北京工业大学
“211工程”资助出版

线性混合效应模型引论

吴密霞 著

科学出版社

北京工业大学“211 工程”资助出版

线性混合效应模型引论

吴密霞　著

科学出版社

北京

内 容 简 介

本书系统阐述了线性混合效应模型的基本理论、方法和应用, 全书共12章. 第1章通过实例引进各种线性混合效应模型. 第2章讨论矩阵论方面的补充知识和线性模型的相关重要定理. 第3章讨论线性混合效应模型的固定效应的估计. 第4章讨论预测问题. 第5~9章系统讨论混合效应模型的方差分量的基本方法与相关理论, 包括: 方差分析估计、极大似然估计、限制极大似然估计、最小范数二次无偏估计、谱分解估计. 第10章讨论估计的最优性问题. 第11章讨论平衡数据情形下的混合效应模型的各种估计的统计性质. 第12章给出了混合效应模型下的假设检验.

本书可作为高等学校数学科学系、数理统计系或统计系、生物统计系、计量经济系等有关专业的高年级本科生及研究生的学位课或选修课教材. 同时可供数学、生物、医学、工程、经济、金融等领域的教师或科技工作者参考.

图书在版编目(CIP)数据

线性混合效应模型引论/吴密霞著. —北京: 科学出版社, 2013
ISBN 978-7-03-035558-4

I. ①线… II. ①吴… III. ①线性模型-高等学校-教材 IV. ①O212

中国版本图书馆 CIP 数据核字(2012)第 217934 号

责任编辑: 李 欣 赵彦超/责任校对: 宋玲玲
责任印制: 钱玉芬/封面设计: 陈 敬

科学出版社出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com

北京佳艺恒彩印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2013 年 1 月第 一 版 开本: B5 (720 × 1000)
2013 年 1 月第一次印刷 印张: 14 1/2
字数: 273 000

定价: 58.00 元

(如有印装质量问题, 我社负责调换)

总　　序

“211工程”是我国建国以来教育领域唯一的国家重点建设工程，面向21世纪重点建设一百所高水平大学，使其成为我国培养高层次人才，解决经济建设、社会发展和科技进步重大问题的基地，形成我国高等学校重点学科的整体优势，增强和完善国家科技创新体系，跟上和占领世界高层次人才培养和科技发展的制高点.

中国高等教育发展迅猛，尤其是1400所地方高校已经占全国高校总数的90%，成为我国高等教育实现大众化的重要力量，成为区域经济和社会发展服务的重要生力军.

在北京市委、市政府的高度重视和大力支持下，1996年12月我校通过了“211工程”部门预审，成为北京市属高校唯一进入国家“211工程”重点建设的百所大学之一.我校紧紧抓住“211工程”建设和举办奥运的重要机遇，实现了两个历史性的转变：一是实现了从单科性大学向以工科为主，理、工、经、管、文、法相结合的多科性大学的转变；二是实现了从教学型大学向教学研究型大学的转变. “211工程”建设对于我校实现跨越式发展、增强服务北京的能力起到了重大的推动作用,学校在学科建设、人才培养、科学研究、服务北京等方面均取得了显著的成绩，综合实力和办学水平得到了大幅度的提升.

至2010年底，我校的学科门类已经覆盖了8个：工学、理学、经济学、管理学、文学、法学、哲学和教育学. 现拥有8个一级学科博士学位授权点、37个二级学科博士学位授权点和15个博士后科研流动站，15个一级学科硕士学位授权点和81个二级学科硕士学位授权点；拥有6种类型硕士研究生专业学位授权资格，工程硕士培养领域19个；拥有3个国家重点学科、16个北京市重点学科和18个北京市重点建设学科.

目前，学校有专任教师1536人，全职两院院士5名，博士生导师220人，有正高职称294人和副高职称580人，专任教师中具有博士学位教师的比例达到54.6%. 有教育部“长江学者”特聘教授4人，国家杰出青年基金获得者6人，入选中组部“千人计划”1人，北京市“海聚工程”3人，教育部新(跨)世纪优秀人才支持计划15人.

