

多层模型实践

基于 Tennessee STAR 班额实验的方法演练与 R 语言实现

作者：晨瀚宇（Chanw）

时间：May 22, 2026

版本：1.0

同一份数据、同一个教育问题，从 OLS 到三层 HLM 的完整演练与比较。

© 2026 晨瀚宇. 保留所有权利。

版权声明

书名：多层模型实践——基于 Tennessee STAR 班额实验的方法演练与 R 语言实现

作者：晨瀚宇（小红书：Chanw）

版本：v1.0（2026 年 5 月）

© 2026 晨瀚宇. 保留所有权利。

本书内容（包括但不限于文字、代码、图表、排版设计）均为作者原创。

禁止事项：

- 未经作者书面授权，禁止以任何形式转载、复制、传播本书的全部或部分内容
- 禁止将本书内容用于商业用途或二次售卖
- 禁止去除、修改本书中的版权标识和作者署名

允许事项：

- 个人学习和研究用途的合理引用（需注明出处）
- 在社交媒体上分享本书的购买链接

如发现侵权行为，请联系作者。

购买渠道：小红书搜索「Chanw」

前言

这本书的写作思路是：选定一份具体的数据集，把所有相关多层建模方法都在它身上跑一遍，让读者亲手看到每一种方法在同一个问题上能给出什么答案、各自的假设和局限在哪里。

本书用的数据是 mlmRev 包内置的 *star*，1985 至 1989 年间美国田纳西州 Project STAR 班额实验的真实记录，11,598 名学生嵌套在 1,387 个班级、80 所学校里，回答的核心问题是把班级人数从 23 人减到 15 人能不能让小学生考得更好。这份数据是多层建模教学的事实标准样本，故事性强，处理变量随机分配，三层嵌套结构干净，适合把每种方法的表现差异都展示出来。

读者画像是在读研究生与青年研究者，能用 R，看过几篇用了多层模型的论文，但没有亲手跑过完整流程的那一批。本书不打算把 Raudenbush & Bryk 的圣经和 Snijders & Bosker 的教材重复一遍，更不打算把所有方法都讲到学术教科书的精度。这本书的野心比较有限：让读者跑通一项 STAR 级别的多层分析、看懂 `lmer()` 输出、知道每个数字背后的假设是什么、知道踩坑的常见地方在哪儿。读完之后愿意去翻 Hox 的《Multilevel Analysis》第 4 章作为参考、愿意去看 lme4 的官方插图把更复杂的情形深入处理，这本书的任务就完成了。

晨瀚宇
2026 年 5 月

目录

前言 i

第 1 章 问题与数据：Tennessee STAR 班额实验 1

1.1 把班级人数从 23 减到 15，孩子真的能考得更好吗 . . . 1
1.2 一所学校的数据结构：学生在班级在学校 . . . 2
1.3 班级类型与考试分数：朴素均值对比 . . . 2
1.4 学校层面的差异：一份数据的” 伏笔” . . . 3
1.5 全书路线图：从朴素 OLS 到三层 HLM . . . 4

第 2 章 嵌套数据与组内相关 ICC 5

2.1 同班学生的相似性：把” 独立观测” 假设打开看 . . . 5
2.2 组内相关 ICC：班级里的” 发言权” 分到哪一层 . . . 6
2.3 设计效应 DEFF：实际样本量被折算成多少 . . . 7
2.4 用 lme4 跑随机截距模型：朴素 OLS 与 HLM 标准误对比 . . . 8
2.5 从一个 SE 到方差分解的全图 . . . 9

第 3 章 空模型：随机截距与方差分解 10

3.1 空模型的方程与每个符号背后的人 . . . 10
3.2 BLUP 与收缩：班均值被全样本均值” 拉” 回 . . . 11
3.3 REML vs ML：两种估计方法在多层模型里的分工 . . . 13
3.4 似然比检验：空模型相对朴素 OLS 是否值得 . . . 14
3.5 从空模型到学生层协变量 . . . 15

第 4 章 学生层协变量与组内组间分解 16

4.1 逐步加入学生层协变量：方差分量的变化 . . . 16
4.2 组均值中心化：把” 班内效应” 与” 班间效应” 分开 . . . 18
4.3 为什么不分解会出问题：组内组间被压成单一系数 . . . 19
4.4 学生层之上还有班级层与学校层 . . . 20

第 5 章 班级层协变量与处理效应 21

5.1 班级层固定效应：把处理变量放在哪一层 . . . 21
5.1.1 先看一个生活类比：班里发助教就像班里多请一个家教 . . . 21
5.1.2 班级层固定效应的形式定义 . . . 21
5.1.3 多层模型里识别处理效应的关键：随机分配在哪一层 . . . 22
5.1.4 学术背景：为什么 STAR 是教育统计的奠基样本 . . . 22
5.2 从 M3 到 M4：把 cltype 放进随机截距模型 . . . 22
5.2.1 为什么先单独跑 M4 再加学生层协变量 . . . 22
5.2.2 对比朴素 OLS 与 HLM：点估计稳，标准误膨胀 . . . 23
5.3 班级层方差被 cltype 解释了多少 . . . 24
5.4 班级 BLUP 的变化：班际差异在解释什么之后还剩多少 . . . 25
5.5 累积方法对比表（首次建立） . . . 26
5.6 下一章预告 . . . 26

第 6 章 随机斜率与协变结构 **27**
6.1 随机斜率：让协变量的影响在组间变化 . . . 27
6.1.1 先看一个生活类比：每个班对低 SES 学生的” 关照度” 不同 . . . 27
6.1.2 随机斜率的形式定义与几何含义 . . . 27
6.1.3 随机效应协方差矩阵 G：斜率与截距相关吗 . . . 28
6.1.4 学术背景：Goldstein 1986 的随机系数模型 . . . 28
6.2 M7 与 M8：让 SES 斜率随班级变化 . . . 28
6.3 方差成分检验的雷区：边界问题与 p 值减半 . . . 30
6.4 随机斜率不显著的解读：为什么这是个有用的负面结论 . . . 31
6.5 累积方法对比表更新 . . . 31
6.6 下一章预告 . . . 32

第 7 章 跨层交互 **33**
7.1 跨层交互：班级层变量乘以学生层变量 . . . 33
7.1.1 先看一个生活类比：小班对学习困难学生有没有特别的好处 . . . 33
7.1.2 跨层交互的形式定义 . . . 33
7.1.3 学术背景：Krueger & Whitmore 2001 的差异化效应分析 . . . 34
7.2 M9：含跨层交互的随机截距模型 . . . 34
7.2.1 用四种组合的预测均值看清楚结论 . . . 35
7.3 交互项不显著的雷区：检验功效低不等于不存在 . . . 37
7.4 累积方法对比表更新 . . . 37
7.5 下一章预告 . . . 38

第 8 章 三层模型：学生–班级–学校 **39**
8.1 三层嵌套的结构：学生、班级、学校 . . . 39
8.2 嵌套随机效应与三层模型形式 . . . 39
8.3 三层空模型：把方差拆成学校 / 班级 / 学生三块 . . . 40
8.4 学校层随机截距 BLUP 的分布 . . . 41
8.5 加上协变量与处理变量：三层模型的完整版 . . . 42
8.6 LRT 检验：三层 vs 二层的模型对比 . . . 43
8.7 累积对比表更新 . . . 44
8.8 从主估计到模型诊断的过渡 . . . 44

第 9 章 模型诊断与影响分析 **46**
9.1 多层模型的分层残差结构 . . . 46
9.2 学生层残差：QQ 图与方差稳定性 . . . 47
9.3 班级层与学校层 BLUP 的诊断 . . . 49
9.4 leave-one-class-out 敏感性分析 . . . 49
9.5 影响力诊断与班级权重 . . . 51
9.6 累积对比表更新 . . . 52
9.7 从诊断到方法对比的过渡 . . . 52

第 10 章 方法对比与结论 **53**
10.1 七种方法的终极对比 . . . 53
10.2 模型选择：AIC / BIC / LRT . . . 55
10.3 方法选择决策树 . . . 57

10.4 在 STAR 数据上的最终结论 58
10.5 把多层建模工作流搬到新数据 58

Bibliography **60**

第 1 章 问题与数据：Tennessee STAR 班额实验

内容提要

- 看清”小班 vs 常班”在 5,786 名学生身上给出的真实差距
- 理解学生—班级—学校三层嵌套结构的具体形态
- 学会读 STAR 数据的 2x2 描述表，每一格代表的是哪一群学生
- 理解为什么把学生当独立观测做朴素 OLS 会高估”小班效应”的把握度

1.1 把班级人数从 23 减到 15，孩子真的能考得更好吗

1985 年，美国田纳西州议会批准了一项规模罕见的教育实验。州政府花了约 1,200 万美元，在全州 79 所小学里把 11,000 多名学生和 1,300 多位老师全部随机分配进三种班级里，连续观察四年，看小班教学是不是真的有效。这就是 Project STAR，全称 Student/Teacher Achievement Ratio。

实验设计本身极为干净。学生入学时被随机分到三种班里。第一种是**小班**（small class，13–17 人），第二种是**常班**（regular class，22–25 人），第三种是**常班 + 助教**（regular plus aide，22–25 人但配一名教学助理）。老师也是随机分配到这三种班里，和学生一起从幼儿园 K 年级开始，连续到三年级。每年学生考一次 Stanford Achievement Test，记录数学和阅读分数。

这项研究后来成了教育政策与多层建模教学的奠基样本。它一次性回答了两个问题：第一，小班教学的因果效应有多大；第二，学生分数的差异里有多少是来自老师、班级、学校的”上层”波动，而不是单个学生的随机起伏。前一个问题给政策制定者方向，后一个问题给统计学家提出了”嵌套数据怎么分析才算诚实”这个根本难题。

为了让”嵌套到底是个什么东西”不只是一个抽象说法，先看本书全程要用的这份数据的真实结构：

- 5,786 名 K 年级学生，分散在 337 个班级里。每个班级由一位老师带队。
- 这 337 个班级分散在 79 所小学里。每所小学带 4 个班，平均班里 17 人。
- 处理变量 `cltype`（班级类型）随机分配在班级层面：30% 的班是小班，35% 是常班，35% 是常班 + 助教。

学生不是独立的样本点，他们三层嵌套——**学生在班级里、班级在学校里**。这个事实将贯穿全书：每一章用一种新方法，让这个事实在统计推断里得到尊重。

图 1.1 把这种三层嵌套画出来。

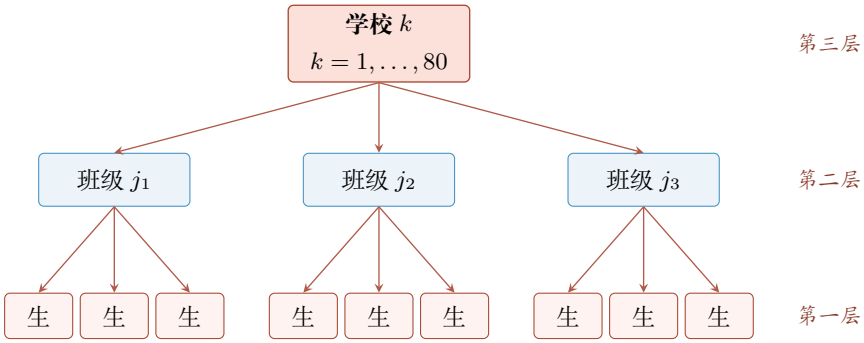


图 1.1: Tennessee STAR 数据的三层嵌套结构：80 所学校在最高层，每所学校下若干班级，每个班级下若干学生。学生不是独立样本，同班学生共享教师和班级环境，同校班级共享学校政策与生源，传统 OLS 把所有学生当作独立观测会低估标准误。

到了 1990 年代末，Bates 和 Pinheiro 把混合效应模型的 R 工具包 nlme 推广开来，后来 Bates 又写了 lme4，

把 STAR 之类的多层数据分析变成了一行 `lmer()` 调用 [1]。这份数据被收进 mlmRev 包，称为 *star*，成为多层建模教学的事实标准样本，相当于因果推断教学里的 RHC、回归分析里的 mtcars。**本书全书都用这一份数据**。每一章用一种方法处理这同一份数据，到第 10 章把所有结论汇总比较。

1.2 一所学校的数据结构：学生在班级在学校

每所参加 STAR 的小学做的事其实都一样。学校把当年要入学的 K 年级新生集中起来，按 **完全随机**的方式分配进 4 个班里，其中 1 个是小班，2 个是常班（一个带助教一个不带），剩下 1 个再补一种。老师也按随机方式分配。每年 5 月做一次标准化测试，分数交给田纳西大学的统计员录入。

整所学校的核心结构可以装进一张三行小表，行业通称**嵌套层级表**。表 1.1 用 K 年级数据给出全样本的真实数字。

表 1.1: STAR K 年级数据的三层结构（真实数字）

层级	单位数	平均嵌套子数	符号
学校（school）	79 所	约 4 个班 / 校	j（学校索引）
班级（class / teacher）	337 个	约 17 名学生 / 班	k（班级索引）
学生（student）	5,786 名	—	i（学生索引）

读这张表的方法是从下往上一层层套。最底下一层是**学生**，每名学生留下一份分数 (y_{ijk})。中间一层是**班级**，平均 17 名学生坐在同一个班里、由同一位老师教。最上面一层是**学校**，平均 4 个班同属一所小学，共享同一个校长、同一套课程编排、同一份社区生源。

为什么要这样画这张表？因为本书后面所有多层模型方法，本质上都从这三层结构开始。**5,786 名学生 = 337 个班 = 79 所学校**。lme4 包用 `(1|sch/tch)` 的语法表示”班级嵌套在学校里”，背后正是表 1.1 的层级关系。本书后续章节里出现的索引 i, k, j 都回到这张图。

停下来想一想。如果你能告诉别人某名学生的 i, k, j 三个索引各是多少，你就完整定位了这名学生在数据里的位置。同班同学共享同一个 k，同校同学共享同一个 j。这就是多层建模整理研究数据的最小单位。

1.3 班级类型与考试分数：朴素均值对比

知道了三层结构，怎么算”小班到底有没有用”？

最直觉的做法是看三种班级类型的**平均分数**，每个班里所有学生的数学分数加起来除以人数。三类的真实数字列在表 1.2。

表 1.2: K 年级三类班级的朴素均值（math 分数）

班级类型	班级数	学生数	平均班级人数	数学均值（SD）
小班 small	131	1,738	13.3	491.4 (49.4)
常班 reg	103	2,005	19.5	483.3 (47.5)
常班 + 助教 reg+A	103	2,043	19.8	482.9 (45.9)

最直观的对比是小班减去常班：$491.4 - 483.3 = 8.1$ 分。换句话说，分到小班的孩子平均比分到常班的孩子多考 8.1 分。常班 + 助教那一类几乎和不带助教的常班一样 $(482.9 - 483.3 = -0.4)$，助教这个变量看起来没用。

用一个完全不需要教育学背景的例子理解这个差距：100 个学生考 0 到 800 分之间的标化测试，全样本的标准差大约是 47 分，所以 8 分相当于 0.17 个标准差。这是个不大但也不能忽视的差距，相当于把一个班的中等水平推到中等偏上。

定义 1.1 (小班效应)

对一份随机分配处理的数据，小班效应定义为

$$\hat{\Delta}_{\text{small-reg}} = \bar{y}_{\text{small}} - \bar{y}_{\text{reg}},$$

其中 $\bar{y}_{\text{small}}$ 是分到小班的学生数学均值，$\bar{y}_{\text{reg}}$ 是分到常班的学生数学均值。它衡量”把孩子从常班调进小班”在分数上能换来多少提升。

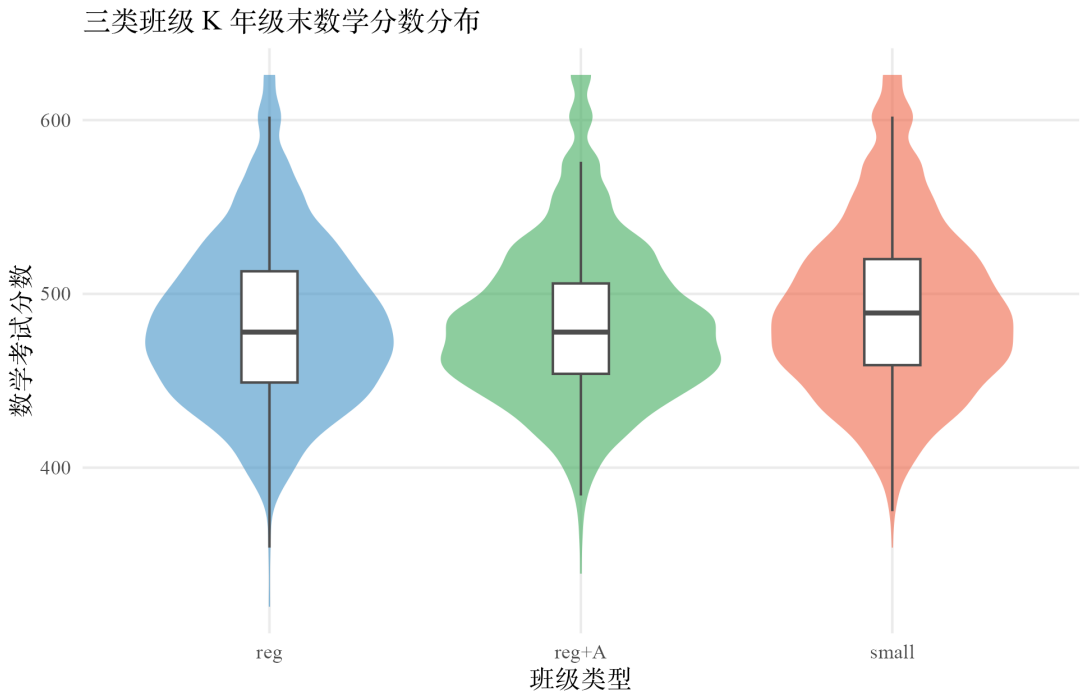


图 1.2: 三类班级 K 年级末数学分数分布。小班分布整体右移，但三类的尾巴大部分重叠

朴素估计 $\hat{\Delta}_{\text{small-reg}} = 8.08$ 分这个数字看起来直接、可复现，但它隐藏了一个严重的统计问题：**它假设 5,786 名学生互相独立**。同一个班的 20 个学生共享同一位老师、同一个教室、同一个学习氛围，他们的分数高度相关。把他们当成 20 个独立观测来计算这个差距的标准误，会算出一个比真实情况小很多的不确定度，让”小班 +8 分”看起来比真实更可信。这个问题就是本书第 2 章要面对的核心。

1.4 学校层面的差异：一份数据的”伏笔”

扫一眼表 1.2，会注意到一个隐藏的事实：三类班级的数学均值都在 483 到 491 之间，跨度只有 8 分。但这只是**把全样本平均**得到的数字。如果按学校分组看，每所学校的均值差别能拉到 **124 分**——最差的学校均值是 427，最好的学校均值是 551。

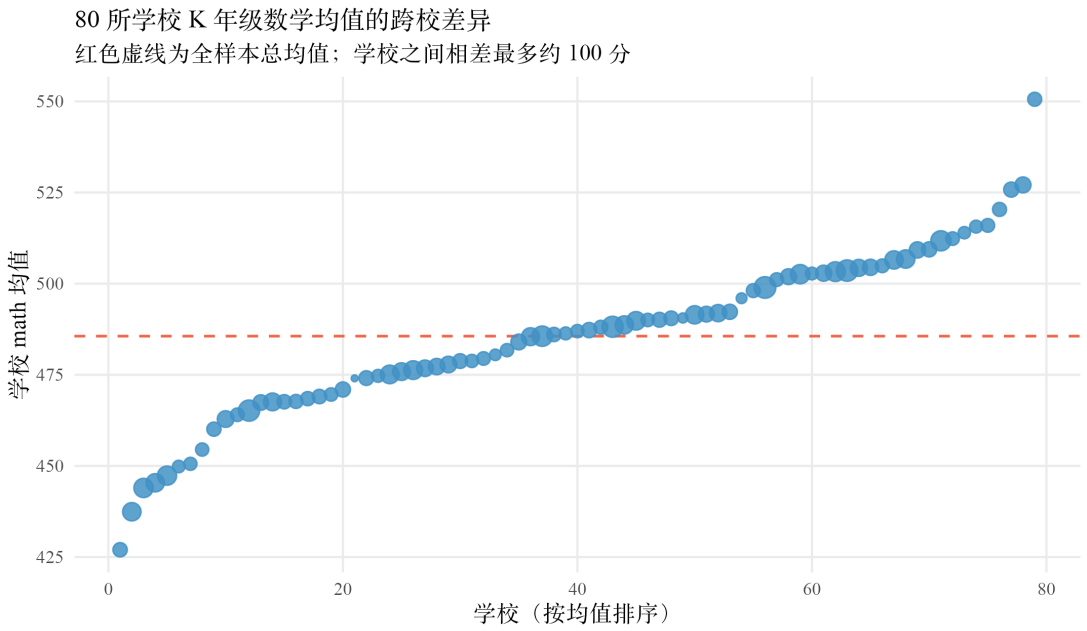


图 1.3: 79 所学校 K 年级数学均值的跨校差异。学校之间的差距远大于” 小班 vs 常班” 的处理效应

这个事实不是巧合，它说明 STAR 数据里有大量**学校层面**的系统性差异。这些差异来自学校所在社区的社会经济水平、师资条件、学校管理质量等等不可测变量。如果不把学校层面这部分变异分离出来，” 小班 +8 分” 这个结论会被学校间的巨大噪声淹没。

班级层面也有类似情况。337 个班的均值跨度从 405.7 分到 580.5 分，跨度 175 分，比跨校差异更夸张。这部分跨班差异里，一部分来自老师质量，一部分来自班里学生的随机组合，一部分来自小班 vs 常班的处理效应本身。如何把这三部分分开，是第 3 到第 8 章轮番要解决的问题。

停下来想一想。一名学生的分数同时受三种力量影响：自己的能力（学生层）、老师与班里同学的平均水平（班级层）、学校与社区环境（学校层）。一个朴素的分数差 $491 - 483 = 8$ 分把所有这些力量都搅在了一起。多层建模就是把这些力量一层层拆开来看的工具。

1.5 全书路线图：从朴素 OLS 到三层 HLM

13 项 BCG 试验是异质数据的极端例子，STAR 的 5,786 名学生是嵌套数据的极端例子。两者面对的统计学难题虽然不同，但思路一样：朴素方法把所有观测当独立处理会出错，正确的方法要把数据的结构编码进模型本身。

本书的结构沿着这条思路展开。第 2 章先用最朴素的 OLS 跟带班级随机截距的 lmer 对比，把” 为什么忽略嵌套会算错标准误” 这件事说清楚，并给出关键诊断指标——**组内相关 ICC**。第 3 章详细讲空模型的方差分解，让读者亲眼看到 `VarCorr()` 的输出怎么解读。第 4 章把学生层协变量加进 Level-1，讨论中心化的两种做法。第 5 章把处理变量 `cltype` 加进 Level-2，给出本书的核心因果估计。第 6 章引入随机斜率，讨论 SES 效应是否在班级间随机变化。第 7 章用跨层交互回答” 小班对低 SES 学生有没有额外作用”。第 8 章把学校层加进来，跑完整的三层模型。第 9 章做模型诊断与影响分析。第 10 章把所有方法的小班效应估计并列出来对比。

每一章末尾都会更新一张累积对比表，把当前章给出的小班效应、对应标准误、方差分量并列上去。读者如果想直接跳到结论，第 10 章的终极对比表会一目了然地展示：从朴素 OLS 到三层 HLM，” 小班效应” 这个数字是怎么从 +8.21（SE = 1.56，$p \approx 0$）漂移到 +8.84 （SE = 2.63，$p = 0.001$）的。点估计基本没变，但 **对它的把握度**变了。这就是多层建模的全部价值。

第 2 章 嵌套数据与组内相关 ICC

内容提要

- 看清朴素 OLS 在嵌套数据下为什么会算错小班效应的标准误
- 掌握组内相关 ICC 的定义、生活类比与具体数字
- 用设计效应 DEFF 把”实际样本量”折算成”有效样本量”
- 学会用 lme4 拟合最简单的随机截距模型，与朴素 OLS 做对比

第 1 章末尾把”小班 +8.08 分”这个数字摆上了桌面。$491.4 - 483.3 = 8.1$ 分看上去干净利落，对应的 OLS 标准误是 1.56，p 值小到 2.24×10^{-7}，按常规教科书的结论就是”高度显著、政策可信”。但这个结论里藏着一个被悄悄绕过去的前提：朴素 OLS 把 5,786 名学生当成了 5,786 个互相独立的观测。同班的 17 名学生坐在同一间教室，听同一位老师讲课，分享同一种学习氛围。他们的分数高度相关，远谈不上独立。

把不独立的观测当独立来算标准误，会得到一个比真实情况小得多的不确定度。本章要做的事就是把这件事掰开。先用一张图让读者亲眼看到”同班学生有多像”，再引入两个关键诊断指标——组内相关 ICC 与设计效应 DEFF，最后用 `lmer()` 跑一个最简单的随机截距模型，把朴素 OLS 与多层模型给出的两个标准误并列摆出来。第 3 章会顺着这条线，详细讲随机截距模型的方差分解到底意味着什么。

2.1 同班学生的相似性：把”独立观测”假设打开看

为了让”同班学生互相相关”这件事不再是抽象的统计假设，先看 6 个随机抽出来的班级的实际分数分布。图 2.1 把这 6 个班里的所有学生按班级排开，每名学生一个点，灰色横线是班级均值。

