平单位残差的比较 .....	(16)
2.6 两水平方差成分模型的扩展 .....	(18)
2.7 各软件拟合例 2.2 模型的操作过程 .....	(18)
2.7.1 MLwiN .....	(18)
2.7.2 SAS .....	(23)
2.7.3 SPSS .....	(25)
<b>3 随机系数模型</b> .....	(29)
3.1 一个实例 .....	(29)
3.2 模型设置及参数含义 .....	(32)
3.3 模型拟合及结果解释 .....	(35)
3.4 用 MLwiN 拟合随机系数模型 .....	(36)
3.5 用 SAS 拟合随机系数模型 .....	(44)
3.6 用 SPSS 拟合随机系数模型 .....	(45)
<b>4 重复测量资料的多水平模型</b> .....	(49)
4.1 重复测量资料的性质 .....	(49)
4.2 测量时间固定的多水平模型 .....	(51)
4.3 交叉试验资料的多水平模型 .....	(54)
4.4 两水平生长模型 .....	(58)
4.5 自相关残差模型 .....	(61)

4.6	用 MLwiN 软件拟合重测量资料的多水平模型 .....	(63)
<b>5</b>	<b>二分类资料的多水平模型</b> .....	(69)
5.1	问题的背景 .....	(69)
5.2	二分类反应变量的单水平模型 .....	(70)
5.3	二分类反应变量的两水平模型 .....	(72)
5.4	一个三水平数据结构的实例 .....	(75)
5.5	参数估计方法 .....	(80)
5.6	文中实例的拟合过程及程序 .....	(81)
5.7	双反应变量的多水平模型 .....	(92)
<b>6</b>	<b>Poisson 分布资料的多水平模型</b> .....	(104)
6.1	Poisson 模型简介 .....	(104)
6.2	两水平 Poisson 模型 .....	(105)
6.3	超 Poisson 方差模型 .....	(114)
6.4	两水平负二项分布模型 .....	(116)
6.5	多水平 Poisson 模型 .....	(118)
6.6	用 MLwiN 软件拟合多水平 Poisson 模型 .....	(122)
<b>7</b>	<b>多水平 Meta 分析模型</b> .....	(126)
7.1	传统 Meta 分析方法 .....	(126)
7.2	正态资料的多水平模型与 Meta 分析的随机效应模型 .....	(129)
7.3	疾病危险因素队列研究资料的 Meta 分析模型 .....	(135)
7.4	疾病危险因素病例对照研究资料的 Meta 分析模型 .....	(140)
7.5	组合聚集数据和个体数据的两水平 Meta 分析模型 .....	(143)
7.6	多水平 Meta 模型的拓展 .....	(147)
7.7	用统计软件拟合实例的操作 .....	(148)
7.7.1	MLwiN .....	(148)
7.7.2	SAS .....	(163)
<b>8</b>	<b>多水平生存时间的统计模型</b> .....	(166)
8.1	具有多水平结构的生存时间数据 .....	(166)
8.2	对数时间模型 .....	(168)
8.3	比例风险模型 .....	(173)
8.4	生存时间模型中的随机效应解释和应用 .....	(177)
8.5	检查比例风险模型的假设 .....	(179)
8.6	离散-时间风险模型 .....	(179)
8.7	用 MLwiN 软件拟合多水平生存时间模型 .....	(181)

参考文献 .....	(187)
中英文索引 .....	(192)
例题数据目录 .....	(198)