2010年学校的到校科研经费为6.2亿元. “十一五”期间，学校承担了国家科技重大专项28项，“973计划”项目16项，“863计划”项目74项，国家杰出青年基金2项，国家自然科学基金重点项目8项、科学仪器专项2项、重大国际合作项目1项、面上和青年基金项目347项，北京市自然科学基金项目180项，获国家级奖励14项. 现有1个共建国家工程研究中心，7个部级或省部共建科研基地，11个北京市重点实验室和3个行业重点实验室.

为了总结和交流北京工业大学“211工程”建设的科研成果，学校设立了“211工程”专项资金，用于资助出版系列学术专著. 这些专著从一个侧面代表了我校教授、学者的学科方向、研究领域、学术成果和教学经验.

展望北工大未来，我们任重而道远. 我坚信，只要我们珍惜“211工程”建设的重要机遇，构建高层次学科体系，营造优美的大学校园，我校在建设国际知名、有特色、高水平大学的进程中就一定能够为国家、特别是为北京市的经济建设和社会发展做出更大的贡献.

中国工程院院士
北京工业大学原校长 左铁镛

2011年6月

前　言

近 20 年, 线性混合模型在生物、医学、经济、金融、环境科学、抽样调查及工程技术领域得到愈来愈广泛的应用. 因此, 线性混合效应模型的基础知识列入了国内外很多所高等院校的数理统计、生物统计、计量经济等专业的高年级本科生及研究生学习或研究的内容. 尽管国外这方面的书陆续出版了许多, 但国内目前几乎没有专门系统介绍混合效应模型的基本方法和相关研究的书. 在我们出版的《线性模型引论》中有一章专门介绍了线性混合效应模型的基本方法, 但限于篇幅, 未涉及相关方法的统计性质, 未能满足深入学习和研究的需要. 本书是为适应上述需要而编写的教材或教学参考书.

全书共分 12 章. 第 1 章介绍线性混合效应模型的相关概念、发展史以及模型形式. 第 2 章讨论矩阵论方面的补充知识和线性模型的相关重要定理. 第 3 章讨论线性混合效应模型的固定效应的估计. 第 4 章预测问题. 第 5~9 章系统讨论混合效应模型的方差分量的基本方法与相关理论, 包括: 方差分析估计、极大似然估计、限制极大似然估计、最小范数二次无偏估计、谱分解估计. 第 10 章讨论估计的最优性问题. 第 11 章讨论平衡数据情形下的混合效应模型的各种估计的统计性质. 第 12 章给出了混合效应模型下的假设检验.

借本书出版之际, 我要向我的恩师王松桂教授表示衷心的谢意, 特别感谢他多年来对我的科研工作给予的指导和鼓励. 同时也要感谢博士后期间的两位指导老师: 美国国家健康研究院Kai-Fun Yu研究员(现为清华大学教授)和Aiyi Liu研究员, 感谢中国科学院数学与系统科学研究院的王启华研究员, 感谢北京工业大学的杨振海教授、张忠占教授、王丽教授、薛留根教授、李寿梅教授、程维虎教授、陈立萍副教授等各位老师多年来给予我的大力支持和帮助. 在此也特别感谢我的爱人孙兵和女儿孙铭岳, 感谢他们一直以来给予我的支持、鼓励和无限的爱.

另外, 本书的写作和出版得到了国家自然科学基金(10801005, 11171011), 北京市自然科学基金(1102010), 北京市人才强教计划“青年骨干教师培养计划”项目(PHR20110820), 北京市优秀人才培养计划(PYZZ090421001156), 教育部留学回国人员科研启动基金, 以及北京工业大学“211工程”专著出版专项基金的资助. 编者愿借此机会表示诚挚的谢意.

由于编者水平所限, 书中错误或不当之处在所难免, 恳请广大读者不吝赐教.