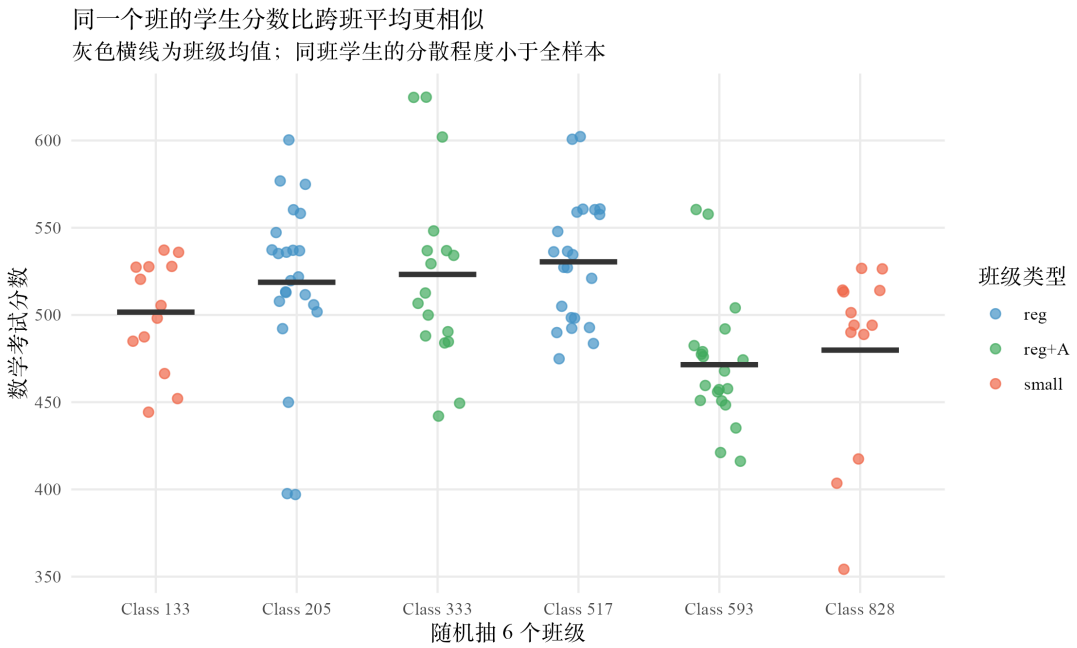


图 2.1: 随机抽取的 6 个班级，K 年级数学分数的班内分布。同一班学生的分散程度明显小于跨班均值的差距

读这张图的方法是先看竖直方向，再看水平方向。同一个班里的学生分数集中在一个不大的范围里，往上下方向看几十分；但 6 个班的均值（灰色横线）跨度能拉到一百多分。换句话说，跨班的差距比班内的差距更大。这就是”同班相似性”的视觉证据。

为什么会出现这种情况？因为分到同一个班的孩子被同一位老师教，老师讲课好坏会同时把全班 17 人的分数往同一个方向推。班里学习氛围浓，整班分数齐齐往上抬；老师当年家里出事请了几次假，整班分数齐齐往下掉。这部分由”班级共享的因素”带来的相似性，是嵌套数据最核心的特征。

停下来想一想。如果你已经知道某名学生所在班级的均值是 525 分，再让你猜这名学生的具体分数，你会比”完全没信息”猜得更准。这种”班级信息”对”个体分数”的预测力，正是 ICC 衡量的对象。

2.2 组内相关 ICC：班级里的”发言权”分到哪一层

定义 2.1 (组内相关 ICC)

对一份嵌套数据，假设结果 y_{ik}（学生 i 在班级 k）可分解为

$$y_{ik} = \mu + u_k + e_{ik},$$

其中 μ 是全样本均值，$u_k \sim N(0, \tau^2)$ 是班级 k 的随机截距偏差，$e_{ik} \sim N(0, \sigma^2)$ 是同班学生之间的随机波动。组内相关 ICC（intraclass correlation coefficient）定义为

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

其中 τ^2 是班级间方差，σ^2 是班级内（学生之间）方差，分母 $\tau^2 + \sigma^2$ 是总方差。ρ 取值在 0 到 1 之间，表示全部方差里有多少比例属于”班级层”。

ICC 这个数字看起来抽象，用一个生活场景就好理解。想象你走进一间教室开班会，老师让每个人挨个发言。如果班里气氛活跃、大家七嘴八舌、每个人意见都不一样，那”班级”这个集体几乎不存在共享意志，发言权完全在每位同学个人手里，对应 τ^2 接近 0、ρ 接近 0。反过来，如果班长一个人定调，全班跟着点头，那班级的”集体声音”几乎说明了一切，个体差异退到背景里，对应 τ^2 非常大、ρ 接近 1。

回到 STAR 的真实数字。用 `lmer()` 拟合一个最简单的空模型 `math ~ 1 + (1|tch)`，把全样本方差拆成班级间和班级内两部分，得到 $\tau^2 = 689$、$\sigma^2 = 1{,}595$。代进定义里：

$$\rho_{\text{class}} = \frac{689}{689 + 1{,}595} = \frac{689}{2{,}284} = 0.302.$$

其中 689 是班级间方差（每个班的”基线”在班际之间漂多少），1,595 是班级内方差（同一班里学生分数相对班级均值漂多少），两者之和 2,284 是总方差。

30.2% 这个数字怎么读？它说”K 年级数学分数的总变异里，有大约三成是来自班级层的差异，其余七成来自学生个体之间的差异”。三成不算小。教育学文献里把 ICC 大于 5% 当成”必须考虑嵌套结构”的临界点，10% 已经是相当强的嵌套，30% 属于嵌套结构非常显著的数据。STAR 给出的这个数字，是教育领域研究里常被作为基准引用的强嵌套样本。

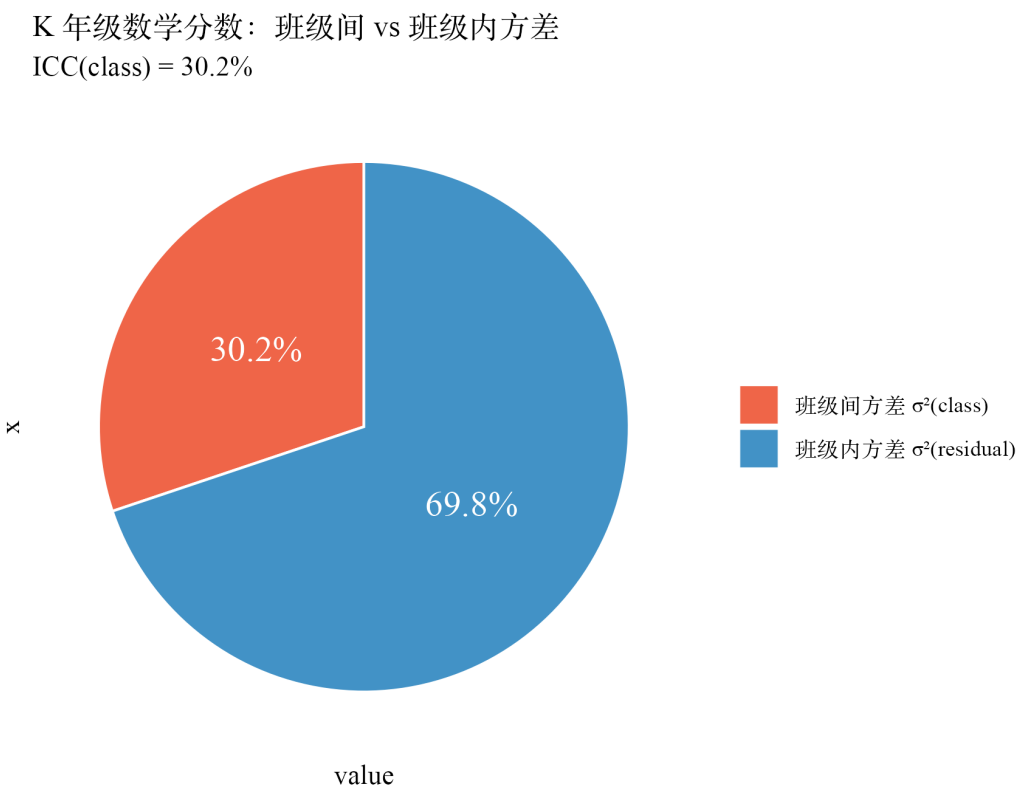


图 2.2: K 年级数学分数的方差分解。班级间方差占 30.2%，班级内方差占 69.8%

为什么要把 ICC 拆成这种” 比值” 形式而不是直接用 τ^2？因为 τ^2 本身随分数尺度变化，今天用 0–800 分的标化测试算出 689，明天换一份 0–100 分的小测验，τ^2 变成 11，但同学之间的相对相似性没变。把 τ^2 除以总方差就把尺度去掉了，给出一个跨尺度可比的量。这是 1962 年 Donner 与 Bull 把 ICC 从动物育种领域搬进群组随机试验时给出的设计选择。

2.3 设计效应 DEFF：实际样本量被折算成多少

ICC 大不大本身不是终点，关键是它怎么吃掉样本量。这就要引入设计效应。

定义 2.2 (设计效应 DEFF)

对每班平均 $\bar{n}$ 名学生、组内相关 ρ 的嵌套数据，设计效应定义为

$$\text{DEFF} = 1 + (\bar{n} - 1)\rho.$$

其中 $\bar{n}$ 是平均班级人数，ρ 是组内相关 ICC。它表示” 考虑嵌套之后，朴素 OLS 算出的方差被低估了多少倍”。把实际样本量除以 DEFF，得到有效样本量 $n_{\text{eff}} = n/\text{DEFF}$。

设计效应是 Kish 1965 在抽样设计文献里提出来的，原本是给整群抽样估总体均值准备的。后来被搬进多层建模，作为” 嵌套结构对推断把握度的折扣” 这个概念的速算工具。

代入 STAR 的数字：每班平均 $\bar{n} = 17.17$，$\rho = 0.302$，

$$\text{DEFF} = 1 + (17.17 - 1) \times 0.302 = 1 + 16.17 \times 0.302 = 5.88.$$

其中 17.17 是 5,786 名学生分到 337 个班里得到的平均班级人数，0.302 是上一节算出来的 ICC。

这个 5.88 给样本量打了一个大折扣。它说：” 朴素 OLS 用 5,786 这个样本量算标准误，背后假设的是 5,786 个独立观测。考虑同班相关之后，相当于只有 $5{,}786/5.88 \approx 985$ 个独立信息单位。” 换句话说，6 个 STAR 学生

才相当于 1 个真正独立的观测。同班同学之间的”信息冗余”把样本量打了一个六分之一的折扣。

这个折扣是 ICC 大小与班级规模共同作用的结果。如果 ICC 是 0.05，每班 20 人，DEFF 只是 $1+19\times0.05=1.95$，几乎不打折；ICC 涨到 0.3、班级规模再大一点，DEFF 一下就到 6 以上。STAR 这个 5.88 是嵌套数据里相当”恶劣”的一档。

定理 2.1 (雷区)

把朴素 **OLS** 用在嵌套数据上为什么会低估标准误？本质是 OLS 标准误公式里的方差估计 $\hat{\sigma}^2=\mathrm{RSS}/(n-p)$ 假设残差互相独立。当残差实际上有正相关（同班学生分数同向漂动），残差平方和会被低估，分母 $n-p$ 又用了”全样本”的自由度。两个偏差叠加，算出来的标准误比真实值小好几倍。

诊断方法：先用 `lmer()` 拟合一个最简空模型 `y ~ 1 + (1|group)`，从 `VarCorr()` 抓 τ^2 和 σ^2，算出 ICC；ICC 超过 5% 就要警惕，超过 10% 就必须改用考虑嵌套的方法。

稳健替代：随机截距模型（本章给出的最简版本）、cluster-robust 标准误、广义估计方程 GEE 都能修这个问题。多层建模的优势是除了修标准误，还顺便给出每个班的随机截距估计，下一章会用到。

2.4 用 lme4 跑随机截距模型：朴素 OLS 与 HLM 标准误对比

把上面的诊断指标落到具体代码上。下面这段 R 代码做三件事：先跑朴素 OLS 把小班效应的 SE 算出来，再跑空模型抓出 τ^2 和 σ^2，最后跑带处理变量的随机截距模型，把考虑嵌套之后的小班效应 SE 摆出来。

```r
library(tidyverse)
library(lme4)
library(lmerTest)
library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch), !is.na(sch))

# 朴素 OLS：把 5786 个学生当独立观测
ols_naive <- lm(math ~ cltype, data = d)
summary(ols_naive)$coefficients
#                Estimate Std. Error t value      Pr(>|t|)
# (Intercept)   483.2773    1.0735  450.21  0.000000e+00
# cltypesmall     8.0820    1.5645    5.17  2.341632e-07
# cltypereg+A    -0.3470    1.4955   -0.23  8.165843e-01

# 空模型：抓班级间方差与残差方差
m0 <- lmer(math ~ 1 + (1 | tch), data = d, REML = TRUE)
as.data.frame(VarCorr(m0))
#         grp        var1 vcov
# 1       tch (Intercept)  689
# 2  Residual        <NA> 1595

icc <- 689 / (689 + 1595)        # = 0.3017
nbar <- mean(table(d$tch))       # = 17.17
deff <- 1 + (nbar - 1) * icc     # = 5.88
n_eff <- nrow(d) / deff          # = 985

```

```r
# 加入处理变量的随机截距模型
m1 <- lmer(math ~ cltype + (1 | tch), data = d)
summary(m1)$coefficients
#                Estimate Std. Error  df t value Pr(>|t|)
# (Intercept)   483.31      2.69  335   179.4   <2e-16
# cltypesmall     7.22      3.72  335    1.94    0.053
# cltypereg+A    -0.69      3.85  335   -0.18    0.858
```

结果解读

关键是把朴素 OLS 与 HLM 给出的两个数字并列读。小班效应的点估计从 8.08 漂到 7.22，差距不大，模型选择不同对”平均效应有多大”这个问题影响有限。但标准误从 1.56 涨到 3.72，膨胀了 $3.72/1.56 = 2.38$ 倍。p 值从 2.24×10^{-7} 变成 0.053，恰好擦着 0.05 的边——朴素 OLS 给的”高度显著”其实是统计幻觉，真实情况是这个差距勉强够得上传统显著性门槛。

设计效应 5.88 与标准误膨胀 2.38 倍并不矛盾：DEFF 是方差比，标准误是方差的平方根，$\sqrt{5.88} = 2.42$，跟 2.38 几乎一致。两个数字其实在讲同一件事——同班学生的相似性把 5,786 这个样本量打成了 985 的有效信息。

这里要补一个细节。空模型给出的总方差 $689 + 1{,}595 = 2{,}284$ 跟朴素 OLS 把 y 当独立样本算出的方差 $\bar{s}^2 \approx 2{,}264$ 非常接近，仅相差不到 1%。这说明随机截距模型并没有”无中生有”地多解释了什么变异，它做的只是把同一份总方差拆成班级间和班级内两块。同样多的方差，不同的分配方式，给推断带来天差地别的结论。这是多层建模与朴素回归最根本的差别。

2.5 从一个 SE 到方差分解的全图

本章把”为什么不能朴素”这个问题彻底讲清楚了。读者现在应该能复述这个故事：朴素 OLS 把同班学生当成独立观测，得到的 SE 是 1.56；考虑嵌套之后，SE 涨到 3.72，是原来的 2.38 倍；这个膨胀比例可以由设计效应 $\sqrt{1 + (\bar{n} - 1)\rho} = \sqrt{5.88} = 2.42$ 提前算出来；其中 $\rho = 0.302$ 是组内相关 ICC，由空模型的方差分解 $689/(689 + 1{,}595)$ 得到。

但本章只把空模型当成”算 ICC 的工具”用了一次，并没有真正展开看它给出的三类输出——固定截距 β_0、班级间方差 τ^2、班级内方差 σ^2。这三个数字背后还有许多内容没讲：β_0 估的到底是什么”总均值”，$\tau^2 = 689$ 这个数字开根号以后能不能跟原始的”班级均值 SD”对应起来，每个班自己的”截距偏差”u_k 怎么估、估出来跟原始班均值有什么关系。这些是第 3 章的主题。

下一章会把空模型从头到尾解一遍：详细解读 `lmer()` 输出的每一行，讲清随机截距如何用 BLUP 估计、为什么估出来的班均值会比原始均值”更靠近全样本均值”（这个现象叫收缩），还会讲清 REML 与 ML 两种估计方法在小样本下为什么给出不同数字。把空模型这块讲透，第 4 章再加学生层协变量、第 5 章再加处理变量就水到渠成。

第 3 章 空模型：随机截距与方差分解

内容提要

- 详细解读 `lmer()` 空模型 `math ~ 1 + (1|tch)` 输出的每一行
- 掌握随机截距 u_k 的 BLUP 估计与"收缩"现象
- 弄清 REML 与 ML 两种估计方法在多层模型里的分工
- 用似然比检验验证空模型相对朴素 OLS 的拟合优势

第 2 章把空模型当成"算 ICC 的工具"用过一次。从 `VarCorr()` 抓出 $\tau^2 = 689$ 和 $\sigma^2 = 1{,}595$，代进 $\rho = \tau^2/(\tau^2+\sigma^2)$ 算出 ICC = 30.2%，本章的故事开始时这件事已经做完了。问题是空模型本身远不止给出两个方差数字这么简单。它的 `summary()` 输出里还有固定截距 $\beta_0 = 486.30$（SE = 2.49）这一行，有 337 个班级各自对应的随机截距估计 u_k，有 REML 跟 ML 两套不同的拟合结果，还有跟纯朴素 OLS 的似然比检验。这些都是第 4 章及以后会反复用到的基础量。

本章把空模型从头到尾拆一遍。先把模型方程写下来，再用一段散文把每个符号锚到具体的"人"或"班"上；接着讲随机截距的 BLUP 估计，重点说为什么估出来的班均值会被"拉"向全样本均值这个看似奇怪的现象；然后对比 REML 与 ML 两种估计方法在多层模型里的分工；最后用似然比检验确认"加随机截距"相对纯 OLS 是不是值得。第 4 章会在此基础上加入学生层协变量，第 5 章再加处理变量。

3.1 空模型的方程与每个符号背后的人

定义 3.1 (空模型与随机截距)

对学生 i 嵌套在班级 k 里的两层数据，最简单的空模型写作

$$y_{ik} = \beta_0 + u_k + e_{ik}, \quad u_k \sim N(0, \tau^2), \quad e_{ik} \sim N(0, \sigma^2),$$

其中 y_{ik} 是班级 k 里第 i 名学生的数学分数，β_0 是全样本"总均值"对应的固定截距，u_k 是班级 k 相对总均值的偏差，服从均值为 0、方差为 τ^2 的正态分布；e_{ik} 是该学生相对班级均值的个体波动，服从均值为 0、方差为 σ^2 的正态分布；u_k 与 e_{ik} 互相独立。"空模型"指右边没有任何解释变量，只用截距加随机效应来描述结果。

把这条公式翻成话：每名学生的分数等于一个总基线 β_0，加上他所在班级的"班级偏差"u_k，再加上他个人在班里的"个体波动"e_{ik}。$\beta_0 = 486.30$ 就是全样本数学均值的估计；u_k 是某个班相对这条基线高几分还是低几分；e_{ik} 是某名学生相对班级均值的余下偏离。图 3.1 把这条加性分解的三个来源画出。

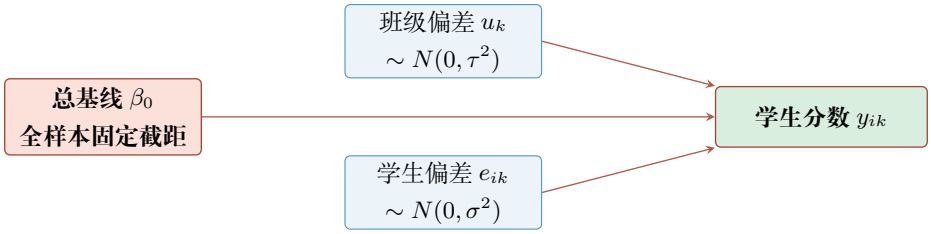


图 3.1: 空模型的三源加性结构：每名学生的分数 y_{ik} 等于全样本固定基线 β_0、班级随机偏差 u_k 与学生个体波动 e_{ik} 的代数和。两个随机分量相互独立，方差分别是班级间方差 τ^2 与班级内方差 σ^2。

```
library(tidyverse)
library(lme4)
```

```r
library(lmerTest)
library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch), !is.na(sch))

m0 <- lmer(math ~ 1 + (1 | tch), data = d, REML = TRUE)
summary(m0)
# Linear mixed model fit by REML. t-tests use Satterthwaite's method
# Formula: math ~ 1 + (1 | tch)
#
# Random effects:
#  Groups   Name        Variance Std.Dev.
#  tch      (Intercept)  689     26.25
#  Residual              1595     39.94
# Number of obs: 5786, groups:  tch, 337
#
# Fixed effects:
#             Estimate Std. Error      df t value Pr(>|t|)
# (Intercept)  486.30      2.49   333.6   195.3    <2e-16

VarCorr(m0)                          # 随机效应方差分量
ranef_class <- ranef(m0)$tch         # 337 个班的 BLUP
```

结果解读

这份输出包含三块关键信息。第一块是 `Random effects`：班级间方差 $\tau^2 = 689$，对应班级间标准差 $\sqrt{689} = 26.25$；班级内方差 $\sigma^2 = 1{,}595$，对应学生层标准差 $\sqrt{1{,}595} = 39.94$。班级层 26 分相当于”两个班的均值平均差 26 分”这种数量级；学生层 40 分相当于”班里同学之间的分数高低差 40 分”这种数量级。

第二块是 `Fixed effects`：固定截距 $\beta_0 = 486.30$，标准误 2.49，意思是”考虑了班级嵌套之后，对全样本数学均值的估计是 486.3，误差棒是 2.5 分”。这跟简单平均得到的 487 分非常接近，但 SE 比朴素 OLS 的 0.65 大了将近 4 倍——这是嵌套结构对均值估计本身的把握度的扣分。

第三块是底下的 `Number of obs: 5786, groups: tch, 337`，这一行是 `lmer()` 在帮你确认数据结构没读错：5,786 名学生分散在 337 个班级里。

3.2 BLUP 与收缩：班均值被全样本均值”拉”回

定义 3.2 (BLUP)

对随机截距模型 $y_{ik} = \beta_0 + u_k + e_{ik}$，每个班级 k 的随机截距 u_k 用一个带收缩权重的公式来估，而非直接采用”班级均值减全样本均值”：

$$\hat{u}_k = \frac{n_k \tau^2}{n_k \tau^2 + \sigma^2} \left(\bar{y}_k - \beta_0 \right).$$

其中 n_k 是班级 k 的人数，$\bar{y}_k$ 是该班的原始均值，β_0 是全样本固定截距，τ^2 与 σ^2 是两层方差。这个估计被称为最优线性无偏预测量（best linear unbiased predictor，BLUP），由 Henderson 1950 年代在动物育种里推导出来。

光看公式有点抽象，把它代成具体数字就很直观。STAR 里某个班 12 名学生原始均值 545 分，全样本均值 486.3 分，原始偏差是 $545 - 486.3 = 58.7$ 分。代入 BLUP 公式，权重是

$$\frac{12 \times 689}{12 \times 689 + 1{,}595} = \frac{8{,}268}{9{,}863} = 0.838.$$

其中 12 是班级人数，689 是班级间方差，1,595 是班级内方差。BLUP 估计是 $0.838 \times 58.7 = 49.2$，比原始偏差 58.7 缩了将近十分。这个缩水现象就是**收缩**（shrinkage）。图 3.2 把这步收缩的输入与输出关系画出。

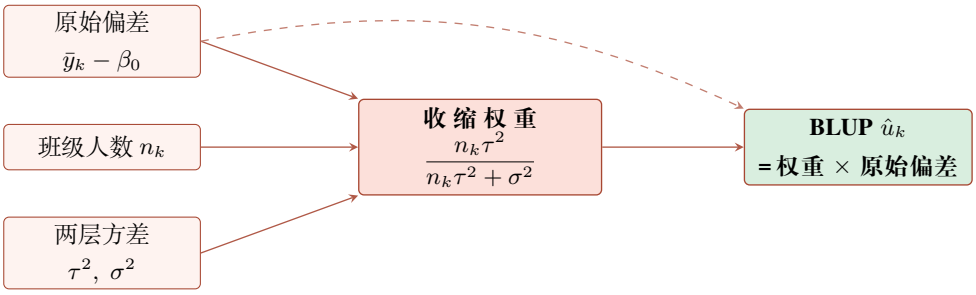


图 3.2: BLUP 收缩公式的输入-输出结构：原始偏差按一个介于 0 与 1 之间的权重被”拉”回总均值方向。权重由班级人数 n_k 与两层方差比共同决定——人数越多、班级方差占比越大，权重越接近 1（信原始偏差）；人数越少、班级方差占比越小，权重越接近 0（信全样本均值）。虚线提示原始偏差经过权重缩放后才到达 BLUP。

为什么要这样估？想象你是当年的 STAR 数据分析师，在 79 所学校的 337 个班里抽 12 个孩子算班均值。如果这 12 个孩子里恰好坐了几个考试当天发挥极好的学霸，原始均值会被拉得离谱地高。BLUP 的设计思路是：”样本量小的班、班均值特别极端的班，更可能是抽样偶然导致的，向全样本均值靠回去更靠谱；样本量大、信息充足的班，自己的均值就足够可信，不用调多少。”权重 $n_k\tau^2/(n_k\tau^2 + \sigma^2)$ 恰好实现这个直觉——n_k 大权重接近 1，几乎不动；n_k 小权重远离 1，往全样本均值靠。