## 1.1 医疗和卫生研究中的多水平数据

多水平数据(multilevel data)或具有多水平层次结构(hierarchical structure)的数据是多水平统计模型(multilevel statistical models)发展和应用的基础。此类数据的主要特征是反应变量的分布在个体间不具备独立性,但存在地理距离内、某行政区划内或特定空间范围内的聚集性(clustering)。例如在医疗卫生服务研究中,采集病人对所在服务区域的保健机构的服务质量的意见,基于各保健机构的医疗资源和医生的经验、态度的不同,所服务病人的意见将有所不同。服务好机构,多数病人会提出好的意见;而服务差的机构将会收到较多的批评意见。如果我们将病人作为一个水平或层次单位,且定义为水平1的单位(level 1 unit);将诊所定为高一个水平即水平2的单位(level 2 unit),所收集到的数据则呈现出在水平2内(诊所内)的相似性(similarity),又称聚集性。换句话说,各病人的意见作为反应变量不具独立性。又例如对高血压病的危险因素作调查分析,不同地区的经济文化背景和人群的长期生活饮食习惯可能导致高血压病症的发病率不同,即高发地区和低发地区之分。如果要研究地区环境因素对高血压的影响,这时发病的地区聚集则是一个不可忽略的数据特征。此数据也呈现了水平1的个体在水平2地区内的相似现象。再例如卫生毒理实验研究中的致畸实验、发育毒性或生殖毒性实验,存在大量数据以“仔鼠”为水平1的观察单位,而窝别为水平2的观察单位。由于遗传因素、妊娠期长短和致畸物的代谢环境的影响,这类数据的窝别效应尤为明显:同窝的仔鼠发生畸形的概率趋于相同,而不同窝别的仔鼠其畸形发生率有较大的差异。

具有以上层次结构和非独立性的数据还大量地存在于临床试验中的重复测量值(repeated measures)和儿童生长发育研究,以及多中心药物疗效研究中。

在实际研究中,收集多水平结构数据常常与要回答的研究问题密切相关。而这些研究问题常与水平2的单位或水平2的变量(level 2 variable)有关。例如,卫生服务质量研究需要病人对服务于他们的诊所或医生做评价,最后的研究结论是在水平2单位上。其中的医生或诊所的背景将是不可忽略的水平2的变量,应该在统计分析中加以考虑。在高血压发病危险因素研究中,确定发病高发区,并鉴别地区水平上的危险因素如人均收入、脂类食物消耗量、饮食习惯等常是研究

者关注的变量。在儿童生长发育研究中，个体儿童作为水平 2 单位，每个儿童在不同时间(年龄)上的体格测量值为水平 1 单位。研究者的兴趣常常在于儿童间生长的个体差异和共同生长模式，以便有效地鉴别正常与异常生长的儿童。

## 1.2 忽略数据多水平结构的后果

传统的多元回归模型是人们常用于研究筛选危险因素或了解有关变量与结果变量关系的方法。例如高血压发病的危险因素研究，最简单的办法是将个体发病与否做结果变量，将所有个人变量和地区变量做协变量拟合 logistic 模型或正态回归模型。但这种方法将有两个问题：其一，传统回归模型的估计方法是建立在个体测量值相互独立的假设上。当这个假设不成立时，回归模型中的各参数估计值的有效性和统计特性均会受到影响，从而最终的统计推断结论将可能偏倚；其二，地区变量对同一地区的每一个体的取值是一样的，这即是说该变量的自由度在同一地区内已经人为地扩大了，其结果无疑会影响该变量参数估计值的统计推断，因为该估计值的标准误会有偏差。

以下一实例可显示所述问题一：

表 1.1 两周患病及影响因素的单水平正态线性模型  
和两水平正态线性模型比较

变量	单水平线性模型 $\hat{\beta}(SE)$	两水平线性模型 $\hat{\beta}(SE)$
年龄 $\geq 65$ (岁)	0.085(0.026)	0.075(0.026)
女	0.036(0.018)	0.035(0.017)
经常饮酒	0.097(0.025)	0.088(0.025)
患有慢性病	0.474(0.026)	0.463(0.026)
-2 对数似然值	2215.2	2180.87
残差标准误平方	0.157(0.005)	0.139(0.005)

表 1.1 数据来自全国第三次卫生服务调查在某省农村抽样的部分数据，含 647 个住户中的 2239 个调查对象。这里的两水平模型以住户为水平 2 单位，个人为水平 1 单位，结果变量是两周内患病与否，实际观察两周内患病与否在家庭内有聚集性。为说明单水平线性模型(single level linear model)，即忽略患病在家庭内的聚集的情况下，其估计值与考虑到此情况的两水平线性模型(2-level linear model)估计值的差别，我们将二分类的结果变量简单处理为线性变量分析。书中第五章将详细介绍此资料的非线性模型的分析方法。