编　者

2011年8月8日

目　　录

符　号　表

$\triangleq$	“定义为”或“记为”
$A \geqslant 0$	A 为对称半正定方阵
$A > 0$	A 为对称正定方阵
$A \geqslant B$	$A \geqslant 0, B \geqslant 0$ 且 $A - B \geqslant 0$
A^-	矩阵 A 的广义逆
A^+	矩阵 A 的Moore-Penrose广义逆
$\mathrm{rk}(A)$	矩阵 A 的秩
$\mid A \mid$	矩阵 A 的行列式
$\parallel A \parallel$	矩阵 A 的范数
$\mathrm{tr}(A)$	方阵 A 的迹
$\lambda_i(A)$	A 的第 i 个顺序特征根
$\mathcal{M}(\ \mathrm{A}\ )$	矩阵 A 的列向量张成的子空间
P_A	向 $\mathcal{M}(\mathcal{A})$ 的正交投影阵, 即 $P_A = A(A'A)^-A'$
Q_A	向 $\mathcal{M}(A)$ 的正交补空间上的正交投影阵, 即 $Q_A = I - P_A$
$\mathbf{1}'_n = (1, \cdots, 1)$	分量皆为1的 $n \times 1$ 列向量
$J_n = \mathbf{1}_n\mathbf{1}'_n$	元素皆为1的 $n \times n$ 列向量
$\mathrm{Vec}(\ A\ )$	将 A 的列向量依次排成的列向量
$\mathrm{Diag}(a_1, \cdots, a_n)$	对角元素分别为 $a_1, \cdots, a_n$ 的对角矩阵
$A \otimes B$	矩阵 A 与 B 的Kronecher乘积
$A \bigoplus B$	矩阵 A 与 B 的直和, 即 $\mathrm{Diag}(A, B)$
$E(X)$	随机变量或向量 X 的均值
$\mathrm{Var}(\ X\ )$	随机变量 X 的方差
$\mathrm{Cov}(\ X, Y\ )$	随机变量或向量 X, Y 的协方差
$u \sim N_p(\boldsymbol{\mu}, \Sigma)$	均值为 $\boldsymbol{\mu}$,协方差阵为 Σ 的 p 维正态向量
LS估计	最小二乘估计
BLU估计	最佳线性无偏估计
MVU估计	最小方差无偏估计
MINQU估计	最小范数二次无偏估计
RSS	回归平方和
SS_e	残差平方和
MSE	均方误差

第 1 章　模 型 概 论

线性混合效应模型是一类非常重要的统计模型. 在处理重复测量数据 (如纵向数据、Panel 数据)、区组数据以及空间相关数据时, 它具有独特的优势. 突破了多元分析中协方差阵无结构假设和线性模型下协方差阵除一个标量外完全已知的苛刻要求, 线性混合效应模型可以根据数据本身的结构特点, 较为灵活地选择其协方差阵的结构. 因此, 线性混合模型愈来愈得到了生物、医学、经济、金融、环境科学、抽样调查及工程技术等领域研究人员的广泛关注和应用. 本书将系统介绍线性混合效应模型统计推断的基本理论和方法.

本章主要简单介绍线性混合效应模型的相关概念、发展简史、模型的基本形式以及相关研究. 目的是让读者对该模型的丰富实际背景以及相关研究有一些了解, 这将有助于对后续章节的理解.

1.1　因子、水平与效应

在试验设计中, 因子是指影响研究变量的各个原因. 水平是指一个因子的不同状态. 效应是指该因子的各水平对所研究变量的影响. 而效应又可分为固定效应和随机效应. 当一个因子的水平是一个有限集合, 而且试验感兴趣的就是水平效应本身时, 该因子的效应为固定效应; 当因子的水平是一个无限集合, 数据中该因子出现的水平仅是这个无限集合的一个随机样本时, 该因子的效应为随机效应. 下面将通过实例来更详细地解释因子、水平和效应的概念.

例1.1.1　小麦品种比较试验.