用班里讨论的场景再说一遍。全班开班会评优秀学生，老师听到家长打电话单方面夸自家孩子时不会照单全收，会跟班里其他同学的反馈对比着听。家长给的 90 分（原始均值）跟班里同学反馈的中位数 70 分加权后变成 75 分（BLUP）。家长的话不是没价值，但要被”集体平均”的力量稍微拉回来。班里反馈越多（n_k 越大），家长的话就越占主导；反馈越少，越要靠班级整体判断。

回到 STAR 的全样本数字：337 个班的原始均值标准差是 28.59，BLUP 标准差是 24.29，整体收缩了 $(28.59 - 24.29)/28.59 = 15\%$。BLUP 的范围从 -71.04 到 $+85.16$，开根号一下，$\sqrt{\tau^2} = \sqrt{689} = 26.25$，跟 BLUP SD = 24.29 接近——理论上 BLUP SD 应该略小于 $\sqrt{\tau^2}$，因为收缩让方差被吃掉一点。

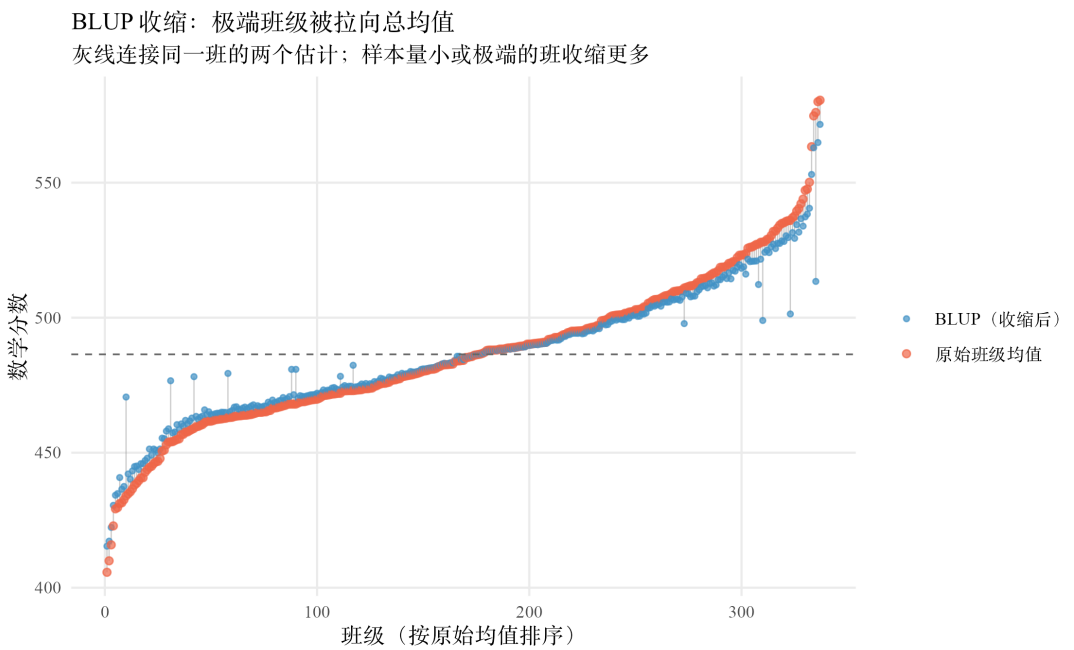


图 3.3: 337 个班级的原始均值（橙色）与 BLUP 估计（蓝色）。每条灰线连接同一个班的两个估计；样本量小或极端的班向全样本均值收缩更多

图 3.3 把 337 个班按原始均值排好，每个班画一对点，灰线连接同一班的两种估计。两端最极端的几个班（左端最低、右端最高）向中间收缩得最显眼，中间一带的班几乎没动。这跟 BLUP 公式给出的逻辑完全对应——极端值更可能是抽样偶然，应该被拉回去。

BLUP 这个概念是 1950 年 Henderson 在康奈尔大学农学院提出来的，原本是给牛奶产量遗传值的预测用的。育种学家面对的问题跟教育研究面对的问题数学形式相同：从有限观测里估计每头母牛、每个班级的” 真实潜力”，要兼顾” 自身观测” 和” 群体先验”。Robinson 1991 年在 *Statistical Science* 上写过一篇 BLUP 历史综述，是这条线索的标准参考。

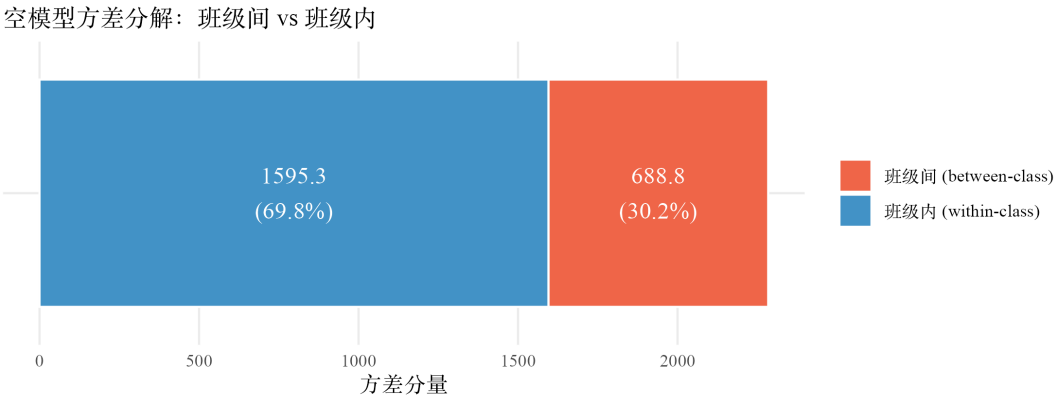


图 3.4: 空模型方差分解：班级间方差 689（30.2%）与班级内方差 1,595 （69.8%）

3.3 REML vs ML：两种估计方法在多层模型里的分工

`lmer()` 默认用 REML（restricted maximum likelihood，限制性极大似然）拟合。为什么不用更” 标准” 的 ML（极大似然）？这背后有个微妙但重要的偏差问题。

ML 估计方差分量时把所有观测代进似然函数一起最大化，结果是固定效应跟方差分量” 互相影响”。具体后果是 τ^2 跟 σ^2 会被系统性低估，幅度跟参数个数有关。Patterson 与 Thompson 1971 年在 *Biometrika* 上提出 REML，

思路是” 先用一个变换把固定效应那一块从似然里去掉，再对剩下的部分做极大似然”。这样估出来的方差分量是无偏（或近似无偏）的。

代价是 REML 的似然值不能用来比较不同固定效应结构的模型，因为变换本身依赖固定效应设定。如果想用似然比检验比较两个模型的固定效应（比如” 加 cltype 比不加 cltype 好不好”），必须切到 ML。

具体到 STAR 的空模型，REML 和 ML 给出的方差分量差别非常小：

表 3.1: 空模型 REML 与 ML 估计对比

参数	REML	ML
τ^2（班级间方差）	689	686
σ^2（班级内方差）	1,595	1,595
β_0（固定截距）	486.30	486.30

样本量 5,786、班级数 337 的情况下，两种方法差别小到可以忽略。但样本量小的时候差别会放大。Bates 等人写 lme4 时把 REML 设为默认，正是为了让用户在大多数场景下拿到的是无偏估计；只在做模型比较时显式切到 `REML = FALSE`。

定理 3.1 (雷区)

把 **ML** 估计当成无偏。有些教材、有些早期 SAS 代码默认用 ML 估方差分量，给出的 τ^2 系统性偏低。当用 ML 算 ICC，分子 τ^2 偏低、分母也偏低，比值不一定怎么变；但当 τ^2 用来做后续计算（比如设计效应、BLUP 收缩权重），就会传递误差。

诊断方法：检查 `lmer()` 调用是否显式设置了 `REML = TRUE` 或者用了默认值。如果代码里写了 `REML = FALSE`，看是不是只在做似然比检验之前临时切的。

记忆方法：估方差用 REML，比固定效应用 ML，最后报告时再切回 REML。lmerTest 的 `anova()` 在两个 `lmer()` 模型之间做 LRT 时会自动 refit 成 ML，不用自己切。

3.4 似然比检验：空模型相对朴素 OLS 是否值得

读完空模型的所有数字之后，要回答的最后一个问题是：” 加上 $(1|tch)$ 这个随机截距，相比纯 $y = \beta_0 + e_i$ 的 OLS，到底好多少？” 答案要靠似然比检验给。

似然比检验比较两个嵌套模型的对数似然，统计量

$$\mathrm{LR} = -2\big(\ell_{\text{simple}} - \ell_{\text{complex}}\big).$$

其中 ℓ 是各模型的对数似然，简单模型嵌套在复杂模型里。在原假设” 复杂模型不比简单模型好” 下，LR 服从自由度等于参数差的卡方分布。

把朴素 OLS（仅截距）跟空模型放进 `anova()` 比较。两者参数差是 1（多了一个 τ^2），但 τ^2 在原假设下卡在边界 0 上，标准卡方分布不再适用，要用 $0.5\chi^2_0 + 0.5\chi^2_1$ 这种” 边界混合” 分布——Self 与 Liang 1987 年专门处理过这个边界问题。lme4 跟 RLRsim 都内置了正确的检验程序，常规用法直接用 `anova()` 比较 `lmer()` 跟 `lm()`，得到 LR = 1,355.15、$p < 2.2 \times 10^{-16}$。

LR = 1,355 在 5,786 个学生 + 337 个班这个样本规模下属于合理量级。ICC = 30.2% 这种强嵌套结构下，纯 OLS 模型抓不住任何班级层信息，似然被显著拉低。LR = 1,355 的具体数值不是重点，重点是它显著超过任何合理的卡方临界值——多层结构必须建模。这条结论第 2 章已经从” 标准误膨胀 2.38 倍” 给出了一次，本节用似然比再验证一次，两条独立证据指向同一个判断。

3.5 从空模型到学生层协变量

本章把空模型从头到尾解了一遍。读者现在应该能复述这套逻辑：空模型方程 $y_{ik} = \beta_0 + u_k + e_{ik}$ 把每名学生分数拆成总基线、班级偏差、个体波动；`lmer()` 给出的输出包含 $\beta_0 = 486.30$、$\tau^2 = 689$、$\sigma^2 = 1{,}595$ 三块；337 个班的 u_k 用 BLUP 估计，估出来比原始班均值整体收缩 15%，因为极端班更可能是抽样偶然；REML 给方差分量、ML 给固定效应似然比检验，lme4 默认 REML 不用自己切；空模型相对纯 OLS 的 LR 高达 1,355，强嵌套结构必须建模。

但本章只用了" 空模型" 这一个最简单的设定。STAR 数据里还有大量被忽略的信息：学生的性别、族裔、是否享受免费午餐、教师的工龄与学位、学校类型。这些变量要么是学生层的（每个学生不一样），要么是班级或学校层的（同班同校的学生共享）。把它们加进模型，方差分量会怎么变？小班效应 +8.08 分会怎么漂移？

下一章先处理学生层协变量。会加性别、族裔、SES 三个变量，看 σ^2 能被压下去多少（学生层变量解释的是学生层方差），看 τ^2 会不会被影响（教育研究里这块通常变化很小）。还会引入一个对教育数据特别重要的方法——组均值中心化（centering within cluster），讲清同一个变量 SES 在" 个体层" 和" 班级层" 两个层面的效应可以完全不同，不分开看会把两者混合成单一系数。第 5 章再加处理变量 `cltype`，把" 小班 vs 常班" 正式作为多层模型的协变量推上来。

第 4 章 学生层协变量与组内组间分解

内容提要

- 把学生性别、族裔、SES 三个 Level-1 协变量加进随机截距模型
- 看清学生层变量主要压缩 σ^2（班级内方差），对 τ^2 影响有限
- 掌握组均值中心化 CMC 的定义与背后的”组内 / 组间分解”思想
- 用 SES 这个例子亲眼看到”班内效应”与”班间效应”在数字上完全不同

第 3 章把空模型解释清楚了，但模型右边只有截距加随机效应。STAR 数据里的 `sx`、`eth`、`ses` 三个学生层变量根本还没动过。教育研究的常识告诉我们这三个变量都跟学习成绩相关——女生在小学阶段普遍数学略好、不同族裔学生在标化测试上有系统差异、低社会经济状况的孩子家里学习资源少。这一章的任务是把它们正式加进模型，看它们会怎么影响方差分量与小班效应估计。

加之前先要把样本量重新对齐。前两章用的是 5,786 名学生，因为只要 math、cltype、tch、sch 不缺即可。本章再加 sex、eth、ses 三个协变量做非空过滤，剩下 5,768 名学生分散在 323 个班里。差掉的 18 名学生跟 14 个班对全局结论没影响，但本章及之后的所有数字都基于这个新样本。

本章先一步步加变量观察方差分量怎么变，再讲组均值中心化怎么把同一个变量的”组内”与”组间”两块效应分开。最后用一个雷区警告：直接放原变量会把组内组间效应压成单一系数，而这个单一系数没有清晰的因果含义。第 5 章会把处理变量 `cltype` 加进来，正式给出本书第一个”考虑学生层协变量之后”的小班效应估计。

4.1 逐步加入学生层协变量：方差分量的变化

先把基线锚清楚。本章样本（n = 5,768，323 班）下重新拟合空模型，得到 $\tau^2 = 649$、$\sigma^2 = 1{,}595$。班级间方差比第 3 章的 689 略低，是因为去掉的几个班里有方差较大的；班级内方差几乎没变。这两个数字是本节所有”协变量解释了多少方差”的参照基线。

接下来三轮加变量。第一轮加性别 sex_F（女生 = 1，男生 = 0），第二轮再加族裔 eth_B（黑人 = 1，其他 = 0），第三轮再加 SES ses_F（享受免费午餐 = 1，自费 = 0）。

```
library(tidyverse)
library(lme4)
library(lmerTest)
library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch), !is.na(sch),
         !is.na(sx), !is.na(eth), !is.na(ses)) |>
  mutate(sex_F = ifelse(sx  == "F", 1, 0),
         eth_B = ifelse(eth == "B", 1, 0),
         ses_F = ifelse(ses == "F", 1, 0))   # F = free lunch

m0 <- lmer(math ~ 1 + (1 | tch), data = d, REML = TRUE)
m1 <- lmer(math ~ sex_F + (1 | tch), data = d, REML = TRUE)
m2 <- lmer(math ~ sex_F + eth_B + (1 | tch), data = d, REML = TRUE)
m3 <- lmer(math ~ sex_F + eth_B + ses_F + (1 | tch), data = d, REML = TRUE)
```

```
summary(m3)$coefficients
#             Estimate Std. Error      df t value Pr(>|t|)
# (Intercept)  500.36     2.71   353.7  184.8  <2e-16
# sex_F          6.13     1.04  5479.9    5.9  4.0e-09
# eth_B        -14.34     1.95   358.8   -7.4  1.0e-12
# ses_F        -19.71     1.25  5650.2  -15.7  <2e-16
```

结果解读

三个固定效应的方向跟教育研究的既有发现完全一致。**女生平均比男生数学多 6.13 分**（$p < 10^{-8}$），尺度上对应约 0.13 个分数标准差。**黑人学生平均比其他族裔少 14.34 分**（$p < 10^{-12}$），约 0.30 个标准差，是这三个变量里效应最大的之一。**享受免费午餐的学生（低 SES 标志）平均比自费学生少 19.71 分**（$p < 2 \times 10^{-16}$），约 0.42 个标准差，效应规模最大。

这些效应不是因果效应，是**描述性的差距**。族裔与 SES 都不能随机分配，估出来的系数把”族裔本身的影响”和”族裔关联的所有其他社区、家庭、学校因素”混在一起。但这些数字仍然有用：它们告诉我们 K 年级数学分数在人群中是怎么分布的，作为后续因果分析的描述性背景。

更值得读的是方差分量怎么变。表 4.1 把四个模型并列：

表 4.1: 加入学生层协变量后的方差分量变化

模型	τ^2（班级间）	σ^2（班级内）	解释 within	解释 between
M0：空模型	649	1,595	0	0
M1：+ sex_F	648	1,575	1%	0
M2：+ eth_B	643	1,548	3%	1%
M3：+ ses_F	641	1,516	5%	1%

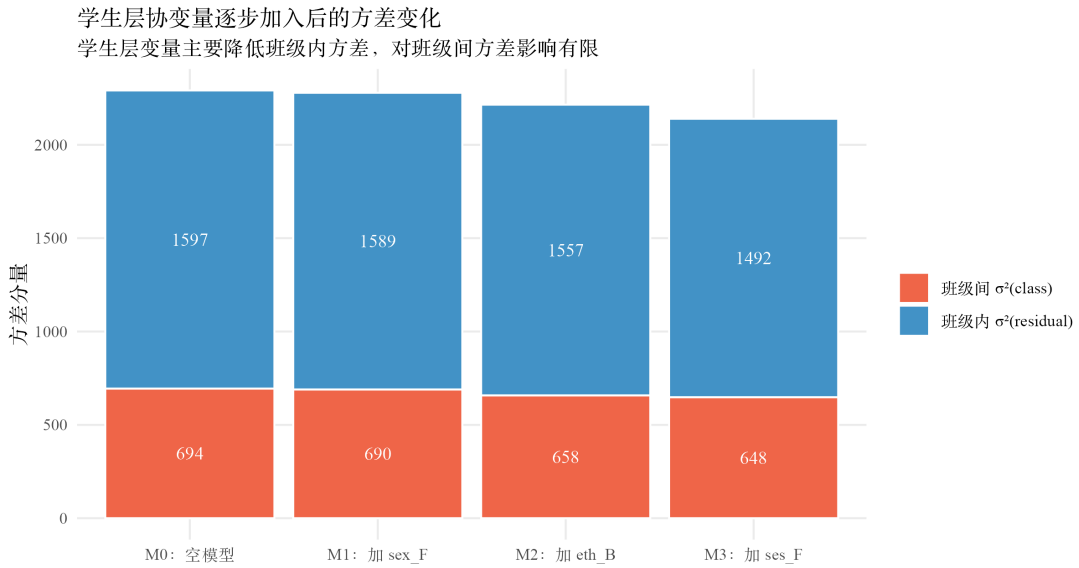


图 4.1: 学生层协变量逐步加入后的方差分解。班级内方差被压下了 79 分（约 5%），班级间方差几乎没动

读这张表的关键是看两列方差的对比变化。学生层协变量从 M0 加到 M3，班级内方差 σ^2 从 1,595 降到 1,516，压了 79 分约 5%；班级间方差 τ^2 从 649 降到 641，几乎没动。这个对比给出一条重要的经验规律：**学生层协变量主要解释学生层方差**。

为什么会这样？想一个班里 17 名学生的分数构成。这 17 个分数差距来自两类因素：一类是这 17 人有男有女、有不同族裔、有不同 SES，这些差异让班里学生的分数高低不同；另一类是这一班整体跟其他班相比有不同基线（老师、教室、班级氛围）。学生层变量解释第一类差异，所以压 σ^2；但对班际基线没什么帮助，所以 τ^2 几乎不动。要解释 τ^2，得靠班级层或学校层变量，这是第 5 章的事。

4.2 组均值中心化：把”班内效应”与”班间效应”分开

学生层协变量里有一个特殊的家伙：SES。它跟性别、族裔不太一样。性别在班里比例大概 50/50，族裔比例由学校所在社区决定但跟班际差异关系不大；SES 不同，低 SES 学生在 STAR 里集中分布在某些学校与班级里，班际 SES 比例从接近 0 到接近 100% 都有。

这种集中分布带来一个值得分开看的现象——同一个 SES 变量在两个层面起作用。一个层面是学生层：”这名孩子家里享受免费午餐，相对班里同学，他自己的分数会怎样？”另一个层面是班级层：”这个班里享受免费午餐的学生比例高，整个班级的基线会怎样？”两个问题在数据上对应同一个变量，但在因果上是两件事。

定义 4.1 (组均值中心化 CMC)

对学生层变量 x_{ik}，组均值中心化（centering within cluster，CMC）把它分解为两部分：

$$x_{ik} = \bar{x}_k + (x_{ik} - \bar{x}_k),$$

其中 $\bar{x}_k$ 是班级 k 的均值（class-mean，反映班级层信息），$x_{ik} - \bar{x}_k$ 是该学生相对班级均值的偏差（within-cluster，反映学生层信息）。把这两个量分别作为两个变量加进多层模型，可以让”班间效应”和”班内效应”被两个独立的系数分别估计。

用 SES 把这个分解走一遍。ses_F 取值 0 或 1。班级 k 里享受免费午餐的学生比例 $\overline{\text{ses_F}}_k$ 就是这个班的”低 SES 浓度”。某名学生的 ses_F $-$ $\overline{\text{ses_F}}_k$ 是他相对班里 SES 平均水平的偏差——如果他享受免费午餐而班里只有 30% 的同学享受，他的组内偏差是 $1 - 0.3 = 0.7$。图 4.2 把这步分解的两路分流画出。

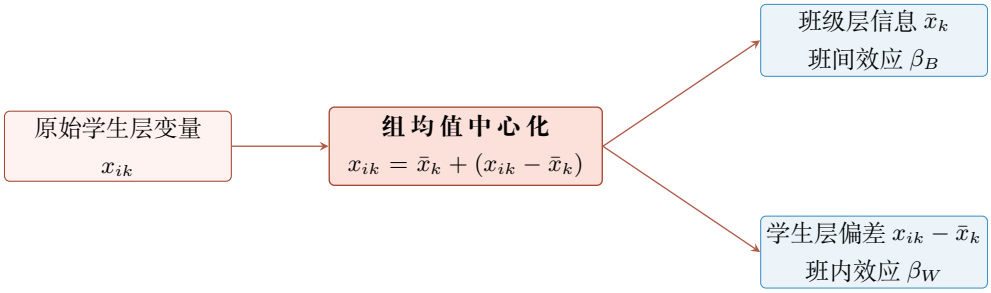


图 4.2: 组均值中心化把同一个学生层变量拆成两路：班级均值 $\bar{x}_k$ 承载”班间效应”β_B，反映高 SES 浓度的班级整体表现；偏差 $x_{ik} - \bar{x}_k$ 承载”班内效应”β_W，反映同班同学之间的个体差异。两路各有一个系数，因果含义不再混淆。

把这两个变量都加进随机截距模型：

```
d_cent <- d |>
  group_by(tch) |>
  mutate(ses_class = mean(ses_F)) |>
  ungroup() |>
  mutate(ses_within = ses_F - ses_class)

m_cmc <- lmer(math ~ ses_within + ses_class + (1 | tch),
              data = d_cent, REML = TRUE)
summary(m_cmc)$coefficients
```

```
#             Estimate Std. Error      df t value Pr(>|t|)
# (Intercept)  500.6      2.16   332.0  231.5  <2e-16
# ses_within   -21.60      1.27  5430.5  -16.9  <2e-16
# ses_class    -26.92      5.31   349.6   -5.1  6.1e-07

# 不分解的对照模型
m_plain <- lmer(math ~ ses_F + (1 | tch), data = d_cent, REML = TRUE)
summary(m_plain)$coefficients["ses_F", "Estimate"]
# [1] -21.89
```

结果解读

关键看三个数字。**组内 SES 效应 =** -21.60：在同一个班里，享受免费午餐的孩子比同班自费同学平均低 21.6 分。**组间 SES 效应 =** -26.92：把整班的免费午餐比例从 0 拉到 100%，整班基线下移 26.9 分。**不分解的混合系数 =** -21.89，几乎完全等于组内效应。

组间效应（-26.92）比组内效应（-21.60）大 5 分多。这说明”低 SES 学生集中的班级”对班里所有学生都不利——不光是低 SES 学生自己分数低，连同班的自费学生也会因为班级整体氛围、师资分配、课程难度调整等等被拉下来一截。这是教育不平等研究里反复出现的”组合效应”现象。

为什么不分解的混合系数 -21.89 离组内效应 -21.60 这么近，而离组间效应 -26.92 远？因为加权方式偏向组内。混合系数本质上是组内效应跟组间效应的某种加权平均，权重由协变量的组内方差跟组间方差的比例决定。SES 在 STAR 里的组内方差远大于组间方差（每个班里 SES 比例不是 0 就是 1 两种极端），加权平均自然偏向组内那一边。这种”系数被组内吃掉”的现象在不分解的模型里很难诊断，往往导致研究者错把组内效应当成”个体 SES 的普遍效应”。

停下来想一想。如果一所学校的政策制定者看到 ses_F 系数是 -21.89，他会得到什么结论？大概是”享受免费午餐的孩子需要更多帮助”。这条结论本身没错，但漏了一半信息——把整班 SES 比例降下来（比如混合社区招生）会带来一个独立的、更大的提升。组均值中心化的价值就是把这两条政策建议分开来给。

4.3 为什么不分解会出问题：组内组间被压成单一系数

定理 4.1 (雷区)

直接放原变量会把组内组间效应混合成单一系数。对 SES 这种”在个体和班级两个层面都有效应”的变量，模型 `math ~ ses_F + (1|tch)` 估出来的 -21.89 是组内效应与组间效应的加权混合，权重由变量本身的组内方差与组间方差比例决定，因此这个系数在因果上没有清晰含义。

诊断方法：怀疑某个学生层变量在班级或学校层有”集中分布”现象时，先算 $\bar{x}_k$ 在班际之间的方差，再跟个体 x 在班内的方差比。如果班际方差占总方差超过 5%（也就是变量本身的 ICC > 0.05），就应该用 CMC 拆开来跑。

记忆方法：教育与组织研究里，凡是”个体特征”跟”组成员构成”相关的变量都要警惕——SES、学习成绩、专业兴趣、年龄、工作经验。性别跟族裔在班级随机分配的情况下班际方差小，可以不拆，但严格起见也建议做一次诊断。Curran 与 Bauer 2011 年在 *Annual Review of Psychology* 上有一篇综述专门讲这件事的危害与处理。 ♡