表 1.1 中的结果显示，两水平线性模型对此数据的拟合优度优于单水平线性模型，表现在较小的残差(residuals)标准误和较小的对数似然值(log-likelihood)。

每一个变量的效应估计值(estimate of effect)的标准误,两模型结果几乎相同,单水平线性模型估计的变量效应值均明显偏大,即在显著性检验中增加了 I 型错误的机会。单水平线性模型的残差标准误偏大的原因是,其中包含了个体间在家庭内的患病差异和家庭间的患病差异。而两水平线性模型则分离了这两部分差异,所得残差标准误只含个人在家庭内的差异部分。

再以卫生服务质量评价一例进一步阐述上面的问题二。传统回归分析可能直接回答的问题是病人的年龄、性别和病情是否影响他们对医生或诊所服务的评价。若要回答哪个医生或诊所做得更好,哪个做得更差,一个分析方法可能是将病人的评价结果在水平 2(诊所内)求平均值,然后拟合水平 2 变量与病人评价均值的回归。如要加入水平 1 的病人变量,也需要对其求平均值或集合值(agggregation),这样的做法无疑是以损失病人间的差异为代价,或数据损失有效信息为前提的,其统计推断结果也失有效性。更重要的是,病人个人背景与地区内病人背景的集合与诊所服务质量的关系可能会不一致或相互独立可加。例如:病人年龄越大,看医生的机会越多,对疾病的耐受性越强,而对服务质量的评价标准越宽,从而反映到容易满足基本服务,即年龄越大,越可能给出较好的评价。但在老年人较多的社区,由于相互之间的言传影响,这里的老年人会比老年人较少的社区的同龄人给出更高的评价,即出现了社区年龄对服务质量评价的附加影响。但也有可能呈现老年人较多的社区卫生资源不足,多数老年人没有受到恰当的服务,结果社区年龄与质量评价呈负相关关系。下表以一例子说明此现象:

表 1.2 2 型糖尿病患者对社会医疗服务的满意程度模型比较

变量	单水平线性回归 $\hat{\beta}(SE)$	单水平集合线性回归 $\hat{\beta}(SE)$	两水平线性回归 $\hat{\beta}(SE)$
个人年龄	0.0032(0.0009)	—	0.0032(0.0008)
社区平均年龄	-0.0298(0.0029)	-0.0231(0.0039)	-0.0288(0.0040)
-2 对数似然值	1311.80	-99.81	1289.02
残差标准误平方	0.1419	0.0138(0.0024)	0.1354(0.0051)

该数据来自 1482 名 2 型糖尿病患者对 69 名社区医生服务的满意度评价,结果变量为满意度分数的对数值。病人年龄从 31~91 岁,而病人所代表的社区病人平均年龄为 54~72 岁。

比较表 1.2 中三个模型拟合结果的对数似然值,我们可以判断两水平线性回归模型相对最好。

尽管表 1.2 中三种模型所估计的水平 2 的变量(社区病人平均年龄)与社区医生服务满意程度均呈相关,并具有统计学意义,然而三种模型对此变量的效应估计值略有不同,其估计值的标准误亦有很大差别。单水平线性回归分析给出的该变量估计值的标准误最小,可认为推断出现 I 型误差(type I error)的可能性最

大。而单水平集合线性回归(aggreated linear regression)分析不能加入病人的变量,只对69个数据点或均数进行回归。其结果是参数估计值偏小,因为数据信息大量损失,这也从负的对数似然值上反映出来。

再比较各模型的残差的标准误(standard error of residuals),单水平线性回归模型的最大,因为其中包含了评价结果在病人间的差异和在社区间的差异。单水平集合线性回归模型的残差标准误最小,因为它仅仅反映了评价结果在社区间的差异。而两水平线性回归模型的残差标准误居中,因为该模型将社区的差异分离出来后,剩下的只是评价结果在病人间的变异度。由此不难看出传统模型在处理含水平结构资料和对不同水平上的变量估计效益的弊端以及多水平模型的合理性。