假定某个农业试验基地引进 a 种小麦品种, 在进行大面积种植前, 先要进行小范围的试验种植, 从中挑选出最适合本地区的优良品种. 这时, 小麦品种就是主要感兴趣的一个因子. 每个具体的品种就是小麦品种这个因子的一个水平. 为了消除土质对产量的影响, 在试验种植时, 首先需找到土质肥沃程度基本一样的一大块田, 将其分成面积相等的 n 块, 其中 n_i 块用来种植第 i 种的小麦, $i=1,\cdots,a$, 这里 $\sum_{i=1}^{a} n_i = n$. 用 y_{ij} 表示第 i 种的小麦在它的第 j 块田上的产量, 则 y_{ij} 可表为

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1,\cdots,a, \quad j = 1,\cdots,n_i, \tag{1.1.1}$$

这里 α_i 为小麦的第 i 品种对产量的影响, 即水平 i 的效应. 我们的感兴趣的是比较这固定的 a 个小麦品种的产量, 故 α_i 为固定效应. 模型 (1.1.1) 就是单向分类模型.

一般情况下, 很难找到土质肥沃程度完全一样的一大块田. 考虑到土质对产量的影响也是不可忽视的, 一般从若干试验田中选取土质肥沃程度较均匀、面积相等的 b 块. 在试验设计中, 把这种块称为区组 (block). 然后再将每块等分成 a 小块, 称为试验单元. 每个试验单元上种植一种小麦. 用 y_{ij} 表示第 i 种的小麦在第 j 区组上的产量, 则 y_{ij} 可表为

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, b. \tag{1.1.2}$$

这里 α_i 与模型 (1.1.1) 中的假设相同, β_j 为第 j 个区组对产量的影响, 即区组 j 的效应. 由于区组非随机选取的, 故 β_j 也为固定效应. 模型 (1.1.2) 就两向分类模型.

例1.1.2 血压数据: 研究人的血压在一天内的变化规律.

在一天内选择 a 个固定时间点 $(T_1, \cdots, T_a)$ 测量被观测者的血压. 假设观测了 b 个人. 这时, 时间和个体就是影响血压的两个因子. 时间点 $T_1, \cdots, T_a$ 就是时间因子的水平, b 个被观测者就是个体因子的水平, a 和 b 分别为时间因子和个体因子的水平数. 用 y_{ij} 表示第 i 个时间点的第 j 个人的血压, 则 y_{ij} 可表为

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, b. \tag{1.1.3}$$

其中, α_i 为第 i 个固定时间点的对血压的影响, 称其为第 i 个固定时间点的效应, 同样称 β_j 为第 j 个人的个体效应. 由于观测的时间点是固定的, 而且感兴趣的正是这固定时间点的人的血压变化, 故 α_i 是固定效应. 关于个体效应 β_j 是否是固定效应, 可以依据如下规则来判断: 如果这 b 个人是感兴趣的特定的 b 个人, 那么 β_j 也是固定效应; 如果我们的研究兴趣只是放在比较不同时间点人的血压高低上, 而被观测的 b 个人是随机抽取的, 这时 β_j 就是随机变量, 称其为随机效应. 进一步, 如果我们关心的问题仅是人一天中的平均血压, 测量的 a 个时间点和被观测的 b 个人都是随机抽取的, 此时 α_i 和 β_j 都是随机效应.

从上面的讨论可以看出,一个效应究竟看做随机的还是固定的, 这取决于研究的目的和样品取得的方法. 如果观测的个体是随机抽取来的, 那么它们的效应就是随机的, 否则就是固定的. 依据模型中所含效应的形式, 可将模型分为固定效应模型、随机效应模型以及混合效应模型. 当数据只包含固定效应的因子时, 则称基于该数据的模型为固定效应模型, 如模型 (1.1.1) 和 (1.1.2). 当数据只包含随机效应的因子时, 则称基于该数据的模型为随机效应模型, 如模型 (1.1.3) 中 α_i 和 β_j 都是随机效应情形. 当数据既包含固定效应的因子又包含随机效应的因子时, 则称基于该数据的模型为混合效应模型, 如模型 (1.1.3) 中 α_i 是固定效应, β_j 是随机效应的情形.