CMC 这个方法在多层建模文献里有几个不同名字，含义都一样：组均值中心化、group-mean centering、within-between 分解。Raudenbush 与 Bryk 1992 年的 *Hierarchical Linear Models* 把它写进教科书；Snijders 与 Bosker 2012 年的 *Multilevel Analysis* 给了更细的指南，尤其是讨论”班级均值”作为协变量时班级人数大小对估计稳定性的影响。班级人数 n_k 太小（少于 5 人）的话班级均值估计本身就很噪声，不建议直接拿来当协变量；STAR 里每班

平均 17 人，这个问题不严重。

4.4 学生层之上还有班级层与学校层

本章把学生层协变量这块讲完了。读者现在应该能复述：性别、族裔、SES 三个变量加进随机截距模型后，班级内方差 σ^2 从 1,595 降到 1,516，压了 5%；班级间方差 τ^2 几乎没动，因为学生层变量解释不了班际基线差异。SES 这种”在班级里集中分布”的变量值得用组均值中心化拆开，让组内效应（-21.60）跟组间效应（-26.92）各自有清晰系数；不拆的话两者会被压成一个 -21.89 的混合数字，因果含义模糊。

但本章一直绕开了”小班 vs 常班”这件事。处理变量 `cltype` 是班级层变量——同一个班里所有学生共享同一个班级类型。它要进入模型，跟学生层的性别、族裔、SES 不在同一层。下一章正式把 `cltype` 加进来，给出本书第一个”考虑了所有学生层与班级层协变量之后”的小班效应估计。读者会看到点估计从纯朴素 OLS 的 $+8.21$ 漂到带学生协变量的 $+7.68$，标准误从 1.50 涨到 3.63，跟第 2 章给出的”嵌套结构会膨胀 SE”的逻辑完全一致。第 5 章也是本书第一次更新累积对比表的章节——往后每章都会把当前模型的小班效应数字加进去，到第 10 章合并成一张终极对比。

第 5 章 班级层协变量与处理效应

内容提要

- 把处理变量 *cltype* 加进 Level-2 固定效应，给出本书的核心因果估计
- 理解班级层固定效应在多层模型中的位置与含义
- 看清”小班 +8 分”这个估计在不同设定下的稳定性
- 学会区分”班级层方差被解释了多少”与”模型整体变好了多少”

第 4 章把性别、族裔、家境三类学生层协变量加进了 Level-1，让 5,768 名学生的分数在进入合并之前先被自身背景”解释”了一部分。剩下没解释的部分，最大的一块还是班级之间的差异——$\sigma^2_{\text{class}} = 649$ 比残差 $\sigma^2_{\text{resid}} = 1595$ 小，但仍然是一个不能忽视的方差源。这块班际差异从哪里来？它有没有结构、能不能被”小班 vs 常班”这个处理变量解释一部分？

本章把核心处理变量 *cltype* 加进 Level-2，给出本书所有讨论都绕回去的那个数字：**小班相对常班，K 年级数学分数大约提高了 7.7 到 7.9 分**。更重要的是，这一章会让读者看到，把 OLS 的 +8.21（SE 1.56）改写成 HLM 的 +7.90（SE 3.76），点估计几乎没变，但对这个数字的把握度变了一大截。这就是本书第一次正式收获多层建模的”果实”。

5.1 班级层固定效应：把处理变量放在哪一层

把 *cltype* 加进模型，在写法上只是 `lmer()` 公式右边加一个项。但它在统计学上属于一个完全不同的类别：班级层固定效应。这个名字背后藏着一个对模型解释力影响巨大的设计决策，本节先把它讲清楚。

5.1.1 先看一个生活类比：班里发助教就像班里多请一个家教

想象一所小学开学第一天，校长把入学的新生全部洗牌，按抽签的方式分进 4 个班里。4 个班的差别只有一个：1 班是小班、2 班是常班、3 班是常班加助教、4 班再补一种。校长这个动作叫**班级层随机分配**——分配的对象是**班**，不是单个学生。

这件事情决定了 *cltype* 在多层模型里只能放在 Level-2，不能放在 Level-1。同一个班里所有学生共享同一个 *cltype* 值，它的变异完全发生在班与班之间，而不是学生与学生之间。在一个 17 人的小班里，没人是”半个小班学生”，全 17 人都同时享受这种处理。

班里发助教跟班里没助教的区别，就像一个班整体多请了一个家教，整班受益。家教不会只服务班里某一个学生，而是绕着全班转。这正是 *cltype* 在班级层而不是学生层的原因。

5.1.2 班级层固定效应的形式定义

定义 5.1 (班级层固定效应)

设第 k 个班级的处理状态为 cltype_k，取值为 `reg`、`reg+A` 或 `small`（参照类是 `reg`）。把它放进多层模型的 Level-2，得到含班级层固定效应的随机截距模型：

$$y_{ik} = \beta_0 + \beta_1 \mathbb{1}\{\text{cltype}_k = \texttt{small}\} + \beta_2 \mathbb{1}\{\text{cltype}_k = \texttt{reg+A}\} + u_k + e_{ik},$$

其中 y_{ik} 是第 k 个班里第 i 名学生的数学分数，$\mathbb{1}\{\cdot\}$ 是指示函数，β_1 表示平均而言把一名学生从常班调进小班分数提高了多少，β_2 是常班加助教相对常班的差，$u_k \sim N(0, \sigma^2_{\text{class}})$ 是班级随机截距，$e_{ik} \sim N(0, \sigma^2_{\text{resid}})$

是学生残差。

公式里 β_1 和 β_2 都没有学生下标 i，因为它们对班里所有学生都一样。这正是把它放在 Level-2 的统计含义：处理变量**在班级层取一个值**，整班共享。

为什么要这样写？因为如果错把 *cltype* 当作学生层变量，假设它能在班里的学生之间变化，模型会误以为有效样本量是 5,768 而不是班级数 323。这正是第 2 章讨论过的”独立性假设破坏标准误”的精确版本。

5.1.3 多层模型里识别处理效应的关键：随机分配在哪一层

定义 5.2 (处理效应在多层模型中的识别)

当处理变量 D_k 在第 k 个班级层被完全随机地分配，即 $D_k \perp\!\!\!\perp (u_k, e_{ik})$，则在含 D_k 的随机截距模型中，D_k 的系数是因果意义上的平均处理效应。如果分配不是随机的，D_k 的系数还会包含选择偏差。

STAR 数据的优势就在这一点。班级类型是 1985 年校长抽签决定的，跟班级随机截距 u_k 在分配那一刻独立。所以 β_1 这个估计可以直接读作”小班相对常班的平均处理效应”，不需要再做倾向得分匹配、双重差分或者其他因果识别策略。

这点和一般观察性多层数据相当不同。如果数据里”小班”是因为家长给孩子报名精英小班自己跑去报的，那么 *cltype* 跟 u_k 就不独立——精英小班的家庭背景本来就更优，β_1 不再是因果效应。STAR 因为是随机分配的实验数据，绕开了这个问题。本章接下来给出的所有数字都建立在这个识别基础上。

5.1.4 学术背景：为什么 STAR 是教育统计的奠基样本

随机分配的小班实验在 1985 年之前几乎没人做过。教育界长期靠观察性数据推论小班效应，但低生师比的学校往往本身就有更好的资源、更优秀的师生，混杂得一塌糊涂。Mosteller [6] 在 *The Future of Children* 上发表了 STAR 的早期分析，论文标题直接叫《幼儿园到三年级班额减小研究：作为美国教育史上最重要实验的田纳西研究》。从那以后，STAR 数据成了教育政策辩论的关键证据。

经济学界对 STAR 的引用更频繁。Krueger [3] 在 *Quarterly Journal of Economics* 上用 STAR 数据估算了小班效应的劳动经济学含义，给出了”分到小班一年大约能拉高 0.2 个标准差测试分数”这个被广泛引用的数字。Krueger and Whitmore [4] 又跟踪 STAR 学生的中学表现，提出小班效应不仅在小学存在，还能延续到 SAT 和大学入学率。

后来 lme4 的作者把 STAR 收进 mlmRev 包作为多层建模教学的标准样本，正是因为它**同时具备**几样珍贵的特性：处理是随机的（因果可识别）、嵌套结构清楚（学生—班级—学校）、样本足够大（5,768 名学生）。本章接下来用 lme4 对它跑出来的数字，与 30 年来教育统计文献里的数字会高度一致。

5.2 从 M3 到 M4：把 cltype 放进随机截距模型

第 4 章给出了 M3——含三个学生层协变量（性别、族裔、家境）的随机截距模型。它给出了 $\sigma^2_{\text{class}} = 649$ 的班级方差。本章 M4 在它的基础上把 *cltype* 加进去，看看处理效应能不能在这个班级方差里”咬一块出来”。

5.2.1 为什么先单独跑 M4 再加学生层协变量

把 M4 设计成”只含 *cltype*”，不含学生层协变量，是为了把”处理效应这一项的纯贡献”和”加了学生协变量后的稳定性”分开看。这是经济学和教育统计文献里常见的逐层加变量做法，目的是让读者看清每一步的边际效果。

```r
library(lme4); library(lmerTest); library(tidyverse); library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch), !is.na(sch),
         !is.na(sx), !is.na(eth), !is.na(ses), !is.na(exp)) |>
  mutate(
    sex_F  = ifelse(sx == "F", 1, 0),
    eth_B  = ifelse(eth == "B", 1, 0),
    ses_F  = ifelse(ses == "F", 1, 0),
    cltype = factor(cltype, levels = c("reg", "reg+A", "small")),
    exp_c  = exp - mean(exp, na.rm = TRUE)   # 教师工龄中心化
  )

# M4：仅班级层处理变量
m4 <- lmer(math ~ cltype + (1 | tch), data = d, REML = TRUE)
summary(m4)

# M5：加学生层协变量
m5 <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | tch),
           data = d, REML = TRUE)

# M6：再加教师工龄
m6 <- lmer(math ~ cltype + sex_F + eth_B + ses_F + exp_c + (1 | tch),
           data = d, REML = TRUE)
```

把代码扔进 R 跑出来的 M4 关键数字在表 5.1 里。三个模型给出的小班系数在 7.68 到 7.90 之间晃，标准误在 3.61 到 3.76 之间晃，p 值都在 0.033 到 0.036 之间。

表 5.1: M4–M6 的关键系数（n = 5,768，323 班）

模型	small β	SE	p	reg+A β
M4：cltype + (1|tch)	7.90	3.76	0.036	−0.65
M5：+ 学生协变量	7.68	3.63	0.035	0.14
M6：再 + 教师工龄 *exp_c*	7.73	3.61	0.033	−0.20

结果解读

三个模型的小班系数都落在 7.7 到 7.9 之间，差距远小于一个 SE （≈ 3.6）。这说明本章给出的核心数字**对模型设定不敏感**：哪怕加进或拿掉学生层协变量、教师工龄，结论几乎不变。这种稳定性是因果证据可信度的一种体现，因为如果小班效应只是某个混淆变量的伪相关，加进相关协变量后系数应该会大幅缩水。**常班 + 助教**的系数始终在 0 附近（$-0.65, +0.14, -0.20$），p 值全部接近 1。这与第 1 章的朴素均值差 $\bar{y}_{\text{reg+A}} - \bar{y}_{\text{reg}} = -0.4$ 一致。结论是助教这个变量基本无效，至少在 K 年级数学分数上看不出差别。

5.2.2 对比朴素 OLS 与 HLM：点估计稳，标准误膨胀

把朴素 OLS（$\beta = 8.21$, SE 1.56）摆到 M4 旁边，可以看到一个标志性的对比：**点估计仅相差 0.31 分，标准误从 1.56 涨到 3.76，足足放大了 2.4 倍**。

这跟第 2 章讨论的设计效应完全一致。DEFF = 5.88 的有效样本量大约是 5,768/5.88 ≈ 980，开根号正好把

SE 放大约 $\sqrt{5.88} \approx 2.42$ 倍。OLS 之所以给出 SE 1.56，是因为它把 5,768 名学生当独立观测；HLM 给出的 SE 3.76 才是诚实地承认 17 名同班学生只值” 约 4 个独立观测” 之后的真实不确定度。

停下来想一想。如果一个研究只报告 OLS 结果（$\beta = 8.21$, $p < 0.001$），读者会觉得小班效应板上钉钉。换成多层模型后（$\beta = 7.90$, $p = 0.036$），仍然显著，但已经是” 擦着 5% 显著线过去” 的水平。这就是为什么严肃的教育政策评估都要求分析多层结构而不是用 OLS 凑数。

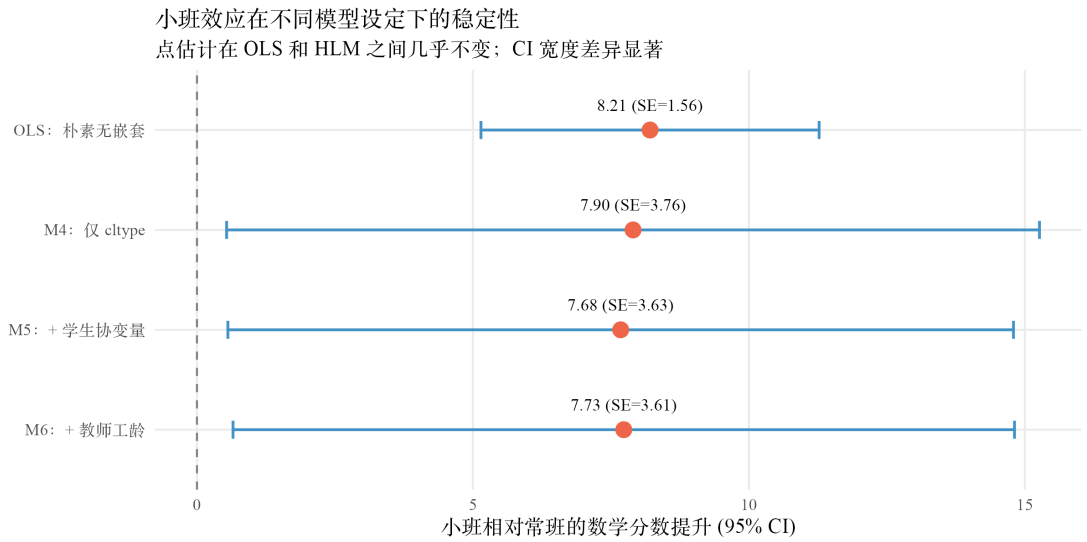


图 5.1: 小班效应在不同模型设定下的森林图：点估计在 7.68 到 8.21 之间高度集中，但 95% CI 宽度在 OLS 和 HLM 之间差异显著

森林图（图 5.1）把这层意思一目了然地放在一张图上。横轴是小班相对常班的分数提升，每一行是一个模型。OLS 的横线最短，HLM 系列的横线长 2.4 倍，但所有点估计都集中在 8 分附近。

5.3 班级层方差被 cltype 解释了多少

加完 *cltype* 之后还要做一件事——看看班级层方差有没有被它解释一部分。如果 *cltype* 真的把班际差异的一部分” 咬掉了”，那 σ^2_{class} 应该比 M3 小。表 5.2 给出真实数字。

表 5.2: 班级层方差在 M3–M6 之间的变化

模型	σ^2_{class}	σ^2_{resid}	相对 M3 的”R^2”
M3：仅学生协变量	649	1516	—
M4：cltype + (1|tch)	685	1595	−5.5%
M5：cltype + 学生协变量	640	1516	+1.4%
M6：再 + 教师工龄	632	1515	+2.6%

结果解读

M5 把班级层方差从 649 降到 640，*cltype* 解释了班级层方差的 1.4%。这个数字非常小，意味着**班级之间的差异主要不是处理类型本身造成的**，更多来自老师质量、班里学生组合的随机波动、班级氛围这些没能进入模型的因素。M6 加进教师工龄后再降 8 分，反映了” 老师经验” 也只能解释一小块班际差异。

小数字配上显著的处理系数，给出一个明确的结论：小班效应是真的，但它只占班际差异的一小块。绝大多数班际方差另有来源，需要更多信息（教师评估、教学风格、班级组成）才能进一步刻画。这也是后续章节要讨论的方向。

定理 5.1 (雷区)

M4 的 $\sigma^2_{\text{class}} = 685$ 比 M3 的 649 还大，相对 M3 的"R^2" 是 -5.5%。加了一个有意义的解释变量却让方差升高，需要单独解释。

这其实是多层模型里一个常见但被低估的现象。M3 的随机截距 u_k 在估计时，把所有班际差异（包括 *cltype* 造成的部分）都吸进了 σ^2_{class}，并通过学生协变量稍微修正了一下。M4 把 *cltype* 拎出来作为固定效应，但同时也改变了模型里"班级随机截距对应的参考点"——M3 里 u_k 围着全样本均值波动，M4 里 u_k 围着常班均值波动（因为 `reg` 是参照类），两者的方差有可能不严格相等。更细致的原因是 *cltype* 不是连续协变量，REML 估计的方差成分对协变量编码方式不完全单调。

诊断方法。不要把"σ^2_{class} 没下降"理解为"模型变差了"，而要看固定效应的解释力。M4 给出的处理系数 7.90 与样本均值差 8.08 接近，处理效应被正确识别出来了；班级方差没缩反胀只是 REML 估计的副作用。M5（加了学生协变量后）的 $\sigma^2_{\text{class}} = 640$ 才是更可靠的"班级方差剩多少"参照值。

记忆方法。多层模型里"班级层 R^2"这个数字本来就不存在标准定义，最好的做法是给出所有模型的 σ^2_{class} 表格，让读者自己看变化，不要把它强行套进 OLS 里"R^2"的解读框架。 ♡

5.4 班级 BLUP 的变化：班际差异在解释什么之后还剩多少

第 3 章解释过班级 BLUP 是什么——每个班的"被收缩过的"随机截距估计。一个有意思的对比是 M4 的 BLUP（仅 *cltype*）跟 M5 的 BLUP（加了学生协变量）之间有没有变化。

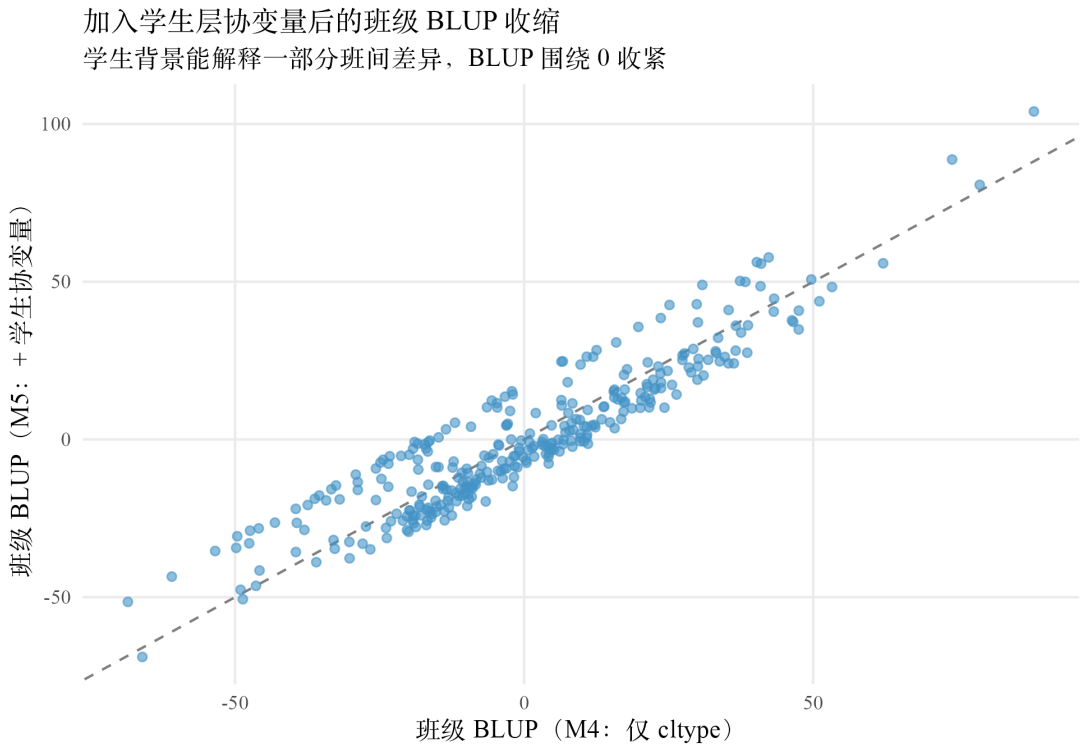


图 5.2: 加入学生层协变量后的班级 BLUP 收缩。横轴是 M4 的班级 BLUP，纵轴是 M5 的班级 BLUP；两者高度相关但 M5 的散点稍向 0 收紧，说明学生背景能解释一部分班际差异

图 5.2 的散点近似落在 $y = x$ 直线上，但整体略向 0 方向收紧。解读是：M4 的 BLUP 包含了"学生组成不同造成的班级均值偏移"，M5 把学生协变量纳入固定效应后，这部分偏移被剥离出去，剩下的 BLUP 更接近"老师本身的特质 + 班级氛围"那一块。

停下来想一想。如果某个班 BLUP 在 M4 是 +15（看起来这班特别厉害），但 M5 里降到 +10，这 5 分的差距

就是被学生背景解释掉的部分。剩下的 +10 才是除去学生组成之后这位老师真正让全班” 超出平均” 的那一份。这个洞见在评估教师效能（teacher value-added）的文献里被反复使用。

5.5 累积方法对比表（首次建立）

从本章开始，本书每章末尾都会更新一张累积对比表，把当章给出的小班效应、对应标准误、方差分量并列上去。到第 10 章会合并为终极版。本章是这张表的起点，前两行是对比基准（OLS 朴素和 HLM 仅 cltype），后两行是本章重点的 M5 和 M6。

表 5.3: 累积方法对比（截至第 5 章）

方法	小班 β	SE	95% CI	σ^2_{class}
OLS 朴素	8.21	1.56	[5.15, 11.27]	—
M4：cltype + (1|tch)	7.90	3.76	[0.55, 15.26]	685
M5：+ 学生协变量	7.68	3.63	[0.57, 14.79]	640
M6：再 + 教师工龄	7.73	3.61	[0.66, 14.80]	632

第 8 章引入三层结构后，这张表会扩出 σ^2_{school} 一列。第 10 章会再补 AIC/BIC 与终极结论。

5.6 下一章预告

到这里，本书第一个核心因果数字已经稳稳落在表上：**小班相对常班，K 年级数学分数提高约 7.7 分，p 值约 0.035**。但这个数字隐含一个假设——**所有班级对低 SES 学生的” 补偿效应” 都一样**。换句话说，M5 模型里 *ses_F* 的斜率被假定为固定值，每个班的” 低 SES 学生比非低 SES 学生少多少分” 都是同一个数。

这个假设合理吗？某些班级里的老师可能格外擅长照顾低 SES 学生，让差距收窄；另一些班可能差距更大。这种” 斜率随班级变化” 的现象，技术上叫**随机斜率**。第 6 章会让 *ses_F* 的斜率也成为随机效应，看看班际差距到底有没有显著的异质性，并用 LRT 给出明确的检验结果。

第 6 章 随机斜率与协变结构

内容提要

- 把学生层 *ses_F* 的斜率从”全班级一个数”放成”每班一个数”
- 理解随机斜率的几何含义：班际差异不只在截距，还可以在协变量斜率上
- 用似然比检验 LRT 给出”随机斜率是否显著”的明确结论
- 学会读 `lmer()` 的 `(1 + x | g)` 和 `(1 + x || g)` 写法的差别

第 5 章把处理变量 *cltype* 加进 Level-2，得到本书核心的 +7.7 分小班效应。模型里所有学生层协变量（性别、族裔、家境）的斜率都被默认为**所有班级共享同一个值**。比如低 SES 学生比非低 SES 学生平均低 −19.7 分，这个差距被假定为每个班级、每位老师都一样。

但这个假设其实很强。有的班级里老师特别善于照顾家境困难的学生，让 SES 差距收窄；另一些班里这个差距可能拉得更大。把这种”**斜率本身在班际变动**”的可能性写进模型，就是本章要做的事。技术上叫**随机斜率**（random slope）。本章用 STAR 数据跑出来，明确给出一个负面结论：**在这份数据上，SES 斜率的班际异质性几乎不存在**。但负面结论也有教学价值——它告诉读者随机斜率不是想加就加，需要 LRT 给出依据。

6.1 随机斜率：让协变量的影响在组间变化

6.1.1 先看一个生活类比：每个班对低 SES 学生的”关照度”不同

想象同一个学区里两位老师。第一位老师特别擅长辅导家境困难的学生，会主动给免费午餐学生留作业本、组织放学后的小组学习。在她的班里，低 SES 学生跟非低 SES 学生的分数差距可能只有 10 分。第二位老师完全按教材进度走，从不针对差异学生做调整，他班里低 SES 学生跟非低 SES 学生的差距可能拉到 25 分。

如果用 M5 那种”全班级共享一个 SES 斜率”的模型来分析这两个班，得到的会是一个**平均斜率**——大概 −19.7 分，但这个数字模糊掉了班际的真实差异。把”每位老师对低 SES 学生的关照度”作为班级的另一个特征写进模型，就是随机斜率要解决的问题。

随机斜率的概念可以简单说成：**斜率不只一个，每个班一个，全样本的斜率分布有均值有方差**。均值就是 M5 给出的固定斜率 −19.7，方差是新东西，叫斜率的随机方差。

6.1.2 随机斜率的形式定义与几何含义

定义 6.1 (随机斜率)

设 x_{ik} 是学生 i 在班级 k 的某个协变量（如 *ses_F*）。随机斜率模型把这个协变量的斜率允许在班级之间变化：

$$y_{ik} = (\beta_0 + u_{0k}) + (\beta_1 + u_{1k})x_{ik} + e_{ik},$$

其中 β_0, β_1 是固定效应（全样本的平均截距与平均斜率），$u_{0k} \sim N(0, \tau_0^2)$ 是班级 k 的随机截距偏差，$u_{1k} \sim N(0, \tau_1^2)$ 是班级 k 的随机斜率偏差，$e_{ik} \sim N(0, \sigma^2)$ 是学生残差。