书中第2、3章将进一步举例介绍两水平线性回归模型的定义和应用。

### 1.3 多水平模型和计算程序的进展

20世纪80年代初期,西方学者在进行学校效益评价的教育研究过程中,提出多水平结构的数据和传统统计模型在处理该数据时出现的问题(Aitkin & Longford, 1986)。与此同时,另有学者就儿童生长发育研究中的重复测量数据结构特征对传统统计模型的分析方法提出了质疑(Goldstein H, 1979),其问题的焦点在于:传统回归模型假设拟合模型后的误差是随机并相互独立完全不适用有多水平结构的数据。而后的误差来源应该与数据的结构相对应,即一个结构水平应对应一个误差来源,如高血压的患病有个人间的差异亦有地区间的差异;医疗服务在地区间有差别,对不同病人亦有差异,忽略此数据特征的统计分析方法所带来不适当的推断结论在上节已有说明。英国伦敦大学教育研究(Institute of Education)统计学家Goldstein H于1985年提出了拟合多水平模型的迭代广义最小二乘法(Iterative Generalized Least Square, IGLS),并推出了专用计算程序ML2(multilevel models for 2-level data),进而进一步推广其计算程序至ML3、MLn、MLwiN (Multilevel models for N level data - Window Version)(Rasbash et al, 2000)。到目前为止,MLwiN已经能够在用户友好界面下高效率地拟合各种变量类型(连续线性、二分类、二项分布、计数、多项有序、多项无序)的多水平统计模型(Jones A, 2005)。

另一英国统计学家Longford N于1987年提出了Fisher-scoring算法,并推出了专用计算程序VARCL。此计算是基于方差分析中的方差成分模型(variance component model)原理而出,但此程序于20世纪90年代早期已停止发展。

同期有美国教育统计学家S. Randenbush在1986年提出的E-M数算法并推出了专用计算程序HLM(Hierarchical Linear Models)。该程序也已由最初的

仅拟合两水平数据的菜单一对话式格式发展到如今的微软窗户式程序，并能友好地拟合多种数据形式的包含三水平结构的模型，如线性模型、Logistic 模型、Poisson 模型和有序/无序分类的多项式模型 (multinomial models) (Yang M, 2003a)。

在以上先驱方法与计算程序的带动下，随着多水平模型在教育研究领域中的应用的深入和在其他应用领域包括社会学和医疗卫生中的推广，现有的各大、中型综合统计软件包亦在 20 世纪 90 年代中、后期逐渐增加了拟合多水平模型的功能块。SAS 在 1996 年新增了 MIXED 过程以拟合含多水平结构的线性模型。到目前为止，它的 NLMIXED 过程可以拟合含多水平结构的非线性模型，如 Logistic、Poisson 和有序/无序分类数据的多项式模型 (Yang M, 2002)。SPSS 软件包版本 11 亦可以拟合多水平的线性模型，包含对重复测量数据的分析模型 (Leyland, 2005)。Stata 统计软件目前亦有相应指令，拟合含两水平结构的基本线性模型和非线性的基本 Logistic 及 Poisson (Mourouga P, 2005)。对于更复杂的三水平模型，Stata 亦有特别的宏程序 GLMMA 可实现模型拟合 (Grilli L, Rampichini C, 2005)。Splus 统计软件包也在 20 世纪 90 年代末期专门设计了子程序 R (Verzilli C, 2003) 来实现对多水平模型的拟合分析。另外它也有相应的程序指令拟合多水平模型。小型统计软件如 Systat (Yang, 2003b) 和 Genstat (Welham S, 2003) 现今都有标准的工具处理多水平数据，拟合基本的多水平统计模型。

由于不同的软件在拟合多水平模型时所用的模型估计方法和计算机算法有差别，使得所拟合模型的结果也略有差别，而且在计算速度、能处理模型的效益和数据量的局限性上都有差别。原英国伦敦大学教育研究所的多水平模型研究中心在 2002~2004 年间组织了多个国家的有关学者，对当前国际流行的所有的统计软件包中的多水平模型功能作了全面和详细的复习，并在各个软件包间作对比。其结果在该中心的万维网页中向国际上的学者开放交流 ([Http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/index.shtml](http://www.cmm.bristol.ac.uk/learning-training/multilevel-m-software/index.shtml))，并还在不断加以补充和完善。