1.2 线性混合效应模型的发展简史

线性混合效应模型最早是由 Airy (1861) 提出的, 距今已有 150 多年的发展历史. 由最初很难被理解到今天被广泛应用, 这归功于几代统计学家在该模型的理论上的突破以及近 20 多年来在相应统计软件的支持. 为了更好地理解后面章节的内容, 本节将主要依据 Searle 等 (1992) 的概括, 介绍线性混合效应模型在理论发展史上的里程碑.

Airy (1861) 在研究天文望远镜的观测数据时. 视同一天晚上对某个天体的几次观测是一个组数据 (clustered data), 将不同天晚上的观测数据放在一起分析, 建立了单向分类随机模型

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, a, \quad j = 1, \cdots, n_i, \tag{1.2.1}$$

这里 μ 为总体均值 (或真值), α_i 和 ϵ_{ij} 随机的. Airy 称 α_i 为常数误差 (constant error), 即针对固定的第 i 晚 α_i 为常数, 这就是第 i 晚的效应, 它是由第 i 晚的特定空气和个人因素引起的. ϵ_{ij} 是第 i 晚条件均值 $\mu + \alpha_i$ 的随机误差. 假设 $\{\epsilon_{ij}\}$ 是独立同分布的, $\{\alpha_i\}$ 独立同分布的, 且 $\{\alpha_i\}$ 和 $\{\epsilon_{ij}\}$ 相互独立, 均值都为零, 方差分别记为 σ_α^2 和 σ_ϵ^2, 称 σ_α^2 和 σ_ϵ^2 为方差分量. Airy首先由第 i 晚数据得到 σ_ϵ^2 的估计

$$\widehat{\sigma}_{\epsilon,i}^2 = \frac{\sum\limits_j (y_{ij} - \bar{y}_{i\cdot})^2}{n_i - 1}, \quad i = 1, \cdots, a, \tag{1.2.2}$$

然后利用这 a 个 $\widehat{\sigma}_{\epsilon,i}^2$ 的平方根的平均值给出了 σ_ϵ^2 的一个估计

$$\widehat{\sigma}_\epsilon^2 = \left[\sum_{i=1}^{a} (\widehat{\sigma}_{\epsilon,i}^2)^{\frac{1}{2}} / a \right]^2, \tag{1.2.3}$$

其中 $\bar{y}_{i\cdot} = \sum_j y_{ij}/n_i$. 值得注意的是(1.2.3)的分子就是现在文献中所指的第 i 组的组内方差, 分母为其自由度.

依据 Scheffé (1956) 的研究, 文献中第二次使用随机效应模型的人为 Chauvenet (1863). 在平衡数据 ($n_i = n$) 下, 他给出模型 (1.2.1) 的样本均值 $\bar{y}_{\cdot\cdot} = \sum_i \sum_j y_{ij}/na$ 的方差为

$$\mathrm{Var}(\bar{y}_{\cdot\cdot}) = \frac{\sigma_\alpha^2 + \sigma_\epsilon^2/n}{a}.$$

但 Airy (1861) 和 Chauvenet (1863) 都没给出 σ_α^2 的估计.

对方差分量的估计做出的奠基工作是 Fisher (1925), 他在 *Statistical Methods for Research Workers* 书中概括了在平衡数据情形下, 如何根据不同来源将随机变

动进行分解来估计方差分量的一般方法. 该方法的思想是在估计完全随机化设计的数据的组内相关系数时体现出来的. 为方便起见, 这里采用模型 (1.2.1) 中的符号来表述 Fisher (1925) 的思想. 该模型下组内相关系数可表示为

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}.$$

因此, 只要得到 σ_α^2 和 σ_ϵ^2 估计就可以得到 ρ 的估计. Fisher (1925) 提出可以直接由组内变差来估计 σ_ϵ^2, 即

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2 = a(n-1)\sigma_\epsilon^2. \tag{1.2.4}$$

由于任意组观测值平均值的方差都包含两部分, 一部分是 σ_α^2, 另一部分为 σ_ϵ^2/n, 故由组间均值的变差有

$$n\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{i\cdot})^2 = (a-1)(n\sigma_\alpha^2 + \sigma_\epsilon^2). \tag{1.2.5}$$

结合 (1.2.4) 便可得到 σ_α^2 的估计.