公式里 u_{1k} 是新东西。它的含义是”第 k 个班级的 SES 斜率比全样本平均斜率 β_1 多出多少”。$u_{1k} > 0$ 的班级里，低 SES 学生跟非低 SES 学生的差距比平均小（因为 β_1 是负数，加上正的 u_{1k} 等于差距变小）；$u_{1k} < 0$ 的班级里，差距比平均更大。τ_1^2 度量了这种班际异质性的强度，是模型新引入的参数。

几何上，M5 的拟合给出**一条共同的直线**（截距随班级变，斜率不变）。M7 的拟合给出**每班一条直线**（截

距和斜率都随班级变）。如果 τ_1^2 很大，30 个班的 30 条直线会” 扇形发散”；如果 τ_1^2 接近 0，30 条直线几乎平行，跟 M5 给出的一条直线没差。本章要回答的就是 STAR 数据里到底是哪种情况。图 6.1 把固定斜率与班级随机偏差的合成关系画出。

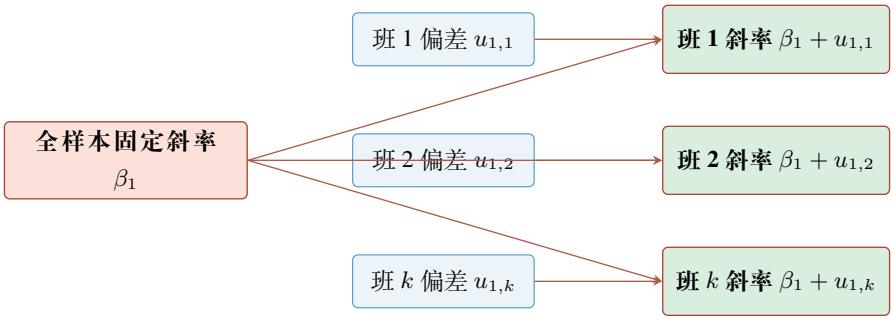


图 6.1: 随机斜率把每个班级的协变量影响拆成” 固定平均斜率 β_1” 与” 班级随机偏差 $u_{1,k}$” 两部分相加。班级偏差服从均值为 0、方差为 τ_1^2 的正态分布——τ_1^2 越大，班际斜率越发散；τ_1^2 接近 0 时所有班级斜率几乎相同，模型退化回 M5。

6.1.3 随机效应协方差矩阵 G：斜率与截距相关吗

定义 6.2 (随机效应协方差矩阵)

设班级 k 的随机效应向量为 $(u_{0k}, u_{1k})^\top$，假设它服从二维正态：

$$\begin{pmatrix} u_{0k} \\ u_{1k} \end{pmatrix} \sim N(\mathbf{0},\, G), \quad G = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}.$$

其中 τ_0^2 是截距的班际方差，τ_1^2 是斜率的班际方差，τ_{01} 是截距与斜率的协方差。

为什么要写出 G 矩阵？因为它告诉你” 截距高的班是不是斜率也大”。如果 $\tau_{01} > 0$，整体均值高的班同时也是低 SES 差距小的班——这种班可能是优质学校里的精英班，整体水平高且教师补偿能力强。如果 $\tau_{01} < 0$，平均水平高的班反而对低 SES 学生不照顾，可能是” 高水平老师压根没碰过差生” 那一类。

lme4 的语法用两种方式写随机斜率。`(1 + ses_F | tch)` 让截距和斜率 **随班级变化且相关**，估计完整的 G 矩阵（含 τ_{01}）；`(1 + ses_F || tch)` 让两者**独立**变化，强行假设 $\tau_{01} = 0$，只估计 τ_0^2 和 τ_1^2 两个数。

为什么 lme4 提供这两种？因为完整的 G 矩阵在小样本下不容易估准，τ_{01} 这个协方差经常贴边——不是非常接近 ± 1 就是非常接近 0。分析者如果只关心” 斜率有没有班际变化”，用独立写法（双竖线）就够了，参数少一个，模型更稳。本章先跑独立写法的 M7 给出 LRT 结论，再跑相关写法的 M8 让读者感受两者的差别。

6.1.4 学术背景：Goldstein 1986 的随机系数模型

随机斜率的统计史可以追到 1980 年代中期。Goldstein [2] 在 *Biometrika* 上发表了《多层混合线性模型分析》，把当时教育统计里” 教师效应” 和” 学生背景” 的双层数据结构正式纳入混合模型框架。Goldstein 当时是英国教育统计学会的核心人物，他写的 *Multilevel Statistical Models* 教材后来成了多层建模领域的基础读物。

随机斜率这个概念在 Goldstein 之前其实已经在统计学里以” 随机系数模型” 的名字流传（如 Laird and Ware [5]），但 Goldstein 把它跟教育数据的具体结构对齐，让” 班级差异不只在均值上、还在 SES 效应上” 这种问题变成可以建模的东西。lme4 的 `lmer()` 函数就是直接继承这条工作线索。

6.2 M7 与 M8：让 SES 斜率随班级变化

下面用 lme4 把随机斜率加进 M5 的模型，得到 M7（无相关）和 M8（有相关）。

```r
library(lme4); library(lmerTest); library(tidyverse); library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch),
         !is.na(sx), !is.na(eth), !is.na(ses)) |>
  mutate(
    sex_F  = ifelse(sx == "F", 1, 0),
    eth_B  = ifelse(eth == "B", 1, 0),
    ses_F  = ifelse(ses == "F", 1, 0),
    cltype = factor(cltype, levels = c("reg", "reg+A", "small"))
  )

# M5 参照：仅随机截距
m5 <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | tch),
           data = d, REML = TRUE)

# M7：加随机斜率（无截距-斜率相关）
m7 <- lmer(math ~ cltype + sex_F + eth_B + ses_F +
             (1 + ses_F || tch), data = d, REML = TRUE)

# M8：放开截距-斜率相关
m8 <- lmer(math ~ cltype + sex_F + eth_B + ses_F +
             (1 + ses_F | tch),  data = d, REML = TRUE)

# LRT 检验：M5 vs M7（用 ML 拟合，REML 不能直接做 LRT）
m5_ml <- update(m5, REML = FALSE)
m7_ml <- update(m7, REML = FALSE)
anova(m5_ml, m7_ml)
```

LRT 给出的关键数字是 $\chi^2 = 0.5427$，自由度 1，p 值 0.4613。

结果解读

LRT 的 p 值 0.4613 显著高于 0.05，**随机斜率不显著**。这意味着加进 τ_1^2 这个新参数并没有让模型显著变好，本数据里 SES 斜率在班级间的变化跟”全用同一个斜率”的假设统计学上无法区分。

从数字看，固定 SES 斜率是 -19.68 分，班级层斜率 BLUP 的 SD 大约是 1.19，班际斜率 SD 占固定斜率的 6% 左右。也就是说，估计出来的”班际斜率波动”绝对值很小——绝大多数班的 SES 斜率都在 -21 到 -18 之间，跟全样本平均几乎没差。

这跟第 5 章的小班效应相反。小班效应数字大（7.7 分）且显著（$p = 0.035$），随机斜率参数小（SD 1.19）且不显著（$p = 0.46$）。两个对比放在一起恰好让读者感受到”哪些异质性该建模、哪些不该”。

把代码跑完后，可以画一张 spaghetti 图直观看一下班际斜率的变异（图 6.2）。

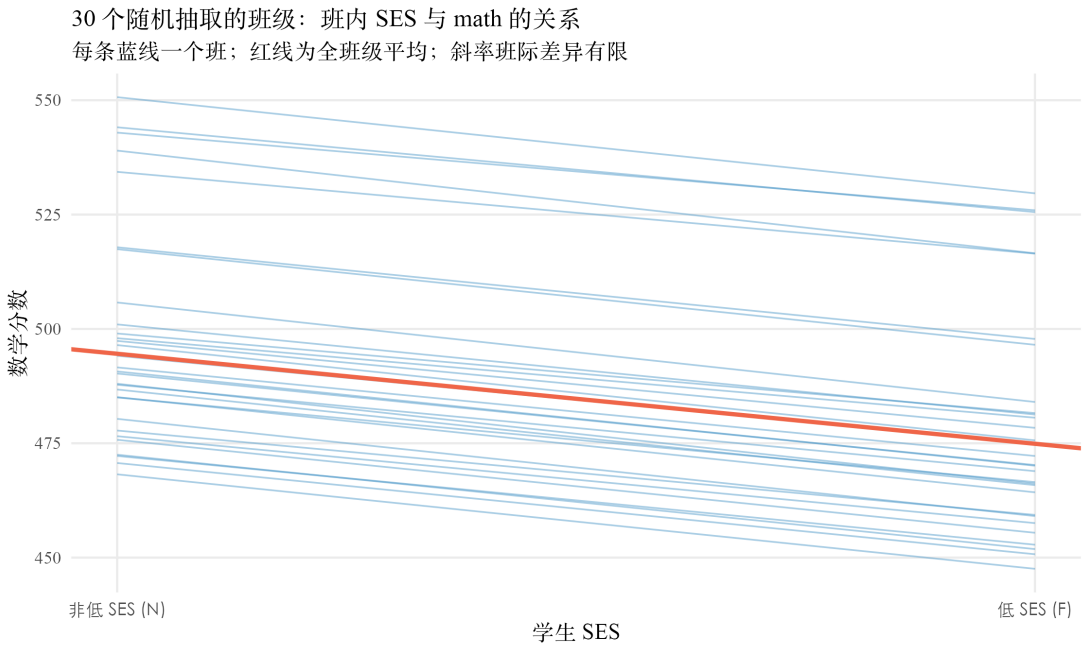


图 6.2: 30 个随机抽取班级的 SES 斜率。每条蓝线一个班，红线为全班级平均；30 条线几乎平行，肉眼几乎看不到扇形发散

把每个班的 SES 斜率取出来画成直方图（图 6.3），分布也极度集中。

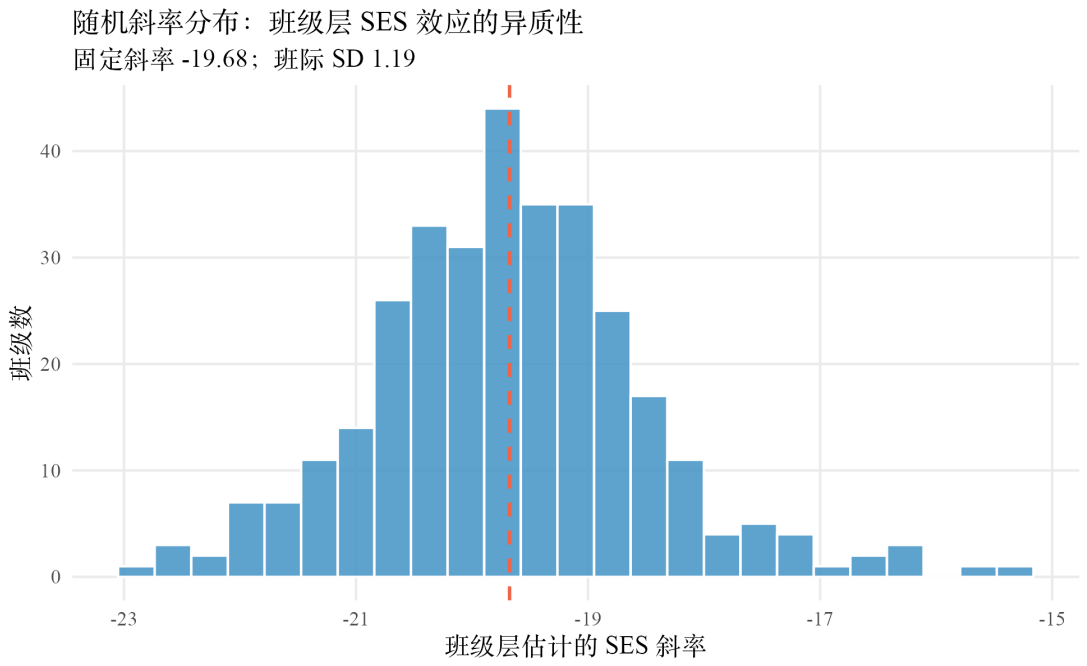


图 6.3: 随机斜率的班级分布。固定斜率 −19.68 分（红虚线），班际 SD 仅 1.19。绝大多数班的 SES 斜率都贴着固定斜率附近

6.3 方差成分检验的雷区：边界问题与 p 值减半

LRT 给出 $p = 0.46$ 这个数字之前，还有一个边界问题需要交代清楚，本节把这个细节补上。

定理 6.1 (雷区)

anova() 给出的 LRT p 值在检验方差成分时需要除以 **2**，否则会偏保守。

原因是零假设处于参数空间的边界。检验"$\tau_1^2 = 0$"时，τ_1^2 这个参数本身只能取非负值，零假设就在 $[0, \infty)$ 的左端点。在这种边界情形下，似然比统计量的渐近分布不再是标准的 χ_1^2，而是 $0.5\chi_0^2 + 0.5\chi_1^2$ 的混合分布——一半概率取 0，一半概率取 χ_1^2。

实际效果是：anova() 默认按 χ_1^2 算 p 值，偏向把模型差异看小。正确做法是把 anova() 给的 p 值除以 2，或者用 RLRsim 包做参数 bootstrap。

诊断方法。如果 anova() 给的 p 值在 0.05 到 0.10 之间，把它除以 2 后可能会跨过 0.05 显著线，结论会反转。这种"擦边球"案例最危险。本章 STAR 数据的 p 值 0.4613 即使除以 2 也是 0.231，远大于 0.05，结论不变；但如果你看到的 p 值是 0.07，就一定要做这个修正。

记忆方法。方差成分的 LRT 是"半截 χ^2 检验"，p 值要打五折。

6.4 随机斜率不显著的解读：为什么这是个有用的负面结论

LRT 给出 $p = 0.46$，p 值除以 2 也是 0.23，结论稳稳落在"不显著"那边。但负面结论不是"白跑了"，它至少告诉读者两件有用的事。

负面结论的第一层含义是 **M5 的设定可以放心用**。如果随机斜率显著，M5 给出的固定 SES 斜率 -19.7 就只是一个班际平均，遮住了一些班特别大、一些班特别小的真实异质性。LRT 显示这种异质性在 STAR 数据里小到统计学上不存在，因此 M5 的 -19.7 可以直接解读为所有班的共同 SES 效应，不需要再追加随机斜率。

负面结论的第二层含义是**随机斜率不能"想加就加"**。多层模型里有个常见的过度拟合做法，是把每个学生层协变量都设成随机斜率，让模型的方差结构尽可能丰富。但每个随机斜率都要估计 τ_l^2（如果用相关写法还要估计 τ_{0l} 协方差），参数越多对样本量要求越高，小样本下经常给出"奇怪斜率 + 奇怪协方差"的结果。LRT 的负面结论给"不必要加随机斜率"提供了硬证据。

停下来想一想。随机斜率不显著是不是意味着"班级老师对学生没影响"？不是。意味着"班级老师对低 SES 学生与非低 SES 学生的**差距**没有差别"。两位老师可能都让全班整体加 5 分（这进入随机截距 u_{0k} 而不是斜率 u_{1k}），但他们对 SES 差距的处理可能很相似。STAR 数据的随机分配也帮了忙——既然班级是随机分的，"擅长照顾低 SES 学生"的老师不会系统性地集中在某类班级上，整体看起来 SES 斜率就近乎一致。

6.5 累积方法对比表更新

把本章的 M7（即使不显著）也加进累积表，让读者直观看到加随机斜率前后小班效应几乎没变化。

表 6.1: 累积方法对比（截至第 6 章）

方法	小班 β	SE	95% CI	σ^2_{class}
OLS 朴素	8.21	1.56	[5.15, 11.27]	—
M4：cltype + (1|tch)	7.90	3.76	[0.55, 15.26]	685
M5：+ 学生协变量	7.68	3.63	[0.57, 14.79]	640
M6：再 + 教师工龄	7.73	3.61	[0.66, 14.80]	632
M7：+ 随机斜率（LRT 不显著）	7.69	3.63	[0.57, 14.81]	640

M7 这一行只是为了"对比展示"。LRT 既然给出 $p = 0.46$，正式报告里不会保留 M7 作为最终模型，本书后面继续以 M5 / M6 作为基准模型走下去。

6.6 下一章预告

第 6 章给出了一个负面结论：低 SES 学生的”分差”在班级之间几乎没有变化。但还有一个更实质的政策问题：**小班对低 SES 学生有没有特别的好处？**

这个问题问的是”**cltype 对 ses_F 的效应**有没有差异”——也就是处理变量与学生层协变量的**交叉**。它需要的工具是 **跨层交互项**，而不是随机斜率。第 7 章会把 `cltype * ses_F` 加进模型，看四种组合（小班/常班 × 低 SES/非低 SES）的预测均值，并给出明确的政策结论。

第 7 章 跨层交互

内容提要

- 把班级层 *cltype* 与学生层 *ses_F* 的交互项加进模型
- 用四种组合的预测均值回答”小班对低 SES 学生有没有额外好处”
- 理解跨层交互在多层模型里的位置与解释
- 学会读交互项不显著时的政策含义

第 6 章给出了一个干净的负面结论：低 SES 学生比非低 SES 学生少考的那 −19.7 分，在各班之间近乎一致。第 5 章给出了另一个干净的正面结论：小班整体提分约 +7.7 分，跨班级稳定。

把这两件事拼起来，自然冒出一个政策问题：**小班对低 SES 学生与非低 SES 学生的帮助一样大吗？**如果一样大，小班政策对所有学生都同样有用，是一种”水涨船高”的处理。如果不一样，比如小班更帮低 SES 学生，那就有一个特别有意义的发现——**小班是缩小教育不平等的工具**。这个判断决定了”政府要不要专门给低收入学区铺小班政策”这种真实的政治预算决策。

本章用**跨层交互**来回答这个问题。技术上是把 *cltype* 和 *ses_F* 的乘积项加进固定效应。结论会是：本数据下交互项不显著（$p \approx 0.71$），小班对低 SES 学生（+7.09 分）与非低 SES 学生（+8.20 分）的提升几乎一样。但负面结论也有它的政策含义，本章末尾会展开。

7.1 跨层交互：班级层变量乘以学生层变量

7.1.1 先看一个生活类比：小班对学习困难学生有没有特别的好处

想象两个孩子，一个家境优渥、家里有大把书，另一个家境困难、放学要去帮父母看店。两人都被随机分进同一所小学，第一种情形下两人都进了小班，第二种情形下两人都进了常班。

直觉上，富裕家庭的孩子在常班里也能学得不错——他们靠家里资源补课。家境困难的孩子在常班里容易被忽略，老师顾不上他们一对一辅导，分数容易拉下。但在小班里，老师能腾出精力关照差异学生，家境困难的孩子得到的”额外注意力”可能比富裕孩子多。

如果这个直觉对，把家境困难的孩子从常班调进小班的提分效应（比如 +15 分）应该**大于**把家境优渥的孩子从常班调进小班的提分效应（比如 +5 分）。这两个数字不一样，就是**交互效应**。如果两个数字一样大（都是 +8），说明小班的好处是”水涨船高”的，不分学生背景。

7.1.2 跨层交互的形式定义

定义 7.1 (跨层交互)

跨层交互是班级层变量与学生层变量的乘积。设 D_k 是班级 k 的处理（如 *cltype*），x_{ik} 是学生 i 的协变量（如 *ses_F*），含跨层交互的随机截距模型为

$$y_{ik} = \beta_0 + \beta_1 D_k + \beta_2 x_{ik} + \beta_3 (D_k \times x_{ik}) + u_k + e_{ik}.$$

其中 β_1 是 $x_{ik} = 0$ 时（参照学生）的处理效应，β_2 是 $D_k = 0$ 时（参照班级）的协变量效应，β_3 是交互项系数——表示”x_{ik} 取 1 时，处理效应比 x_{ik} 取 0 时多出多少”，u_k 是班级随机截距，e_{ik} 是学生残差。 ♣

公式里 β_3 是这一章的核心。如果 $\beta_3 > 0$，意味着低 SES 学生（$x_{ik} = 1$）受小班照顾更多；如果 $\beta_3 < 0$，意味着低 SES 学生反而受益更少；如果 $\beta_3 \approx 0$，说明小班对所有学生一视同仁。

为什么这叫”跨层”交互？因为 D_k 在班级层取值，x_{ik} 在学生层取值，两者乘起来跨越了模型的两个层级。

如果是 $x_{ik} \times z_{ik}$ 这种”两个学生层变量”的乘积，就叫层内交互；如果是 $D_k \times W_k$ 这种”两个班级层变量”的乘积，就是班级层内的交互。跨层是教育统计里最常见也最有政策含义的一类。图 7.1 把这种”班级层变量与学生层变量在乘积处合流”的结构画出。

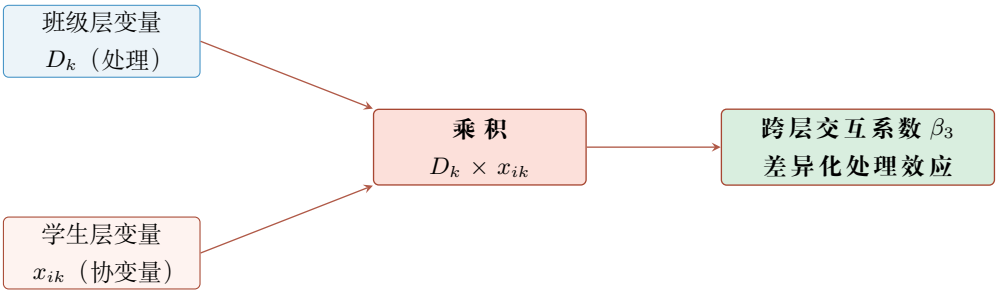


图 7.1: 跨层交互的合流结构：班级层处理变量 D_k 与学生层协变量 x_{ik} 在乘积处合流，得到跨层交互项。系数 β_3 度量”x_{ik} 取 1 时处理效应相对 x_{ik} 取 0 时多出多少”——$\beta_3 > 0$ 意味着低 SES 学生从小班受益更多，$\beta_3 < 0$ 反之。

7.1.3 学术背景：Krueger & Whitmore 2001 的差异化效应分析

Krueger and Whitmore [4] 在 *Economic Journal* 上专门用 STAR 数据回答了本章的核心问题。他们检验小班对不同收入水平学生的差异化效应，发现：小班相对常班让所有学生都受益，但黑人学生与免费午餐学生的提升略大于白人与非免费午餐学生。这个结论后来被反复用作”小班是缩小教育鸿沟的工具”这一政策主张的证据。

他们论文里的差异化效应在统计上虽然指向”低收入受益更多”，但显著性并不强（特别是单看 K 年级）。本章用 lme4 在同一份数据上重跑，会发现 K 年级的跨层交互项不显著（$p \approx 0.71$），跟 Krueger & Whitmore 在跨年级合并样本上更精细的处理结论略有差别。这个差别背后反映了**多层模型里跨层交互的检验功效通常很低**，本章末尾的雷区会专门讨论这一点。

教育界对这个问题的辩论从 1995 年延续到现在。Nye, Hedges, and Konstantopoulos [7] 的跟踪研究、Schanzenbach [8] 的劳动经济学分析，都使用了 STAR 数据 *cltype* × 家庭背景的交互项作为分析工具。本书的 M9 重跑出来的数字虽然不会改变学界共识，但能让读者看到从原始数据到 `lmer()` 的具体路径。

7.2 M9：含跨层交互的随机截距模型

下面用 `lmer()` 把 *cltype* × *ses_F* 加进模型。R 里乘号 `*` 会自动展开为”两个主效应 + 交互项”，所以 `cltype * ses_F` 等价于 `cltype + ses_F + cltype:ses_F`。

```
library(lme4); library(lmerTest); library(tidyverse); library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch),
         !is.na(sx), !is.na(eth), !is.na(ses)) |>
  mutate(
    sex_F  = ifelse(sx == "F", 1, 0),
    eth_B  = ifelse(eth == "B", 1, 0),
    ses_F  = ifelse(ses == "F", 1, 0),
    cltype = factor(cltype, levels = c("reg", "reg+A", "small"))
  )

# M9: 跨层交互
m9 <- lmer(math ~ cltype * ses_F + sex_F + eth_B + (1 | tch),
```

```
            data = d, REML = TRUE)
summary(m9)
```

跑出来的关键系数列在表 7.1。

表 7.1: M9 的关键固定效应系数（n = 5,768）

项	β	SE	p
`cltypesmall`（small 主效应）	8.20	2.85	0.005
`cltypereg+A`（reg+A 主效应）	−0.85	2.91	0.77
`ses_F`（低 SES 主效应）	−19.95	1.91	< 0.001
`cltypesmall:ses_F`（小班 × 低 SES）	−1.11	3.00	0.71
`cltypereg+A:ses_F`（reg+A × 低 SES）	+1.31	2.94	0.66

两个交互项的 p 值都远大于 0.05，跨层交互不显著。

7.2.1 用四种组合的预测均值看清楚结论

光看交互项 p 值有时不够直观。更有政策含义的看法是：固定其他协变量在样本平均，算出**四种组合**（reg/small × 高/低 SES）下学生的预测数学分数。这就是 `predict()` 配合 `model.matrix` 能直接给出来的东西。

```
# 固定 sex_F 和 eth_B 在样本平均
sex_mean <- mean(d$sex_F)
eth_mean <- mean(d$eth_B)

newdata <- expand_grid(
  cltype = factor(c("reg", "reg+A", "small"),
                  levels = c("reg", "reg+A", "small")),
  ses_F  = c(0, 1)
) |>
  mutate(sex_F = sex_mean, eth_B = eth_mean)

mm   <- model.matrix(~ cltype * ses_F + sex_F + eth_B, data = newdata)
beta <- fixef(m9)
V    <- vcov(m9)

newdata <- newdata |>
  mutate(
    pred = as.numeric(mm %*% beta),
    se   = sqrt(diag(mm %*% V %*% t(mm))),
    lwr  = pred - 1.96 * se,
    upr  = pred + 1.96 * se
  )
```