1995 年英国伦敦大学教育研究所学者在与北京中央教育科学研究所的学术交流活动中首次介绍了多水平统计模型及其在教育研究中的应用，随之在国内出现了多水平统计模型在医疗卫生研究中的应用报告 (李晓松，倪宗瓚，1998；杨珉，1997；曹秀堂，徐勇勇，姚晨，1995)。随着四川大学华西医学中心公共卫生学院卫生统计学研究者将 Goldstein H 的《多水平模型在社会学中的应用》一书翻译成中文在全国发行 (李晓松，杨珉，1999)，多水平统计模型的应用价值被越来越多的应用学者所重视，对此方法的深入学习了解和正确应用已成为当前教育、医疗卫生及社会学研究人员的迫切需要。

(杨珉 李晓松)

在文献中,多水平模型(multilevel model)有各种不同的称谓,如“随机系数模型(random coefficient model)”、“层次线性模型(hierarchical linear model)”以及“混合效应模型(mixed-effects model)”或“混合模型(mixed model)”。这些文献所提及的模型或许在算法或在应用领域的普及程度上有差别,但它们均用于处理具有层次结构的数据或非独立误差数据。本章将介绍多水平模型中最简单的方差成分模型(variance component model)。包括模型的基本定义、与传统方差分析方法的联系、实际应用和结果解释。本章的叙述从固定效应(fixed effects)和随机效应(random effects)方差分析模型开始。

## 2.1 固定效应和随机效应方差分析模型

通常,在试验设计和调查设计中,涉及不同处理组时,会提及固定效应和随机效应这两个概念。某研究中有多个不同处理因素,若研究者感兴趣的各种处理都设计在研究当中,则认为这一因素具有固定效应,如以下例 2.1 中对小白鼠给予三种不同的营养素;若处理包含的各个组别是从更大的总体中得到的随机样本,则认为该处理因素具有随机效应,如以下例 2.2 中病人对社区医生服务的满意度研究。

**例 2.1** 为了解不同营养素增重的效果,设计上将 24 只小白鼠随机分成 3 组,每组 8 只:第 1 组用 A 营养素进行喂养,第 2 组用 B 营养素进行喂养,第 3 组用 C 营养素进行喂养。研究者仅关心小白鼠经三种不同营养素喂养后所增体重有无差别。

由此推论 3 个处理组别对试验结果的影响。这里假定三种处理效应是固定不变的,即只关心 A、B、C 三种营养素对体重增加的影响。该例可采用完全随机设计单因素方差分析(one-way ANOVA)对三种营养素的增重效果进行比较。

单因素方差分析的目的在于比较  $J$  个组数据的总体均数有无差别,其无效假设为:  $\mu_1 = \mu_2 = \dots = \mu_j = \mu$ ,  $j = 1, 2, \dots, J$ , 第  $j$  组样本数  $n_j$ ,  $\sum_j n_j = N$ 。

假定  $\mu$  是  $J$  个组总体均数的加权平均值,即  $\mu = \sum_j n_j \mu_j / N$ , 各组均数与总体均数的差值为  $\alpha_j = \mu_j - \mu$ , 各组第  $i$  个个体测量值  $y_{ij}$  与各组均数  $\mu_j$  的差值  $e_{ij} = y_{ij} - \mu_j$ , 则单因素方差分析的线性模型可表示为:

$$y_{ij} = \mu_j + e_{ij} \text{ 或 } y_{ij} = \mu + \alpha_j + e_{ij} \quad (2.1)$$

模型(2.1)的期望值为  $E(y_{ij}) = \mu + \alpha_j$ ,

方差为  $var(y_{ij}) = var(\mu + \alpha_j + e_{ij}) = var(e_{ij}) = \sigma_e^2$ , 且  $e_{ij} \sim N(0, \sigma_e^2)$

此模型中, 各组均数与总体均数的差值  $\alpha_j$  为固定数值, 理论上不存在变异, 因此有  $var(\mu + \alpha_j) = 0$ , 故称为固定效应模型(fixed effect model)。模型中含有四个基本假设: 变量间相互独立、来自正态分布总体、组间效应与误差线性可加及方差齐性。