Tippett (1931) 将 Fisher 的工作首次推广到了无交互效应的两向分类随机模型的情形, 并将 Fisher 提出的估计方差分量的方法定义为方差分析 (analysis of variance, ANOVA) 方法. 此外, Tippett 的另一个贡献是采用线性模型的框架, 将数据中由多个随机因子引起的随机变动用其线性函数的形式引入模型, 给出了随机效应模型的一般表达形式.

Neyman 等在比较随机区组设计和拉丁方设计的相对效率时, 扩展性地使用了混合模型, 并首次引入了“方差分量 (variance component)”这个术语. Jackson (1939) 提出了一个两因子的混合模型, 其中一个因子是随机的, 另一个因子是固定的, 即今天我们所熟悉的不带交互效应的两向分类混合模型

$$y_{st} = A + B_s + C_t + z_{st},$$

这里 A 是对全部个体的一般效应的度量, B_s 是对试验效应的度量, C_t 是对个体效应的度量, z_{st} 是测量误差, 并假设随机效应和误差具有正态分布. 在此模型描述中“效应(effect)”这个术语首次被引入. 不过当时该模型并没有被称为混合模型 (mixed model). “混合模型”这个术语直到 Eisenhart (1947) 才被引入, 固定效应模型和随机效应模型 (即 Eisenhart 模型 I 和 Eisenhart 模型 II) 才正式被区分开来.

由于数据并非总是平衡数据, 事实上, Airy 所考虑的模型 (1.2.1) 就是不平衡数据情形. 然而从 Fisher 在平衡数据情形下提出了 ANOVA 方法之后, 统计界关于非平衡数据下方差分量的估计问题的研究就很少了, 直到 Cochran (1939) 该类

问题的研究在文献中才再次出现. 在该文献中, Cochran 考虑了不平衡数据下单向分类随机效应模型 (1.2.1), 但没有特别关注方差分量的估计问题. Winsor 和 Clarke (1940) 给出了模型 (1.2.1) 下, 随机效应方差 σ_α^2 和误差方差 σ_ϵ^2 的估计, 即可由下式给出

$$E\sum_{i=1}^{a} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2=(a-1)\left[\frac{\sum\limits_{i=1}^{a} n_i-\sum\limits_{i=1}^{a} n_i^2/\sum\limits_{i} n_i}{a-1}\sigma_\alpha^2+\sigma_\epsilon^2\right], \tag{1.2.6}$$

$$E\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{ij}-\bar{y}_{i\cdot})^2=\left(\sum_{i=1}^{a} n_i-a\right)\sigma_\epsilon^2, \tag{1.2.7}$$

这里 $E(\cdot)$ 是求期望的运算符号.

Anderson 和 Bancroft (1952) 出版了 *Statistical Theory in Research*, 在此书中, 他们全面总结了当时平衡数据下方差分量的估计方法, 并介绍了不平衡数据下套结构随机效应模型的方差分量估计方法. Henderson (1953) 在不平衡数据下分别针对随机效应模型、混合效应方差分析模型以及一般的方差分量模型提出了三种的方差分量的估计方法, 这便是著名的 Henderson 三方法. 至此, ANOVA 方法才真正成为了一种成熟的估计方差分量的方法.

此外, Henderson 在线性混合效应模型发展中作出的另一个大的贡献是引进一个著名的混合模型方程组, 参见 Henderson (1950), 该方程组形式上类似于正则方程组, 然而它却能同时给出固定效应可估函数的最佳线性无偏 (best linear unbiased, BLU) 估计, 以及随机效应的最优线性无偏预测(best linear unbiased prediction, BLUP). 由于在线性混合效应模型下的 BLU 估计和 BLUP 往往依赖于未知的方差分量, 故在实际应用中它们所包含的方差分量需要用其估计来替代. 因此方差分量的估计问题是一个非常重要的问题, 关于它的研究在线性混合效应模型发展史上占了很大比重.