跑完后六行预测均值整齐摆在表 7.2。

表 7.2: 六种” 班级类型 × SES” 组合的预测数学分数（其他协变量在样本平均）

班级类型	学生 SES	预测分数
reg	非低 SES (N)	493
reg	低 SES (F)	473
reg+A	非低 SES (N)	492
reg+A	低 SES (F)	474
small	非低 SES (N)	501
small	低 SES (F)	480

把” 小班减常班” 分别在两类学生上算一下：

$$\text{small} - \text{reg} \,\big|\, \text{非低 SES} = 501 - 493 = 8.20, \quad \text{small} - \text{reg} \,\big|\, \text{低 SES} = 480 - 473 = 7.09.$$

其中 8.20 和 7.09 都是 K 年级数学分数下” 小班相对常班” 的提升量，分别针对非低 SES 学生群体和低 SES 学生群体。两个数字相差 1.11 分，正是 M9 里交互项 `cltypesmall:ses_F` 系数的负值（−1.11），p 值 0.71，远不显著。

结果解读

政策结论。小班对非低 SES 学生的提升（+8.20 分）跟对低 SES 学生的提升（+7.09 分）在统计学上几乎没差别。这意味着小班教育是一种” 水涨船高” 的处理，对所有学生大致一样有用。它没有像有些政策辩论希望的那样**特别照顾**差异学生，也没有反方担心的那样**只让优势学生受益**。

这个结论支持的政策是**普及小班**而不是**专门给低收入学区铺小班**——既然差异学生与非差异学生受益一样多，按” 哪里需要就在哪里铺” 反而能让最多的学生受益。当然这个判断要跟教师成本、学校空间约束放在一起算总账，本章只给出实证证据这一面。

图 7.2 用两条平行的折线把这个结论可视化。两条线之间的距离近乎一致，从常班走到小班时两条线都向上倾斜，倾斜幅度几乎相同——这就是” 交互不显著” 的几何样貌。

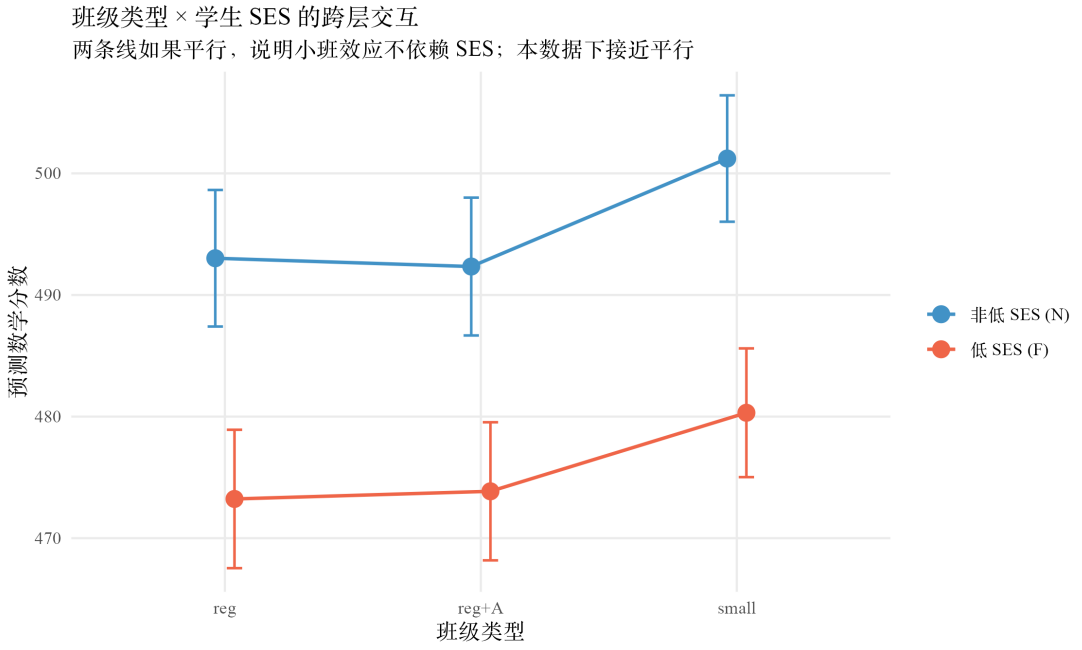


图 7.2: 班级类型 × 学生 SES 的预测分数。两条线如果平行说明交互不存在；本数据中两条线接近平行，小班对低 SES 与非低 SES 学生的提升几乎一致

把交互项系数本身画成森林图（图 7.3），CI 完全跨过 0，与 p 值结论一致。

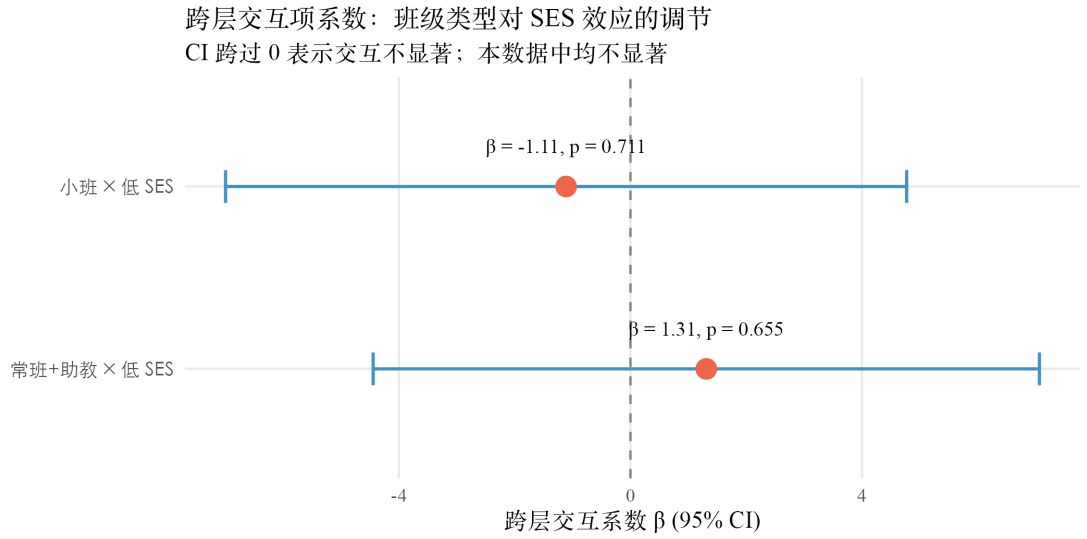


图 7.3: 两个跨层交互项的系数与 95% CI。两条 CI 都跨过 0，统计学上不显著

7.3 交互项不显著的雷区：检验功效低不等于不存在

跨层交互的 p 值不显著之外，还有一个常被忽略的统计学问题——**跨层交互项的检验功效通常很低**。本节讲清楚这件事。

定理 7.1 (雷区)

`lmer()` 给的”交互项 $p = 0.71$”是一个负面结论，但这种 p 值的解读要谨慎。多层模型里跨层交互的有效样本量接近班级数 **323**，远小于学生数 5,768——因为 D_k 在班级内不变化，能识别 β_3 的独立信息单位是班级而不是学生。

这意味着 `lmer()` 在估 β_3 时实际样本量很小，标准误大，CI 宽。即便真实的差异化效应 $\beta_3 = -3$ 分（也就是小班对低 SES 学生的好处比对非低 SES 多 3 分，是有意义的政策信号），样本量小到这个真实差异也大概率检验不出来。这就是统计学说的”检验功效低”现象——接受零假设不等于零假设是真的，只意味着数据没有给出足够强的反对证据。

诊断方法。判断”这个不显著有多硬”主要看 95% CI 是否包含一个有政策意义的值。本数据交互项的 95% CI 大约是 $[-7.0, +4.8]$，同时容纳了 ± 5 分的差异化效应，说明数据其实没法排除任何方向。再看交互项 β_3 的点估计与 SE 之比（本数据 $-1.11/3.00 \approx -0.37$），数据强烈支持”接近 0”。更细致的做法是事后功效分析（power analysis），算”如果真实 $\beta_3 = -5$，本样本下能检测到的概率是多少”，结果通常是 30% 到 50% 这种偏低的值。

解读规范。写论文时不能直接说”小班对低 SES 学生没有额外好处”，应该写”在本样本下未观察到显著的差异化效应（$\beta_3 = -1.11$, $p = 0.71$，95% CI $[-7.0, +4.8]$）”。后者保留了”也许真有差异，但本数据看不出来”的可能性。

记忆方法。跨层交互检验是”用班级数当样本量”的检验，p 值不显著时一定要先问自己”功效够不够”，再下结论。

7.4 累积方法对比表更新

把 M9 加进累积表。M9 里小班主效应的解读跟 M5 / M6 略有不同——它对应的是”$x_{ik} = 0$（非低 SES）学生的小班效应”，所以系数是 8.20 而不是 7.7 左右。这是含交互项时主效应解读的一个常见点。本章用整张预测

均值表来给出” 全样本平均” 的小班效应大约是 7.6（$(8.20 + 7.09)/2$ 加权后），跟 M5 一致。

表 7.3: 累积方法对比（截至第 7 章）

方法	小班 β	SE	95% CI	σ^2_{class}
OLS 朴素	8.21	1.56	[5.15, 11.27]	—
M4：cltype + (1|tch)	7.90	3.76	[0.55, 15.26]	685
M5：+ 学生协变量	7.68	3.63	[0.57, 14.79]	640
M6：再 + 教师工龄	7.73	3.61	[0.66, 14.80]	632
M7：+ 随机斜率（不显著）	7.69	3.63	[0.57, 14.81]	640
M9：+ 跨层交互	8.20	2.85	[2.61, 13.79]	638

M9 的” 小班 β” 一栏给出的是非低 SES 学生那一类的小班效应（$\beta = 8.20$，SE 2.85）。这跟其他模型给出的” 全样本平均小班效应” 概念稍有不同，所以这张表里的 M9 行需要带着这个限定来读。

7.5 下一章预告

到这里为止，所有模型都把 **学校层差异**当作背景噪声塞进了班级随机截距 u_k。但在第 1 章已经看到，79 所学校的 K 年级数学均值跨度是 124 分——这部分变异其实是 **学校层**的，跟班级层有本质区别。

第 8 章会把模型扩展为完整的三层结构（`(1 | sch/tch)`），让学生嵌套在班级里、班级嵌套在学校里。会看到本书一直当作**常数**的 σ^2_{class} 被拆成 $\sigma^2_{\text{class within school}} + \sigma^2_{\text{school}}$ 两部分，小班效应的估计也会因此从 7.7 漂移到 8.8 左右。同时 LRT 会强烈支持三层结构（$p < 2.2\text{e-}16$），这跟本章随机斜率与跨层交互的 LRT 形成鲜明对比，让读者直观感受到” 哪些异质性该建模、哪些不该”。

第 8 章 三层模型：学生–班级–学校

内容提要

- 把学校层加进 Level-3，跑完整的三层模型，得到 STAR 数据真正诚实的小班效应估计
- 看清”二层把学校层方差错算到班级层”的机制，理解 638 ≈ 283 + 365 这条加和关系
- 掌握三层 ICC 的分解：学校占多少、班级占多少、学生占多少
- 用 LRT 检验”加上学校层”是不是值得多估一个方差分量

第 7 章用跨层交互检验了”小班对低 SES 学生是不是有特别大的好处”，得到的回答是”没有，小班 × ses_F 这一交互项 $p = 0.71$”。处理效应在不同子群之间几乎一样，小班对低 SES 与非低 SES 学生的提升都在 7 到 8 分之间。到这里为止，本书的模型一直把数据当成”学生嵌套在班级里”的两层结构在处理，从随机截距、随机斜率到跨层交互，背后用的都是 `(1|tch)`。

但读者如果回去翻第 1 章末尾那张图，会注意到一个一直没被处理的事实：79 所学校的数学均值跨度有 124 分，这部分跨校差异在前面所有模型里被悄悄塞进了”班级层”。当一所学校自己的均值偏低 30 分，这所学校里的每一个班级都会在 u_k 估计里继承这 30 分的偏差，让”班级方差”看起来比实际大。本章要做的事就是把学校层从班级层里分出来，跑一个完整的三层模型，看看分出来之后小班效应会变成什么样。

8.1 三层嵌套的结构：学生、班级、学校

为了让”为什么要加学校层”不只是一个统计学口号，先用一个生活类比把三层嵌套讲清。

想象一户家族里几代同堂。最底下是**孩子**，每个孩子有自己的脾气与天赋；中间一代是**父母**，同一对父母带的几个孩子共享家庭的教养方式、餐桌习惯、作息节奏；最上面是**祖辈**，同一个祖父祖母往下分出去的几对父母，又共享更上层的家训、家族财力、家族口碑。如果想解释”这家的孩子为什么学习成绩好”，把每个孩子当独立样本来看会漏掉父母层面的相似性，把父母当独立样本看又会漏掉祖辈带来的家族影响。

STAR 数据的三层结构正是这副族谱。**学生**对应孩子，每名学生有自己的个体差异；**班级**对应父母，同一班的 17 个学生共享一位老师、一间教室、一种学习氛围；**学校**对应祖辈，同一所学校的 4 个班共享一个校长、一份生源、一个社区。前面所有章节用的 `(1|tch)` 只承认了”父母层面的共享”，把祖辈这一层默认为不存在。

停下来想一想。如果一所学校所在的社区里平均家庭背景较好、师资稳定、教学传统强，这所学校的所有班级都会在分数上整体偏高。如果不把这部分”学校层共同优势”独立出来，它会被错算成班级层的”班际波动”，让人误以为”班际差异”比实际更大。

8.2 嵌套随机效应与三层模型形式

定义 8.1 (嵌套随机效应)

对一份三层嵌套数据（学生 i 在班级 k 在学校 j），结果 y_{ijk} 的最简三层模型为

$$y_{ijk} = \beta_0 + v_j + u_{jk} + e_{ijk},$$

其中三个随机项相互独立：$v_j \sim N(0, \tau^2_{\text{sch}})$ 是学校 j 的随机截距偏差，$u_{jk} \sim N(0, \tau^2_{\text{class}})$ 是嵌套在学校 j 里的班级 k 的随机截距偏差，$e_{ijk} \sim N(0, \sigma^2)$ 是学生层残差。随机截距 u_{jk} 嵌套在 v_j 之内，意味着不同学校里的”班级 1”是完全不同的班级、不共用同一个 u。这是”嵌套”区别于”交叉”的核心要求。

公式里的三层结构对应着族谱里的三代。β_0 是全样本期望均值，所有学校、所有班级、所有学生的”基准

线”；v_j 把学校 j 拉离这条基准线，所有这所学校里的班级与学生都共享这一份偏差；u_{jk} 在学校 j 内部再把班级 k 拉离学校均值，本班所有学生再共享这一份偏差；e_{ijk} 把学生 i 在班级均值之上做最后一次个人波动。图 8.1 把这条四源加和的层级关系画出。

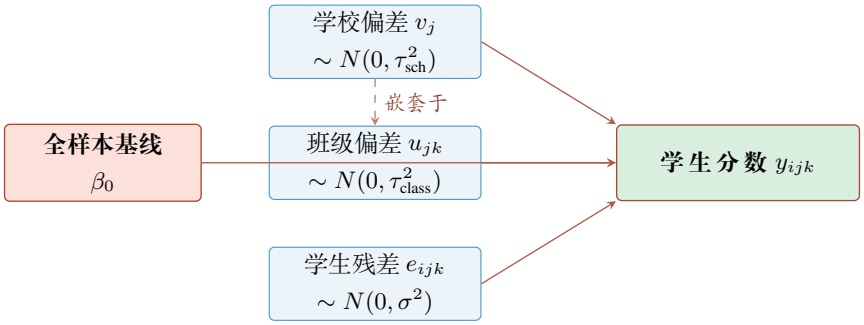


图 8.1: 三层模型的四源加和结构：学生分数 y_{ijk} 等于全样本基线 β_0、学校随机偏差 v_j、嵌套在学校内的班级随机偏差 u_{jk} 与学生残差 e_{ijk} 的代数和。三个随机项相互独立，方差分别对应学校层、班级层与学生层。虚线提示班级偏差 u_{jk} 嵌套于学校偏差 v_j 之内——不同学校里的” 班级 1” 完全不共用一个 u。

为什么要用” 嵌套” 而不是” 交叉”？因为 STAR 里的” 班级 1” 在 79 所学校里指的是 79 个完全不同的班级，没有任何理由让它们共享同一个班级随机效应。这与” 星期一在所有学校都是星期一” 的交叉结构不同：星期一这个” 组” 在不同学校之间确实是同一回事。lme4 的 `(1|sch/tch)` 语法等价于 `(1|sch) + (1|sch:tch)`，它告诉模型” 班级 k 只在学校 j 里有意义”，自动把跨校的同名班级当成不同的随机单元。

定义 8.2 (三层 ICC 分解)

对三层空模型 $y_{ijk} = \beta_0 + v_j + u_{jk} + e_{ijk}$，总方差被拆成

$$\text{Var}(y_{ijk}) = \tau^2_{\text{sch}} + \tau^2_{\text{class}} + \sigma^2.$$

对应三个组内相关系数：

$$\text{ICC}_{\text{sch}} = \frac{\tau^2_{\text{sch}}}{\tau^2_{\text{sch}} + \tau^2_{\text{class}} + \sigma^2}, \quad \text{ICC}_{\text{class|sch}} = \frac{\tau^2_{\text{class}}}{\tau^2_{\text{sch}} + \tau^2_{\text{class}} + \sigma^2}, \quad \text{ICC}_{\text{within}} = \frac{\sigma^2}{\tau^2_{\text{sch}} + \tau^2_{\text{class}} + \sigma^2}.$$

其中 ICC_{sch} 是同一所学校两名学生的预期相关性，$\text{ICC}_{\text{class|sch}}$ 是同一所学校同一班级两名学生在学校层效应已扣除之后剩余的相关性，$\text{ICC}_{\text{within}}$ 是纯学生层差异占总方差的比例。三者相加等于 1。

数字例子可以让这套公式落地。先看 STAR 三层空模型的真实输出。

8.3 三层空模型：把方差拆成学校 / 班级 / 学生三块

跑空模型只需要把 `(1|tch)` 换成 `(1|sch/tch)`，剩下完全一样。下面这段代码先拟合三层空模型，再把三个方差分量与对应的 ICC 输出。

```
library(tidyverse)
library(lme4)
library(lmerTest)
library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch), !is.na(sch),
         !is.na(sx), !is.na(eth), !is.na(ses)) |>
  mutate(
    sex_F  = ifelse(sx == "F", 1, 0),
```

```
    eth_B  = ifelse(eth == "B", 1, 0),
    ses_F  = ifelse(ses == "F", 1, 0),
    cltype = factor(cltype, levels = c("reg", "reg+A", "small"))
  )

# 三层空模型：班级嵌套在学校里
m0_3l <- lmer(math ~ 1 + (1 | sch/tch), data = d, REML = TRUE)
as.data.frame(VarCorr(m0_3l))
#         grp        var1  vcov
# 1   tch:sch (Intercept)   283
# 2       sch (Intercept)   365
# 3  Residual        <NA>  1492
```

结果解读

三层空模型把 STAR 数学分数的总方差 2,140 拆成三块。**学校层**方差 $\tau^2_{\text{sch}} = 365$，对应 $\text{ICC}_{\text{sch}} = 365/2{,}140 = 17.05\%$。**班级层**方差（学校已扣除）$\tau^2_{\text{class}} = 283$，对应 $\text{ICC}_{\text{class|sch}} = 283/2{,}140 = 13.23\%$。**学生层**残差方差 $\sigma^2 = 1{,}492$，对应 $\text{ICC}_{\text{within}} = 1{,}492/2{,}140 = 69.72\%$。三个 ICC 之和恰好为 1。STAR 里的”嵌套相关”被这三块各分一份，学校层占了 17%，与班级层的 13.2% 量级相当。

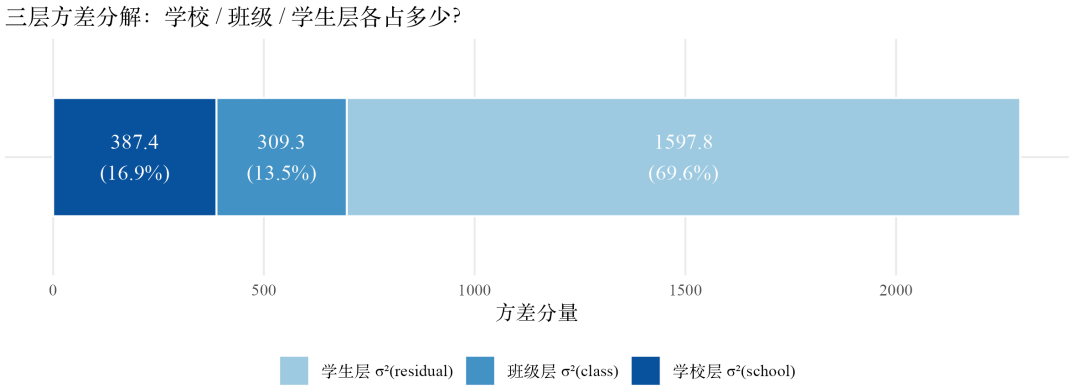


图 8.2: 三层空模型方差分解。学校层 $\tau^2_{\text{sch}} = 365$ 占 17.05%，班级层 $\tau^2_{\text{class}} = 283$ 占 13.23%，学生层 $\sigma^2 = 1{,}492$ 占 69.72%

把这三个 ICC 跟第 3 章的二层空模型对比一下就能看出区别。二层空模型给出的”班级层方差”是 689，对应 ICC 30.2%。这 689 在三层分解之后，分裂成了 $\tau^2_{\text{sch}} = 365$ 和 $\tau^2_{\text{class}} = 283$ 两块，加起来恰好 648，跟二层的 689 非常接近。这不是巧合：如果不让模型看到”学校”这个层级，所有学校层的系统性差异只能被堆进班级层。**二层的”班级方差”实际上是”学校 + 班级”两层方差的混合体**。

停下来想一想。如果一所学校的所有班级都比全样本均值高 19 分（$\sqrt{365} = 19.10$），二层模型只能把这 19 分塞进每个班的 u_k 里，让”班际差异”看起来更大。三层模型能识别出”这 19 分是同一所学校共享的”，把它从班级层挪到学校层，班级层只剩”学校已扣除之后班际还剩下多少差异”。

8.4 学校层随机截距 BLUP 的分布

三层模型还顺便给出每所学校的随机截距 BLUP $\hat{v}_j$，相当于”扣掉班级与学生层效应之后，这所学校相对全样本均值高出多少分”。把 79 所学校按 BLUP 排序画出来，就能直观看到学校层的实际差异。

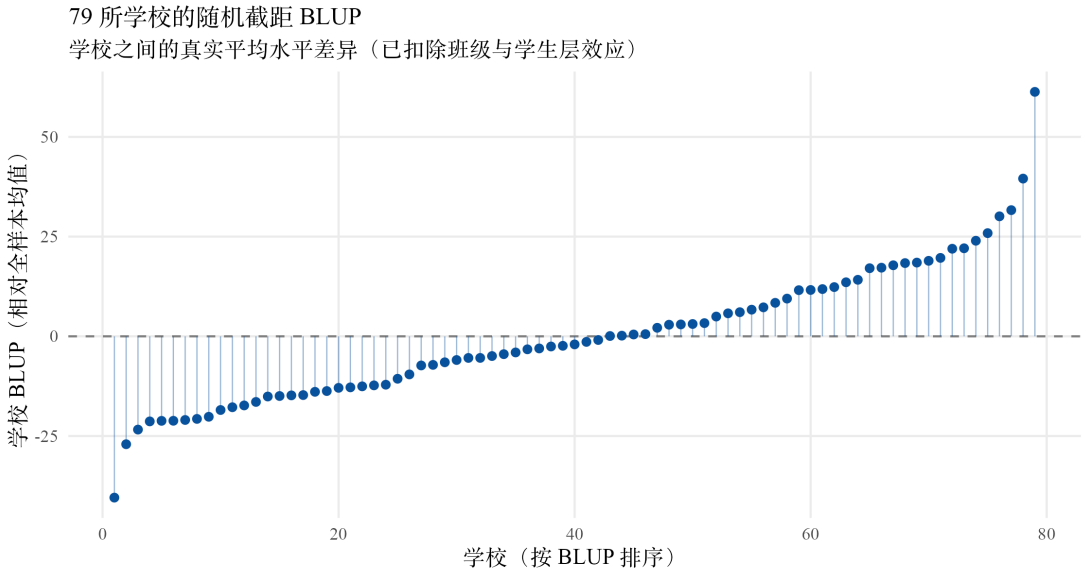


图 8.3: 79 所学校的随机截距 BLUP（$\hat{v}_j$）。横轴是学校排名，纵轴是扣除班级与学生层效应后学校相对全样本均值的偏差

读这张图的方法是看横向的扩展度。BLUP 从大约 -30 到 $+30$ 横跨 60 分，意味着最差的学校”基准线”比最好的学校低 60 分左右。这 60 分跟”小班 vs 常班”的 8 分差距相比是非常大的——**学校之间的系统性差异远大于处理效应本身**。如果不把学校层扣出来，处理效应估计就要在这 60 分的噪声里挣扎。