基于上述模型和假设, 用最小二乘法可获得现有的 I 型单因素方差分析(type I ANOVA)的算法, 求解  $\alpha_j$  的估计值的线性模型即为固定效应模型。其计算方法在几乎所有的统计教科书中均有介绍。

但在很多实际研究中, 所观测的  $J$  组可能只是一个随机样本。研究者所关心的不再是只比较  $J$  组均数之间的差别, 而在于了解这  $J$  组均数所在的总体情况, 并根据这  $J$  组的样本对其总体做出统计推断, 如例 2.2。

**例 2.2** 某研究调查 2 型糖尿病患者对社区医疗服务的满意程度。研究采用两阶段随机抽样, 先在某地随机抽取 69 名社区医生, 然后在每名医生的诊所抽取若干 2 型糖尿病患者, 共计纳入 1482 名患者。每名入选医生需填写个人相关资料, 包括个人背景、从业时间等。每名入选患者除提供基本资料(年龄、性别、病史等)外, 需完成一份满意度调查的问卷。该问卷总分为 100 分, 分数越高, 表示患者对医疗服务的满意程度越高。数据列于表 2.1。

表 2.1 1482 名糖尿病患者满意度调查结果

病人 ID	医生 ID	病人年龄(岁)	病人满意度(分)	医生从业时间(年)
100001	001	59.84	92.86	8
100004	001	59.56	60.71	8
100005	001	73.24	53.57	8
100006	001	46.42	100.00	8
100007	001	68.41	62.50	8
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
103358	081	57.18	73.21	10
103360	081	45.86	60.71	10
103361	081	73.78	64.29	10

这里 69 名社区医生是某地随机选择的样本, 研究目的不是估计患者对这 69 名医生的满意程度, 而是估计这 69 名社区医生代表的该地所有社区医生的情况。因此, 对该地所有社区医生而言, 这 69 名医生的调查结果是随机的。若分析患

者对不同医生医疗服务的满意程度有无差别, 这时无效假设是各组间的变异度没有差别, 即:  $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_j^2$ 。

用线性模型来表达上述研究问题, 有:

$$y_{ij} = \mu + \alpha_j + e_{ij}, \quad \alpha_j = u_j - u \quad (2.2)$$

$$j=1, 2, \dots, J, \quad i=1, 2, \dots, n_j, \quad E(y_{ij}) = \mu$$

此模型假定  $\alpha_j$  和  $e_{ij}$  均为正态分布的随机变量,

$$\text{var}(y_{ij}) = \text{var}(\mu + \alpha_j + e_{ij}) = \text{var}(\alpha_j + e_{ij}) = \sigma_a^2 + \sigma_e^2$$

$\alpha_j \sim N(0, \sigma_a^2)$ ,  $e_{ij} \sim N(0, \sigma_e^2)$ , 且  $\alpha_j$  和  $e_{ij}$  相互独立, 即  $\text{cov}(\alpha_j, e_{ij}) = 0$ 。

$\sigma_a^2$  又称为组间变异度 (within group variance),  $\sigma_e^2$  为组内变异度 (between group variance)。针对无效假设, 这类分析的核心是检验组间变异度  $\sigma_a^2$  是否为 0。由于这里的组间效应  $\alpha_j$  为随机变量, 模型(2.2)又称为随机效应模型 (random effect model), 所对应的方差分析为 II 型方差分析 (type II ANOVA), 又称方差成分模型 (variance component model)。组间均方的期望值为  $\sigma_a^2 + n\sigma_e^2$ 。

模型(2.2)假设反应变量有两个方差成分  $\sigma_a^2$  和  $\sigma_e^2$ , 即对应于组间和组内的随机误差, 反应变量不具有独立性, 即同组的测量值有相似性。相似的程度可用组内相关系数 (intra-class correlation, ICC)  $\rho$  衡量。