对于 ANOVA 估计的性质的研究, Graybill 和 Wortham (1956) 指出在平衡数据的随机效应模型下方差分量的 ANOVA 估计为最小方差无偏估计 (minimum variance unbiased, MVU); Graybill 和 Hultquist (1961) 给出了一般的随机效应模型下方差分量的 ANOVA 估计为 MVU 的充要条件; Albert (1976) 研究了 ANOVA 估计为最佳二次无偏估计和 MVU 估计的条件; Brown (1978, 1984) 考虑了 ANOVA 估计的不变最小方差无偏 (minimum variance invariant unbiased, MVIU) 估计的性质. Wald (1940) 在构造方差分量比的置信区间时, 发现了方差分量的 ANOVA 估计不唯一. Ganguli (1941) 和 Crump (1946) 分别将 ANOVA 方法应用于随机效应带套误差结构的混合效应模型和带交互效应的混合效应模型中, 他们注意到 ANOVA

估计可能取负值这个缺陷.

由于方差分量的 ANOVA 估计可能会取负值和不一定唯一这两个不可避免的缺点, 故寻找新的估计方法成为了当时一个热点问题. 随之出现了极大似然 (maximum likelihood, ML) 方法、限制极大似然 (restricted maximum likelihood, REML) 方法、最小范数二次无偏 (minimum norm quadratic unbiased, MINQU) 估计方法.

ML方法最早出现于 Crump (1947, 1951) 估计平衡数据和非平衡数据下单向分类随机模型的方差分量. 但该方法真正成为一种成熟的估计方法要归功于 Hartley 和 Rao (1967). 他们基于模型的矩阵表达, 得到了似然方程组, 将极大似然方法发展成为适用于各类线性混合效应模型在平衡数据情形或着非平衡数据情形下的估计方法. 然而由于 ML 估计的计算首先涉及协方差阵 $V(\sigma^2)$ 的求逆运算和行列式运算, 故其计算量比较大. 应用平衡随机效应的设计阵的特殊结构, Smith 和 Hocking (1978), Searle 和 Henderson (1979) 给出了平衡数据下 $V(\sigma^2)$ 的逆和行列式以及特征根的表达式, 大大简化了计算. Szatrowski (1980) 和 Szatrowski 和 Miller (1980) 给出了方差分量模型 ML 估计存在显式表达式的充要条件.

ML 估计克服了 ANOVA 方法两个弱点. 但该方法在当时也存在一个使用上的限制: ML 估计往往没有显式表达形式, 需要迭代计算. 此外, 在导出方差分量的 ML 估计的过程中, 极大似然方法并没有考虑由估计固定效应所引起的自由度的减少, 从而使得 ML 估计是有偏估计.

Patterson 和 Thompson (1971) 提出了 ML 估计的一种修正估计: 限制极大似然 (REML) 估计, 即基于原模型的最小二乘 (least squares, LS) 估计残差的分布求方差分量的 ML 估计, 故该估计也被称为边缘极大似然估计, 对模型的固定效应参数具有不变性.

与 ML 估计相比, REML 估计的偏差减少很多, 且对于许多平衡混合方差分析模型, REML 估计方程组的解与方差分析估计相等. 但在通常情况下, REML 估计方程组的求解仍依赖于迭代算法, 其迭代的收敛性问题依然存在.

Rao (1970 ~ 1972) 提出了最小范数二次无偏 (minimum norm quadratic unbiased, MINQU) 估计. 本质上类似于线性模型下固定效应的 LS 估计, MINQU 估计也是通过极小化一个二次范数得到的. 与 ANOVA 估计类似, MINQU 估计仅依赖于随机效应和随机误差的一阶矩和二阶矩, 不需要假设它们的分布, 也不需要迭代, 可通过求解一个线性方程组得到. 但缺点是依赖于先验值, 故受主观影响比较大. 此外, MINQV 估计与 REML 估计也有很密切的关系, 只要将 MINQU 估计的方程组中方差分量的先验值换成方差分量本身就得到了 REML 估计的似然方程组.