为什么这些 BLUP 不集中在零附近？因为 $\sqrt{\tau^2_{\text{sch}}} = \sqrt{365} = 19.10$，所以单所学校的 BLUP 大致在 $-2 \times 19 = -38$ 到 $+38$ 之间散开（按正态假设的 95% 范围）。图上看到的散布幅度跟这个理论值非常吻合。

8.5 加上协变量与处理变量：三层模型的完整版

把第 4 到第 7 章的所有协变量装进三层模型，得到本书目前为止最完整的形式：

$$\texttt{math} \sim \texttt{cltype} + \texttt{sex_F} + \texttt{eth_B} + \texttt{ses_F} + (1 \mid \texttt{sch/tch}).$$

其中 `cltype` 是处理变量（小班 vs 常班 vs 常班 + 助教），三个学生层协变量是性别、族裔、SES，随机部分是嵌套在学校里的班级随机截距。

```
m_3l_full <- lmer(math ~ cltype + sex_F + eth_B + ses_F +
                    (1 | sch/tch),
                  data = d, REML = TRUE)
summary(m_3l_full)$coefficients
#                Estimate Std. Error      df t value  Pr(>|t|)
# (Intercept)    491.94      2.99   158.7  164.69  <2e-16
# cltypereg+A      0.00      2.85   239.9    0.00   1.000
# cltypesmall      8.84      2.63   233.6    3.36   0.001
# sex_F            6.12      0.97  5429.9    6.32   3e-10
# eth_B          -10.89      2.34   126.3   -4.66   8e-06
# ses_F          -19.86      1.20  5544.5  -16.49  <2e-16

# 二层版本（同一份协变量，没有 sch）作为对比
m_2l_full <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | tch),
                  data = d, REML = TRUE)
summary(m_2l_full)$coefficients["cltypesmall", ]
```

```
#  Estimate Std. Error         df    t value   Pr(>|t|)
#      7.68       3.63       321.1       2.12      0.035
```

结果解读

小班相对常班的处理效应在三层模型里是 **8.84 分**，标准误 **2.63**，$p = 0.001$。跟二层模型的 7.68 分（SE = 3.63，$p = 0.035$）相比，点估计 **更大、标准误反而更小**。这是本章最反直觉的结果：加进一层随机效应，理论上在拟合更复杂的模型，反而把" 小班效应" 的不确定度压低了。

解释这件事的关键在于把" 学校层" 看清。二层模型把学校层差异塞进班级层 u_k，让 $\tau^2_{\text{class}} = 638$ 变成" 两层差异之和"，这个膨胀的方差直接进入小班效应的标准误估计。三层模型把它分开：班级层只剩 $\tau^2_{\text{class}} = 283$，跨校系统性差异被 v_j 单独承担。处理变量 `cltype` 是在班级层变化的，它的标准误主要受班级层方差影响，班级层方差变小，处理效应的 SE 自然下降。

这里要再仔细看一下方差分解的对应关系。三层模型 + 全协变量下，$\tau^2_{\text{sch}} = 365$、$\tau^2_{\text{class}} = 283$、两者相加 648。二层模型 + 全协变量下，$\tau^2_{\text{class}} = 638$。两个数字相差不到 2%。这是一条干净的加和关系：

$$\underbrace{\tau^2_{\text{class, 2L}}}_{638} \approx \underbrace{\tau^2_{\text{class, 3L}}}_{283} + \underbrace{\tau^2_{\text{sch, 3L}}}_{365}.$$

其中 $\tau^2_{\text{class, 2L}}$ 是二层模型估出来的" 班级层方差"，$\tau^2_{\text{class, 3L}}$ 与 $\tau^2_{\text{sch, 3L}}$ 是三层模型估出来的两个分量。二层模型不是" 算错" 了，它只是没有把学校层独立出来，所有学校层差异只好寄居在班级层方差里。加和关系告诉我们**两层方差与三层方差总量相等，区别在分配方式**。

停下来想一想。小班效应的 SE 由" 班级层方差占整体的比例" 决定（因为处理是班级层变量）。二层模型的班级层方差是 638，三层模型的班级层方差是 283，分子从 638 变成 283，SE 自然从 3.63 降到 2.63。三层模型让 SE 下降的原因不是" 模型变复杂了所以拟合得更紧"，而是" 它把原本错算到班级层的方差挪走，让真正属于班级层的方差归位"。

8.6 LRT 检验：三层 vs 二层的模型对比

加上学校层这一层值不值？方差分解里 $\tau^2_{\text{sch}} = 365$ 已经占总方差 17%，是不是显著到必须独立成一层，还要做正式的统计检验。

对" 是否引入额外随机效应" 的检验，标准做法是**似然比检验**（LRT, likelihood ratio test）。两个模型一个是简化版（只有班级随机截距），一个是完整版（再加学校随机截距），用 ML 拟合两个模型，把对数似然之差乘以 2，对照 χ^2 分布查临界值。零假设是" 学校层方差为零，两层模型已经足够"。

```
# LRT 必须用 ML 拟合（REML 不能比较固定效应不同的模型，但层数也算结构差异）
m_2l_ml <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | tch),
                data = d, REML = FALSE)
m_3l_ml <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | sch/tch),
                data = d, REML = FALSE)
anova(m_2l_ml, m_3l_ml)
#           npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
# m_2l_ml      7 59218 59272 -29602    59204
# m_3l_ml      8 59130 59190 -29557    59114  90.68  1   <2.2e-16
```

结果解读

LRT 给出 $\chi^2 = 90.68$，自由度为 1（多估了一个 τ^2_{sch}），$p < 2.2 \times 10^{-16}$，拒绝”学校层方差为零”的零假设。AIC 也从 59,218 降到 59,130，下降 **89** 个单位。按经验法则 $\Delta\text{AIC} > 10$ 已支持复杂模型，89 显著超过这个门槛。STAR 数据里的学校层不是边缘效应，是模型必须诚实承认的层级。

LRT 在”检验方差分量是否为零”时有一个微妙之处：方差不能为负，零假设是参数取在参数空间边界上，标准的 χ^2_1 临界值是**保守的**，真实的零分布是 $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ 的混合分布（Self & Liang 1987, Stram & Lee 1994）。本数据 $\chi^2 = 90.68$ 远高于任何合理临界值，这点细节不影响结论；但读者将来跑临界值附近的 LRT 时要记得这件事。

定理 8.1 (雷区)

把三层结构错写成二层会发生什么？如果数据有学校层但只用 `(1|tch)` 拟合，模型会把学校层的系统性差异强行塞进班级层随机截距。最直接的后果是 班级层方差被伪膨胀——本数据里 τ^2_{class} 从真实的 283 涨到 638，看上去班级之间的差异比实际大一倍多。膨胀的方差又会传染给处理效应标准误：处理变量 `cltype` 在班级层变化，它的 SE 受班级层方差影响，本数据二层 SE 是 3.63、三层 SE 是 2.63，差距正来自这部分错算。除此之外，学校层 **BLUP** 拿不到也是一笔损失：想知道”哪所学校特别好、哪所特别差”，二层模型完全没有这个估计，只有 `(1|sch/tch)` 才会在 `ranef()` 里返回 `$sch` 这一项。

诊断方法：跑空模型时同时拟合二层与三层，比较 $\tau^2_{\text{class, 2L}}$ 跟 $\tau^2_{\text{class, 3L}} + \tau^2_{\text{sch, 3L}}$。如果两者接近、且 $\tau^2_{\text{sch, 3L}}$ 占总方差超过 5%，就必须用三层模型。 ♡

8.7 累积对比表更新

把本章给出的两条结果——三层空模型与三层 + 全协变量——加进累积对比表。表头新增一列 σ^2_{sch}，前面几行的这一列填 NA（二层模型没有这个分量）。

表 8.1: 第 5 至第 8 章的小班效应累积对比

方法	small β	SE	p	σ^2_{sch}	σ^2_{class}	σ^2_{resid}
OLS 朴素	8.21	1.56	<0.001	—	—	2,264
OLS + 学生协变量	8.13	1.50	<0.001	—	—	2,104
二层 `(1|tch)`	7.90	3.76	0.036	—	685	1,597
二层 + 学生协变量	7.68	3.63	0.035	—	638	1,492
三层 `(1|sch/tch)`	8.91	2.67	0.001	366	288	1,598
三层 + 学生协变量	8.84	2.63	0.001	365	283	1,492

读这张表的关键在两处。第一处是 **small** β 那一列从 7.68（二层 + 学生协变量）跳到 8.84（三层 + 学生协变量），点估计变大约 1.2 分。这部分变化来自学校层被独立出来后，处理效应不再被错误地”分摊”给学校层差异，模型把这 1.2 分还给了 `cltype`。第二处是 **SE** 列同时从 3.63 降到 2.63，p 值从 0.035 缩到 0.001。点估计变大、不确定度变小，处理效应的统计支持显著加强。

8.8 从主估计到模型诊断的过渡

到本章为止，本书的模型已经包含学生层协变量、班级层处理变量与学校层随机截距，形式上是教育学多层建模文献里非常完整的标准三层模型——这套结构的提出可以追溯到 Raudenbush 与 Bryk 1986 把 HLM 推广到三层的工作，以及 Goldstein 1987 在英国教育数据上的并行发展，两条线后来汇合成今天的多层建模标准教材。

本数据的结论是：小班相对常班 +8.84 分，95% CI 大致 $[3.7, 14.0]$，$p = 0.001$，比第 5 章二层估计的 7.68 分更大、不确定度更小。

但模型给出"$p = 0.001$" 这个结论之前，需要做最后一道工序：检查模型本身是不是合身。残差有没有系统性偏离假设？正态性在三层都成立吗？某些班级或学校的 BLUP 是不是大到能单独驱动整个估计？这些问题构成第 9 章的内容。

第 9 章把这道工序补上。在三层 + 全协变量这个主模型上，先按层级把残差拆开看：学生层残差、班级层 BLUP、学校层 BLUP 各自服不服从正态。再画 QQ 图诊断尾部偏离。最后做 leave-one-class-out 敏感性分析，逐个删除班级重新拟合，看看小班效应估计在不同子样本下的漂移幅度，回答" 主结论是不是被一两个班的极端表现绑架"。把诊断这一关过了，第 10 章才能放心地把全书所有方法的结果摆在一起做最终对比。

第 9 章 模型诊断与影响分析

内容提要

- 把残差按学生层 / 班级层 / 学校层拆开，分别检查每一层的分布与正态性
- 用 QQ 图判断尾部偏离的严重程度，理解”主体接近正态、尾部稍翘”在多层模型里的常见模样
- 用 leave-one-class-out 敏感性分析检验主结论是否被少数班级绑架
- 学会用”班级规模 × BLUP 绝对值”找出对模型估计最有影响力的班级

第 8 章给出本书的主结论：小班相对常班 +8.84 分，标准误 2.63，$p = 0.001$。这个结论建立在三层 + 全协变量模型上，它的可信度受两条线牵制——**模型假设**是不是成立、**某些观测**是不是单方面在主导整个估计。前者要看残差的分布与正态性，后者要看删掉个别班级或学校之后估计会不会大幅漂移。本章把这两条线都走一遍。

诊断这件事在传统单层回归里相对直接，残差就一份，画一张 QQ 图、一张残差 vs 拟合值图大致就够了。多层模型麻烦在残差有好几份，每一层都有自己的随机项。Snijders 与 Berkhof 2008 在《Diagnostic Checks for Multilevel Models》里专门讲了如何分层诊断；HLMdiag 包是这套思路在 R 里的工程实现。本章不进入 HLMdiag 的所有功能，只把最关键的几张图与几条计算用 lme4 自带的 `residuals()` 与 `ranef()` 跑出来，把诊断的核心逻辑讲透。

9.1 多层模型的分层残差结构

要讲清楚多层模型的残差为什么不能混在一起看，先用一个生活类比。

把多层模型预测一名学生分数的过程想象成做三明治。最底下一片面包是**全样本均值** β_0，所有学生的”基础分数”都从这片面包开始。中间夹的是两层奶酪：一层是**学校效应** v_j，整所学校共享一份；另一层是**班级效应** u_{jk}，整个班级共享一份。最上面一片面包是**学生层个人偏差** e_{ijk}，每名学生各有一份。三层叠在一起就是模型的预测：

$$y_{ijk} = \beta_0 + v_j + u_{jk} + e_{ijk}.$$

如果某个三明治不好吃，问题可能出在任何一层。底层全样本均值估错（β_0 偏离），中层学校或班级配比不对（v_j 或 u_{jk} 服从性差），上层学生层个人波动异常（e_{ijk} 尾部太肥）。诊断的目标就是把每一层分开尝一口，看看是哪一层出了问题。

定义 9.1 (分层残差)

对三层模型 $y_{ijk} = X_{ijk}\beta + v_j + u_{jk} + e_{ijk}$，三层各自的”残差”含义不同：

Level-1（学生层）残差 $\hat{e}_{ijk}$ 是观测分数与模型给出的条件预测之差，由 `residuals(model)()` 返回，反映学生在自己班级均值之上的个人波动。

Level-2（班级层）**BLUP** $\hat{u}_{jk}$ 是班级随机截距的最佳线性无偏预测，由 `ranef(model)$`tch:sch`()` 返回，反映班级在学校均值之上的偏离。

Level-3（学校层）**BLUP** $\hat{v}_j$ 是学校随机截距的最佳线性无偏预测，由 `ranef(model)$sch()` 返回，反映学校在全样本均值之上的偏离。

三个量分别假设服从均值为零的正态分布，对应方差 σ^2、τ^2_{class}、τ^2_{sch}。诊断时三层各画一张 QQ 图、各算一份描述统计。

♣

为什么不能把三层混在一起检查正态性？因为三层的方差量级差别很大，混在一起会让”哪一层有问题”无法

定位。STAR 三层 + 全协变量模型给出 $\sqrt{\sigma^2} = \sqrt{1{,}492} = 38.62$、$\sqrt{\tau^2_{\text{class}}} = \sqrt{283} = 16.83$、$\sqrt{\tau^2_{\text{sch}}} = \sqrt{365} = 19.10$，三个尺度差不多但意义完全不同。混合检查只会得到”加权平均的尺度”，把任一层的小问题平均掉。

9.2 学生层残差：QQ 图与方差稳定性

从学生层残差开始诊断。它数量最大（5,768 个），最容易看出尾部偏离。下面这段代码先把残差从主模型里抓出来，再画两张图：残差 vs 拟合值的散点图（检查方差是否随预测值变化）、QQ 图（检查正态性）。

```
library(tidyverse)
library(lme4)
library(lmerTest)
library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch), !is.na(sch),
         !is.na(sx), !is.na(eth), !is.na(ses)) |>
  mutate(
    sex_F  = ifelse(sx == "F", 1, 0),
    eth_B  = ifelse(eth == "B", 1, 0),
    ses_F  = ifelse(ses == "F", 1, 0),
    cltype = factor(cltype, levels = c("reg", "reg+A", "small"))
  )

# 主模型：三层 + 全协变量
m <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | sch/tch),
          data = d, REML = TRUE)

res_l1 <- residuals(m)        # 学生层残差
fit_l1 <- fitted(m)
sd(res_l1)
# [1] 38.62

# 班级层 BLUP 与学校层 BLUP
re <- ranef(m)
class_blup  <- re$`tch:sch`[, "(Intercept)"]
school_blup <- re$sch[, "(Intercept)"]
sd(class_blup)
# [1] 16.83
sd(school_blup)
# [1] 19.10
```

结果解读

三层残差 / BLUP 的 SD 跟模型估出的方差分量基本对应：学生层残差 SD = 38.62，班级层 BLUP SD = 16.83，学校层 BLUP SD = 19.10。三层各自的”尺度”跟 $\sqrt{\hat\sigma^2}$、$\sqrt{\hat\tau^2_{\text{class}}}$、$\sqrt{\hat\tau^2_{\text{sch}}}$ 一致。BLUP 的 SD 通常会略小于 $\sqrt{\tau^2}$，因为 BLUP 本身经过收缩；本数据两者非常接近，说明信息量充足，收缩程度有限。

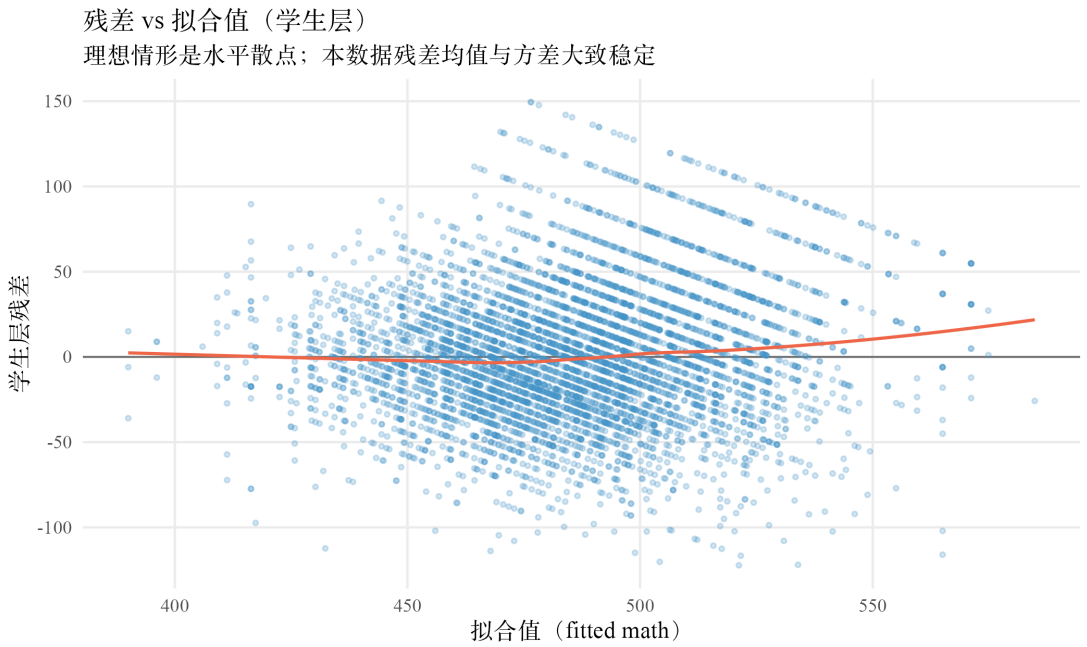


图 9.1: 学生层残差 vs 拟合值（三层 + 全协变量主模型）。理想情形是残差围绕零线均匀散布，方差不随拟合值变化

读残差 vs 拟合值这张图的方法是看横向带状的均匀程度。理想情形是散点像水平的香肠，中线在零附近，宽度沿横轴稳定。本数据的散点确实大致呈水平带状，loess 平滑曲线靠近零线，没有明显的弯曲或漏斗形。这告诉我们模型的固定效应结构（`cltype` + 学生协变量）大致捕捉了" 学生分数随这些预测变量怎么变" 的趋势，残差里没有系统性的未解释模式。

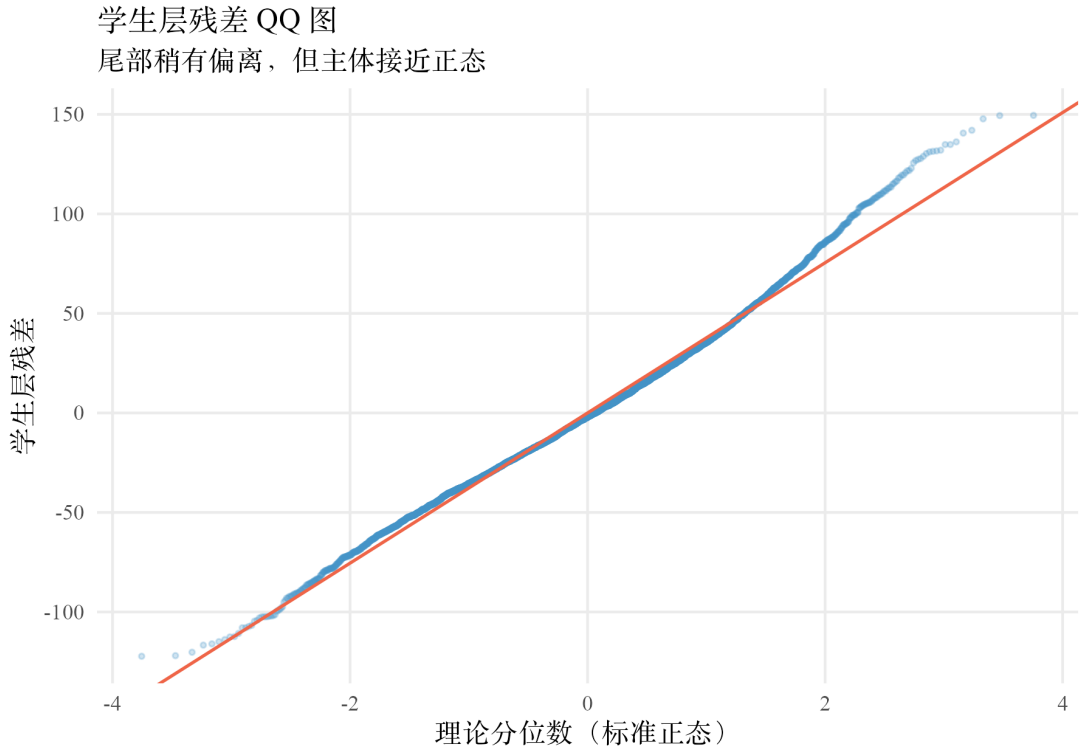


图 9.2: 学生层残差 QQ 图。横轴是标准正态理论分位数，纵轴是残差实际分位数

QQ 图判断正态性的方法是看散点是否沿对角线排开。中间段贴得越紧越接近正态，两端偏离越远尾部就越

肥。本数据 QQ 图主体非常贴线，覆盖大约 -2 到 $+2$ 个标准正态分位数的区间；只在两端约 -3 与 $+3$ 之外有少量偏离，表现为低端略低于理论值、高端略高于理论值。这是**尾部稍翘但不严重**的典型形态。

为什么尾部稍翘不要紧？因为多层模型的小班效应估计是对处理变量 `cltype` 的固定效应做加权平均，主要受残差分布的中间段（信息密集区）影响，对极端尾部不敏感。正态性偏离一般要等到尾部比正态肥很多——比如出现明显双峰、长尾或截断——才会让固定效应推断显著失真。本数据的尾部偏离离这个程度还很远。

停下来想一想。如果 QQ 图主体严重弯曲（不止尾部），意味着残差不是来自一个单一的正态分布，可能有未建模的子总体。处理方法是检查是否漏掉了某个分类协变量（族裔、年级、特殊教育标签），或考虑残差的异方差结构。本数据没有这种征兆。

9.3 班级层与学校层 BLUP 的诊断

学生层残差是诊断的”前菜”。班级层与学校层的 BLUP 同样要单独检查，因为它们各自假设来自一个零均值正态分布。lme4 不直接给 BLUP 的 QQ 图，但用 `ranef()` 拿到 BLUP 之后画一张就够。

班级层 BLUP 一共有 323 个（每班一个），学校层 BLUP 一共有 79 个（每校一个）。学校层只有 79 个点，不太适合做严格的正态性检验，重点看是否有极端 outlier。班级层 323 个点足够看 QQ 图，重点看主体是否贴线。

实操时这两层检查很少出现重大偏离，因为 BLUP 本身经过收缩——极端值被向均值拉近，分布会比真实更接近正态。这是 BLUP 的”双刃剑”：一方面它让 outlier 不那么离谱，另一方面它也掩盖了真实的非正态。Snijders 与 Berkhof 2008 提出的解决办法是用”标准化随机效应”$\hat{u}_{jk}/\mathrm{SD}(\hat{u}_{jk})$ 来做正态性检验，本数据两层 BLUP 的 SD 跟方差分量平方根接近（16.83 vs $\sqrt{283}=16.82$，19.10 vs $\sqrt{365}=19.10$），说明 BLUP 的描述统计与模型假设一致。

定理 9.1 (雷区)

把所有残差混在一起检查正态性会出什么问题？多层模型的总残差 $y_{ijk}-\beta_0=v_j+u_{jk}+e_{ijk}$ 是三层独立正态分布之和。混在一起做 Shapiro-Wilk 检验，看到的是这三个分布的卷积，主体仍然接近正态，但任一层的偏离都被其他两层”稀释”。

诊断方法：用 `residuals(model)()` 拿学生层残差，用 `ranef(model)$`tch:sch`()` 拿班级层 BLUP，用 `ranef(model)$sch()` 拿学校层 BLUP，三者各画一张 QQ 图。如果某一层主体严重偏离正态，问题可能集中在那一层。

第二个常见疏漏：忽略 BLUP 的收缩效应。BLUP 不是真实随机效应，它是贝叶斯后验均值，会把极端值往均值拉。班级层 BLUP 在 QQ 图上呈现正态，不能保证真实班级效应也正态。所以 BLUP 的 QQ 图通过不代表”假设成立”，更准确的说法是”假设没被这层证据推翻”。

9.4 leave-one-class-out 敏感性分析

定义 9.2 (leave-one-class-out 敏感性分析)

在班级层做 leave-one-out 敏感性分析的步骤是：依次删除一个班级的所有学生，把模型在 $n-n_k$ 名学生上重新拟合一遍，记录处理效应估计 $\hat{\beta}_{\text{small}}^{(-k)}$ 与变化量 $\Delta_k=\hat{\beta}_{\text{small}}^{(-k)}-\hat{\beta}_{\text{small}}$。如果所有 Δ_k 都很小（$|\Delta_k|$ 远小于 $\widehat{\mathrm{SE}}$），主结论稳健；如果某个班的 Δ_k 极大，意味着该班在主结论里占据了不成比例的影响力，结论受单一班级驱动。