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \quad (2.3)$$

以下我们用例 2.2 解释模型(2.2)的 II 型方差分析 (即随机效应方差分析) 思想。

例 2.2 中,  $y_{ij}$  为第  $j$  名医生的第  $i$  名患者的满意程度,  $\mu$  为患者满意程度的总体均数,  $\mu_j$  为患者对第  $j$  名医生的满意程度的均数 (组均数),  $\alpha_j = \mu_j - \mu$ , 因此  $\alpha_j$  表示组均数与总体均数的差值,  $e_{ij}$  表示第  $j$  名医生的  $i$  名患者的满意程度的随机误差。

$y_{ij}$  的方差是两个方差成分的合计, 即:

$$\sigma_y^2 = \sigma_a^2 + \sigma_e^2 \quad (2.4)$$

其中  $\sigma_a^2$  为组间方差 [本例为不同医生之间满意程度与满意程度总均数的方差  $\text{var}(\alpha_j)$ ],  $\sigma_e^2$  为组内方差 [本例为同一医生的不同患者之间满意程度的随机误差的方差  $\text{var}(e_{ij})$ ], II 型方差分析 (随机效应方差分析) 就是通过样本数据得到  $\sigma_a^2$  和  $\sigma_e^2$  的估计值。

其估计方法是在完全随机设计方差分析表的基础上, 通过组间均方  $MS_a$  和组内均方  $MS_e$  的数学期望值  $EMS$  (Expected Mean Square), 获得  $\sigma_a^2$  和  $\sigma_e^2$  的估计值, 如表 2.2:



表 2.2 单因素随机效应方差分析表

变异来源	SS	df	MS	EMS
组间	$SS_a$	$J-1$	$MS_a$	$n_0\sigma_a^2 + \sigma_e^2$
组内	$SS_e$	$N-J$	$MS_e$	$\sigma_e^2$
总变异	$SS_{\text{总}}$	$N-1$	—	—

注:  $\sigma_e^2 = MS_e$ ,  $\sigma_a^2 = \frac{1}{n_0}(MS_a - MS_e)$ ,  $n_0 = \frac{1}{J-1}(N - \frac{\sum n_i^2}{N})$

其中  $n_i$  为各个组内的数据个数,  $N$  为全部数据个数。

对表 2.1 资料作单因素随机效应的方差分析, 结果如表 2.3:

表 2.3 糖尿病患者随机效应方差分析表

变异来源	SS	df	MS	方差成分
组间	12.741	68	0.187	$n_0\sigma_a^2 + \sigma_e^2$
组内	182.779	1413	0.129	$\sigma_e^2$
总变异	195.520	1481	—	—

$$\hat{\sigma}_e^2 = 0.129, n_0 = \frac{1}{69-1}(1482 - \frac{21^2 + 25^2 + \dots + 23^2}{1482}) = 21.42$$

$$\hat{\sigma}_a^2 = \frac{0.187 - 0.129}{21.42} = 0.0027$$

因此  $\sigma_y^2 = \sigma_a^2 + \sigma_e^2 = 0.129 + 0.0027 = 0.1317$ ,  $ICC = 0.0027 / 0.1317 = 0.021$ ,

即总变异中 2.1% 的变异是由医生之间的变异引起的, 97.9% 的变异是由患者之间的变异引起的。

## 2.2 带协变量的随机效应方差分析模型

例 2.2 的研究并非只满足于了解病人对医生服务水平的满意程度的不同有多少是由于医生服务水平的差异所致。进一步的问题是:

问题一: 社区医生对病人的服务水平有无差别?

问题二: 病人对医生的评价是否受病人年龄、医生从业时间的影响?

问题一可转化为医生间的满意度平均得分有无差别, 可根据表 2.3 方差成分结果计算方差分析的  $F$  统计量:

$$F = \frac{n_0 \hat{\sigma}_a^2}{\hat{\sigma}_e^2} = \frac{0.0578}{0.14} = 0.448, \text{查 } F \text{ 界值表可得 } P > 0.05. \text{结论为, 可认为社区}$$

医生对病人的服务水平无差别。但如果用固定效应的单因素方差分析得  $F =$

$$\frac{MS_{\text{组间}}}{MS_{\text{组内}}} = 1.450, P < 0.05, \text{则结论相反。}$$