线性混合效应模型发展史上的另一个重要的里程碑是 Laird 和 Ware (1982) 提出了混合效应模型的更一般形式

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + U_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \cdots, m, \tag{1.2.8}$$

这里, $\boldsymbol{y}_i$ 是个体单元上的因变量的观测向量, X_i 和 U_i 分别是 $n_i \times p$ 和 $n_i \times q$ 的设计矩阵, $\boldsymbol{\beta}$ 为 $p \times 1$ 的总体的未知参数, $\boldsymbol{b}_i$ 为 $q \times 1$ 的未知的个体效应, 是随机的, $\boldsymbol{\varepsilon}_i$ 是随机误差. 假设 $\boldsymbol{b}_i \sim N_q(\boldsymbol{0}, D_0)$, $\boldsymbol{\varepsilon}_i \sim N_{n_i}(\boldsymbol{0}, R_i)$, 且 $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_m, \boldsymbol{\varepsilon}_1, \cdots, \boldsymbol{\varepsilon}_m$ 相互独立, 其中 D_0 为 $q \times q$ 的半正定矩阵, R_i 为 $n_i \times n_i$ 的正定矩阵. 为简单起见, 应用中通常又假设 $R_i = \sigma_{\varepsilon}^2 I_{n_i}$.

近 20 年来, 线性混合效应模型被越来越广泛地应用到生物、医药、经济、金融、气象、社会科学等应用领域的数据研究中, 成为分析遗传数据、纵向数据 (longitudinal data)、面板数据 (panel data) 等各类重复测量数据的重要的模型之一. 如在 Baltagi (1995) 编著的*Panel Data*、朱军(1997) 编著的《遗传模型分析方法》、Verbeke 和 Molonberghs(2007) 编著的*Linear Mixed Models for Longitudinal Data*、Diggle 等 (2002) 编著的*Analysis of Longitudinal Data* 以及 Demidenko (2004) 编著的 *Mixed Models: Theory and Applications* 等许多著作中, 线性混合效应模型都是其数据分析最为重要的一类模型. 此外, 在线性混合效应模型的理论研究方面, 也取得到许多新的成果和推断方法, 如 SD 估计、广义 P-值检验、EM 算法等.

1.3 模型形式

线性混合效应模型是一类应用非常广泛的模型, 可以用来分析处理纵向数据 (longitudinal data)、组数据和面板数据 (panel data) 等各类重复测量数据. 突破了传统线性模型索要求的观测值是彼此独立且等方差的条件限制, 线性混合效应模型对观测值的协方差阵的结构有了更加灵活的假设. 如针对组数据, 通常可以假设同一组内数据是相关的, 不同组数据之间是独立的; 针对多层分组数据, 线性混合效应模型可以通过引进多水平的随机效应给出观测值的协方差阵的一个简单且合理结构假设. 本节主要给出混合效应模型最一般的表达形式, 以及具有广泛应用的各类混合效应模型之间的结构关系.

混合效应模型最一般的表达形式为

$$\boldsymbol{y} = X\boldsymbol{\beta} + U\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{1.3.1}$$

这里 $\boldsymbol{y}$ 是 $n \times 1$ 观测向量, X 为 $n \times p$ 已知设计阵, U 为 $n \times q$ 设计阵矩阵, $\boldsymbol{\beta}$ 为 $p \times 1$ 未知参数向量, $\boldsymbol{\xi}$ 为 $q \times 1$ 随机向量, $\boldsymbol{\varepsilon}$ 为 $n \times 1$ 随机误差向量. 假设 $E(\boldsymbol{\xi}) = 0$, $\mathrm{Cov}(\boldsymbol{\xi}) = D$, $E(\boldsymbol{\varepsilon}) = 0$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = R$, 且 $\boldsymbol{\varepsilon}$ 与 $\boldsymbol{\xi}$ 独立, 其中, D 为非负定矩阵, R 为正定矩阵. 则 $\boldsymbol{y}$ 的协方差阵为

$$\mathrm{Cov}(\boldsymbol{y}) = UDU' + R.$$