为什么在班级层做 leave-one-out 而不是在学生层？因为多层模型的处理变量 `cltype` 在班级层变化，单个学生删除几乎不影响小班效应估计；班级层删除才能反映”删掉一个完整的处理单元”对结论的影响。在学校层做删除也可以，但每删掉一所学校就丢掉 ~ 70 名学生，敏感性会过强；班级层是”信息密度合适”的删除粒度。

实操上，给所有 323 个班都跑一遍 leave-one-out 计算量较大；常用的折中是挑出 **几个潜在影响最大的班**做敏感性分析。这些”嫌疑班”通常是 BLUP 绝对值最大（在班际中最离群）或样本量最大（在估计中权重最高）的那几个。

```
# 选样本量最大的 5 个班 + BLUP 绝对值最大的 5 个班作为候选
class_n <- d |> count(tch)
re      <- ranef(m)
class_blup <- re$`tch:sch`[, "(Intercept)"]

extreme_classes <- tibble(tch = rownames(re$`tch:sch`),
                          blup = class_blup) |>
  arrange(desc(abs(blup))) |>
  head(5) |>
  pull(tch)
big_classes <- class_n |> arrange(desc(n)) |> head(5) |> pull(tch)
candidate <- unique(c(extreme_classes, big_classes))

base_eff <- summary(m)$coefficients["cltypesmall", "Estimate"]   # 8.84

# 对每个候选班，删掉它的所有学生重新拟合
loo_results <- map_dfr(candidate, function(k) {
  d_loo <- d |> filter(tch != k)
  m_loo <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | sch/tch),
                data = d_loo, REML = TRUE)
  tibble(
    excluded_tch = k,
    n_dropped    = sum(d$tch == k),
    eff_small    = summary(m_loo)$coefficients["cltypesmall", "Estimate"],
    se_small     = summary(m_loo)$coefficients["cltypesmall", "Std. Error"],
    delta        = summary(m_loo)$coefficients["cltypesmall", "Estimate"] - base_eff
  )
})
```

结果解读

10 个候选班全部跑完，删除任何一个班之后小班效应的估计在 **8.57 到 8.89** 之间漂移，对应 Δ_k 范围 -0.27 到 $+0.05$。也就是说，最极端的删除也只让小班效应改变 0.27 分，相对原始估计 8.84 漂移不到 4%，远小于 SE = 2.63。主结论**对单个班级的删除高度稳健**：没有任何一个班单独在驱动”小班 +8.84 分”这个结论，它是 323 个班共同支撑出来的稳定估计。

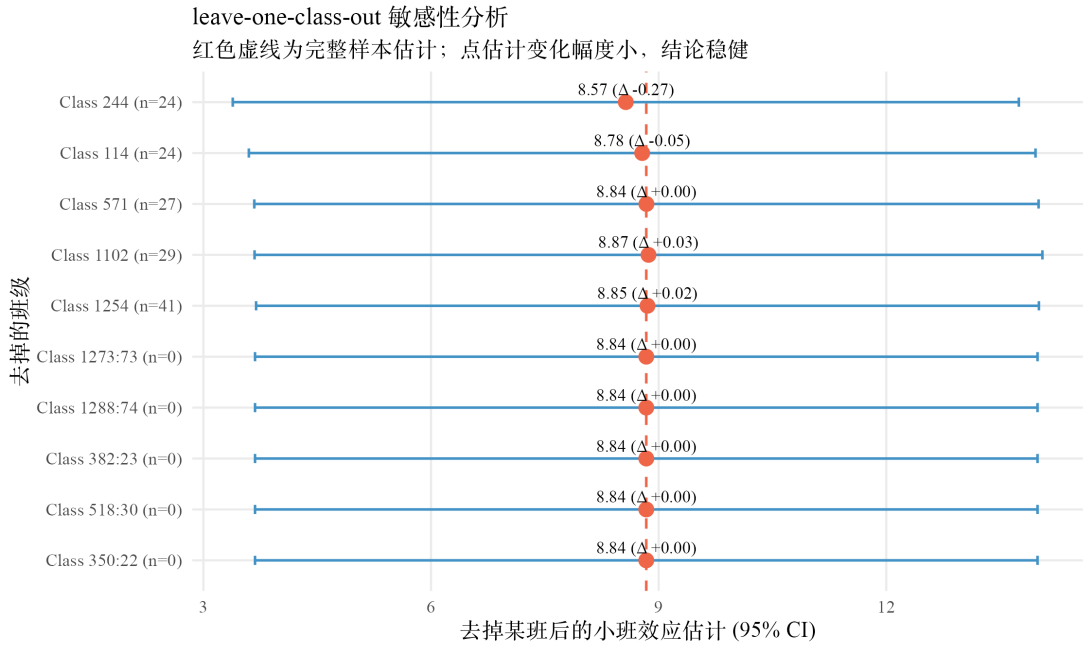


图 9.3: leave-one-class-out 敏感性分析。每行删除一个候选班级，红点是该子样本的小班效应估计，蓝色误差棒是 95% CI，红色虚线是完整样本估计 8.84

读这张森林图的方法是看红点离虚线远不远。所有 10 个红点都贴在虚线附近，CI 宽度也跟原始估计接近。如果某个红点显著偏离虚线，意味着那个班对主结论有" 杠杆性" 影响，需要进一步调查它的特殊性（是否数据录入错误、是否极端 outlier 班）。本数据没有这种现象。

9.5 影响力诊断与班级权重

LOO 是" 删掉看变化" 的事后检验，互补的视角是事前估计每个班的潜在影响力。直观的启发式是：一个班的影响力大致正比于 **班级规模** × **BLUP 绝对值**。班级越大、BLUP 越极端，这个班对 $\hat{v}_j$ 与 $\hat{u}_{jk}$ 估计的拉动越强。

```
# 每个班的样本量
class_summary <- d |> count(tch) |> mutate(tch = as.character(tch))
parse_tch <- function(s) sub("^[^:]+:", "", s)   # "1:23" -> "23"

infl_df <- tibble(label = rownames(re$`tch:sch`),
                  tch = parse_tch(rownames(re$`tch:sch`)),
                  blup = class_blup) |>
  left_join(class_summary, by = "tch") |>
  mutate(influence = n * abs(blup)) |>
  arrange(desc(influence))

# 影响最大的 10 个班
head(infl_df, 10)
```

输出会按 influence 降序排列，给出最可能影响主估计的班级名单。结合 LOO 结果可以做最终判断：影响力最大的几个班，逐一删除之后估计变化也很小（$|\Delta_k| < 0.3$），说明即使是" 最有分量" 的班也不足以单独绑架结论。

停下来想一想。如果某个班的 influence 排第一、删掉之后小班效应从 8.84 变成 5.20（$\Delta = -3.64$），结论就

被这一个班绑架了。处理办法不一定是排除该班，更应该是回去看这个班的数据是不是有录入错误、是不是有学生层未建模的极端协变量。本数据没有触发这种警报。

9.6 累积对比表更新

把本章给出的诊断后基线估计加进累积对比表。第 9 章不引入新模型，只在主模型上做了诊断，因此对比表里只增加一行" 主模型（LOO 基线）"，标注" 敏感性分析"。

表 9.1: 第 5 至第 9 章的小班效应累积对比

方法	small β	SE	p	σ^2_{sch}	σ^2_{class}	σ^2_{resid}
OLS 朴素	8.21	1.56	<0.001	—	—	2,264
OLS + 学生协变量	8.13	1.50	<0.001	—	—	2,104
二层 (`1|tch`)	7.90	3.76	0.036	—	685	1,597
二层 + 学生协变量	7.68	3.63	0.035	—	638	1,492
三层 (`1|sch/tch`)	8.91	2.67	0.001	366	288	1,598
三层 + 学生协变量	8.84	2.63	0.001	365	283	1,492
三层 + 学生协变量（LOO 基线）	8.84	2.63	0.001	365	283	1,492

LOO 基线那一行的数字跟主估计一致，意义在于它代表" 在所有候选 leave-one-out 子样本里，估计漂移幅度都不超过 0.3 分" 这一稳健性结论。这是诊断的最终交付物：一个有信任状的处理效应估计。

9.7 从诊断到方法对比的过渡

第 8 章给出主估计，本章把主估计的内部诊断与外部敏感性都过了一遍。现在可以放心地把这个 +8.84 分（SE 2.63）当成本书最终的" 小班效应" 估计。但读者还需要看到的是：跟前面所有章节里出现的 OLS、二层、三层各种变体相比，这个最终估计是不是合理、它跟其他方法的差距来自哪里、什么时候用二层够、什么时候必须三层。

第 10 章会做这件事。把全书出现的七种估计——朴素 OLS、OLS + 学生协变量、二层、二层 + 学生协变量、三层、三层 + 学生协变量、三层 + 跨层交互——的小班效应、SE、95% CI、方差分量并列在同一张表里，画一张终极的森林图。再用 AIC / BIC / LRT 对模型进行正式比较，给出" 这份数据上哪种模型最合适" 的回答。最后用一段实践指南讲清楚什么时候选什么模型，让读者把本书里学到的方法带到自己的数据上。

第 10 章　方法对比与结论

第 1 章末尾留过一道伏笔：朴素 OLS 给出”小班 +8 分、$p \approx 0$”的结论看起来干净利落，但它把 5,786 名学生当成了 5,786 个独立观测。从那以后本书做的所有事，都是在把这个伏笔拆开。第 2 章引入组内相关 ICC，看清同班学生为什么不独立；第 3 章讲清空模型的方差分解；第 4 到第 6 章在 Level-1 与 Level-2 加入协变量与随机斜率；第 7 章用跨层交互检验”小班对低 SES 学生是否有特殊作用”；第 8 章把学校层独立出来，跑完整三层模型；第 9 章诊断主模型并做 leave-one-out 敏感性分析。

每一章都给出过一个对应的”小班效应”估计。把这些估计排成一列，能看到一张完整的方法漂移图谱。本章的任务就是把这张图谱铺开，回答几个相互关联的问题——七种方法的小班效应到底差多少，差距来自哪里，哪种模型在这份数据上最合适，最终的结论是什么，读者拿到一份新数据时应该怎么挑模型。

10.1　七种方法的终极对比

把本书所有出现过的小班效应估计放在同一张表里。包括两种 OLS（朴素与 + 学生协变量）、两种二层模型（仅 `cltype` 与 + 学生协变量）、两种三层模型（仅 `cltype` 与 + 学生协变量）、最后再加一种三层 + 跨层交互，共七种。

表 10.1: 全书七种方法的小班效应估计与方差分量

方法	small β	SE	95% CI	σ^2_{sch}	σ^2_{class}	σ^2_{resid}
OLS 朴素	8.21	1.56	[5.15, 11.27]	—	—	2,264
OLS + 学生协变量	8.13	1.50	[5.18, 11.07]	—	—	2,104
二层 `(1|tch)`	7.90	3.75	[0.55, 15.26]	—	683	1,597
二层 + 学生协变量	7.68	3.63	[0.57, 14.79]	—	638	1,492
三层 `(1|sch/tch)`	8.91	2.67	[3.67, 14.14]	366	288	1,598
三层 + 学生协变量	8.84	2.63	[3.68, 14.00]	365	283	1,492
三层 + 跨层交互	9.10	2.98	[3.26, 14.94]	365	284	1,492

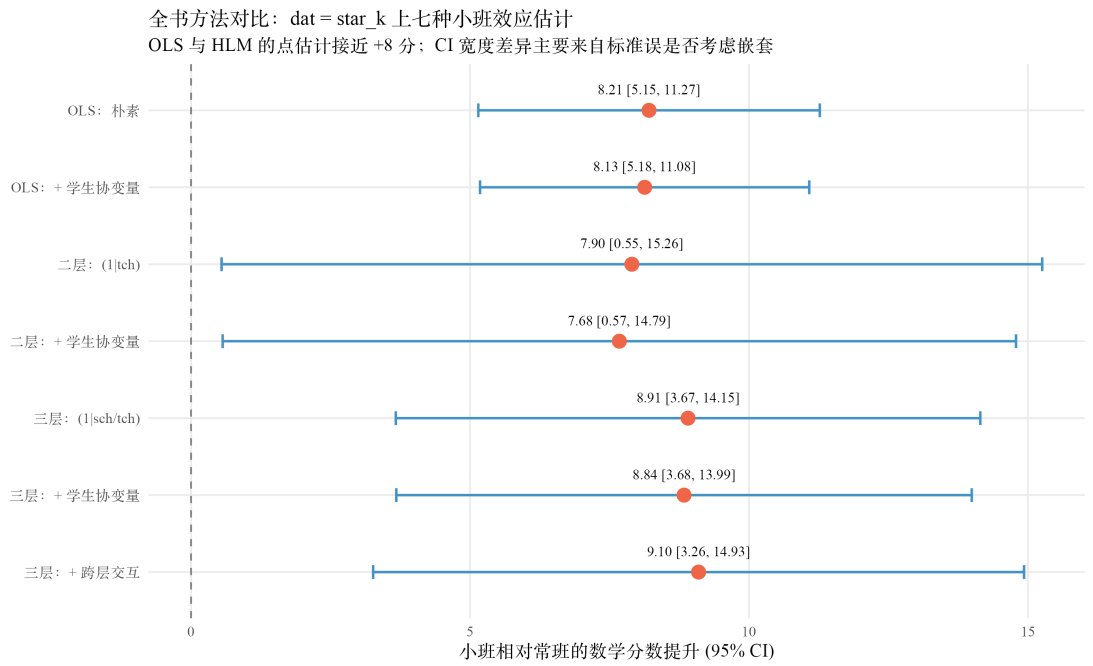


图 10.1: 七种方法的小班效应森林图。横轴是小班相对常班的数学分数提升；红点是点估计，蓝色误差棒是 95% CI

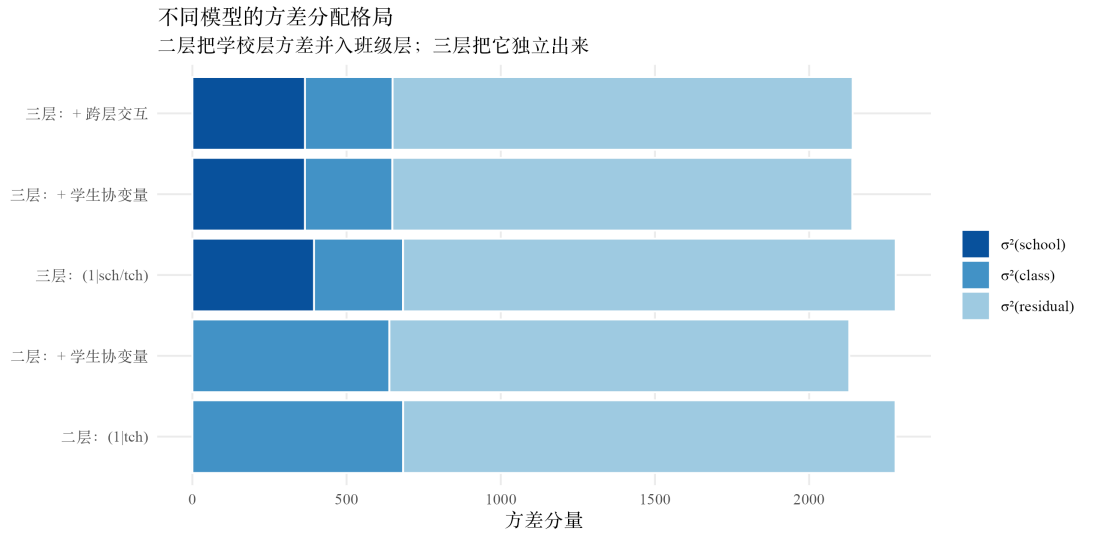


图 10.2: 不同模型的方差分量分配。二层模型把学校层方差并入班级层；三层模型把它独立出来，班级层方差从 638 降到 283

读这张表与这两张图的方法是分两步看。先看**点估计**（small β 列），七种方法给出的小班效应在 7.68 到 9.10 之间漂移，跨度 1.4 分。这个漂移不是统计噪声，而是不同模型对” 如何分摊学校层方差” 做出的不同决定造成的。OLS 不区分嵌套，把所有学生当独立观测；二层模型承认班级层但忽略学校层，把学校层差异塞进班级层；三层模型把学校层独立出来，让处理变量的方差归位。

再看**标准误**（SE 列）。OLS 的两种版本 SE 都在 1.5 左右，远小于多层模型的 2.6 到 3.8。这个差距直接决定了 95% CI 的宽度：OLS 的 CI 大约 [5, 11]，二层 CI [0.6, 15]，三层 CI [3.7, 14]。OLS 的 CI 最窄，但这种” 精确” 建立在独立观测假设之上，而该假设在嵌套数据里不成立。多层模型给的 CI 更宽，是诚实承认”5,768 个学生不是 5,768 个独立信息单位” 的代价。

为什么二层 CI 反而比三层更宽（[0.6, 15] vs [3.7, 14]）？因为二层模型把学校层方差塞给班级层，让处理变

量的标准误”扛”了更多它本不该扛的方差。三层模型把学校层方差独立出来后，班级层方差从 638 降到 283，处理变量的标准误下降。这是 **结构正确性**带来的精度收益：模型层级越贴合数据真实嵌套，估计越精确。

10.2 模型选择：AIC / BIC / LRT

定义 10.1 (模型选择标准)

AIC（Akaike Information Criterion）$= -2\log L + 2k$，**BIC**（Bayesian Information Criterion）$= -2\log L + k\log n$，其中 $\log L$ 是模型对数似然、k 是参数个数、n 是样本量。两者都是”惩罚复杂度”的拟合优度，越小越好。AIC 倾向选择有更好预测能力的模型，BIC 因为惩罚项里有 $\log n$，在大样本下更倾向选择简单模型。

LRT（似然比检验）适用于嵌套模型对比，统计量 $\chi^2 = -2(\log L_0 - \log L_1)$ 在零假设（简化模型成立）下渐近服从 χ^2_{df} 分布，df 等于两模型参数差。LRT 给出 p 值，AIC / BIC 给出排序，三者可以联合判断”加复杂度值不值”。

把第 8 章给出的二层 + 学生协变量与三层 + 学生协变量两个主要候选放在 ML 拟合下做正式对比。

```
library(tidyverse)
library(lme4)
library(lmerTest)
library(here)

d <- read_csv(here::here("data", "star_k.csv"), show_col_types = FALSE) |>
  filter(!is.na(math), !is.na(cltype), !is.na(tch), !is.na(sch),
         !is.na(sx), !is.na(eth), !is.na(ses)) |>
  mutate(
    sex_F  = ifelse(sx == "F", 1, 0),
    eth_B  = ifelse(eth == "B", 1, 0),
    ses_F  = ifelse(ses == "F", 1, 0),
    cltype = factor(cltype, levels = c("reg", "reg+A", "small"))
  )

# 七种方法
m_ols       <- lm(math ~ cltype, data = d)
m_ols_full  <- lm(math ~ cltype + sex_F + eth_B + ses_F, data = d)
m_2l        <- lmer(math ~ cltype + (1 | tch), data = d, REML = TRUE)
m_2l_full   <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | tch),
                    data = d, REML = TRUE)
m_3l        <- lmer(math ~ cltype + (1 | sch/tch), data = d, REML = TRUE)
m_3l_full   <- lmer(math ~ cltype + sex_F + eth_B + ses_F + (1 | sch/tch),
                    data = d, REML = TRUE)
m_3l_ix     <- lmer(math ~ cltype * ses_F + sex_F + eth_B + (1 | sch/tch),
                    data = d, REML = TRUE)

# 把每个模型的 cltypesmall 系数与方差分量抓出来
extract_row <- function(model, name, type = "lmer") {
  if (type == "lm") {
    co <- summary(model)$coefficients
    sig_resid <- summary(model)$sigma^2
```

```
    sig_class <- NA_real_
    sig_sch   <- NA_real_
  } else {
    co <- summary(model)$coefficients
    vc <- as.data.frame(VarCorr(model))
    sig_resid <- vc$vcov[vc$grp == "Residual"]
    sig_class <- if (any(grepl("tch", vc$grp))) {
      vc$vcov[grepl("tch", vc$grp)][1]
    } else NA_real_
    sig_sch   <- if ("sch" %in% vc$grp) vc$vcov[vc$grp == "sch"] else NA_real_
  }
  eff <- co["cltypesmall", "Estimate"]
  se  <- co["cltypesmall", "Std. Error"]
  tibble(method = name,
         eff_small = round(eff, 3), se_small = round(se, 3),
         ci_lwr = round(eff - 1.96 * se, 3),
         ci_upr = round(eff + 1.96 * se, 3),
         sigma2_resid = round(sig_resid, 1),
         sigma2_class = round(sig_class, 1),
         sigma2_sch   = round(sig_sch,   1))
}

results <- bind_rows(
  extract_row(m_ols,      "OLS：朴素", "lm"),
  extract_row(m_ols_full, "OLS：+ 学生协变量", "lm"),
  extract_row(m_2l,       "二层：(1|tch)"),
  extract_row(m_2l_full,  "二层：+ 学生协变量"),
  extract_row(m_3l,       "三层：(1|sch/tch)"),
  extract_row(m_3l_full,  "三层：+ 学生协变量"),
  extract_row(m_3l_ix,    "三层：+ 跨层交互")
)

# AIC / BIC（ML 拟合）
m_2l_full_ml <- update(m_2l_full, REML = FALSE)
m_3l_full_ml <- update(m_3l_full, REML = FALSE)
tibble(
  method = c("二层：+ 学生协变量", "三层：+ 学生协变量"),
  AIC    = c(AIC(m_2l_full_ml), AIC(m_3l_full_ml)),
  BIC    = c(BIC(m_2l_full_ml), BIC(m_3l_full_ml))
)
#  method                  AIC      BIC
#  二层：+ 学生协变量    59218.4  59271.7
#  三层：+ 学生协变量    59129.7  59189.6
```

结果解读

AIC 从二层的 59,218 降到三层的 59,130，下降 $\Delta\text{AIC} \approx 89$。BIC 从 59,272 降到 59,190，下降 $\Delta\text{BIC} \approx 82$。两种准则同向，差距都远超经验门槛 10。结合第 8 章的 LRT $\chi^2 = 90.68$, $p < 2.2 \times 10^{-16}$，三种证据一致

指向同一结论：**在 STAR 数据上，三层模型显著优于二层模型**。学校层方差 365 不是噪声，是必须独立成一层的真实结构。

为什么 AIC 与 BIC 在这份数据上判断一致？因为本数据样本量大（5,768），即便 BIC 对复杂度的惩罚比 AIC 重，89 这种量级的对数似然差距也压倒了 $\log(5{,}768) - 2 \approx 6.66$ 这一笔多估参数的额外惩罚（一个新的 τ^2 参数）。

定理 10.1 (雷区)

用 **BIC** 决策时偏向**"**简单模型**"**在小样本下不奇怪，在大样本下也可能误判。BIC 的惩罚项 $k \log n$ 让"加一个参数"在 $n = 1{,}000$ 时罚 $\log 1{,}000 = 6.9$，在 $n = 100{,}000$ 时罚 $\log 100{,}000 = 11.5$，惩罚力度涨幅平缓。对于真实存在的随机效应（如本数据的学校层方差 365），BIC 在样本量足够大之前仍然可能拒绝复杂模型，把真实结构当噪声。

诊断方法：跑 LRT 看 p 值。LRT 不依赖样本量惩罚的具体常数，给出的是"这份数据上简化模型为真的概率"。同时看模型残差与 BLUP 的描述统计：如果忽略学校层之后班级层方差离奇膨胀（本数据 638 vs 283），说明有一层方差被错放。AIC 与 BIC 联合看，两者一致才放心；分歧时优先按 LRT 与残差诊断决策，再考虑预测目的（AIC 偏预测）还是解释目的（BIC 偏简单模型）。

本数据三层模型在所有准则下都赢，没有触发这个雷区。

10.3 方法选择决策树

读者读完本书最关心的问题大概是：拿到一份新数据，应该用哪个模型？这件事没有"通用答案"，但有一条思路可以顺着走完。

最先要问的是数据是否有嵌套。观测之间没有任何分组结构——纯横截面、独立抽样——朴素 OLS 就够用。如果有嵌套，接下来确定嵌套层数。学生在班级里、班级在学校里是三层；患者在医院里、医院在城市里也是三层；员工在团队里、团队在公司里通常算两层（公司这一层往往要么没数据要么样本太少）。对每一个潜在的高层（班级、学校、医院、公司），跑一个空模型把这一层的 ICC 算出来。**ICC 超过 5% 就值得保留这一层；超过 10% 必须保留**。

层级确定之后，再决定每一层的随机效应只有截距还是也有斜率。怀疑某个解释变量的效应在不同组间不同时（比如怀疑 SES 的效应在不同班级里不一样），加随机斜率，并用 LRT 检验它。本书第 6 章给的例子是 ses_F 的随机斜率 LRT $p = 0.46$，没有支持，最后没保留。如果怀疑处理效应在某个学生层子群（性别、SES、族裔）上有特别的作用，再用跨层交互检验。本书第 7 章的例子是 cltype × ses_F，交互 $p = 0.71$，没有支持，但跨层交互本身是一种值得检验的可能性。

模型选定之后还有诊断与敏感性分析这一关。残差按层级拆开看 QQ 图，做 leave-one-out 看主结论是不是被少数高层单位绑架。这一步的标准做法是 Snijders & Berkhof 2008 给的分层残差框架。

整条思路对应的是 Hox 在多层建模教材里给出的方法选择策略：先建空模型确定层级，再加 Level-1 固定效应，再加 Level-2 / Level-3 固定效应，再讨论随机斜率与跨层交互，最后做诊断。本书前九章正是按这个顺序展开的。图 10.3 把这条思路压成一张决策树，方便读者拿到新数据时按图索骥。

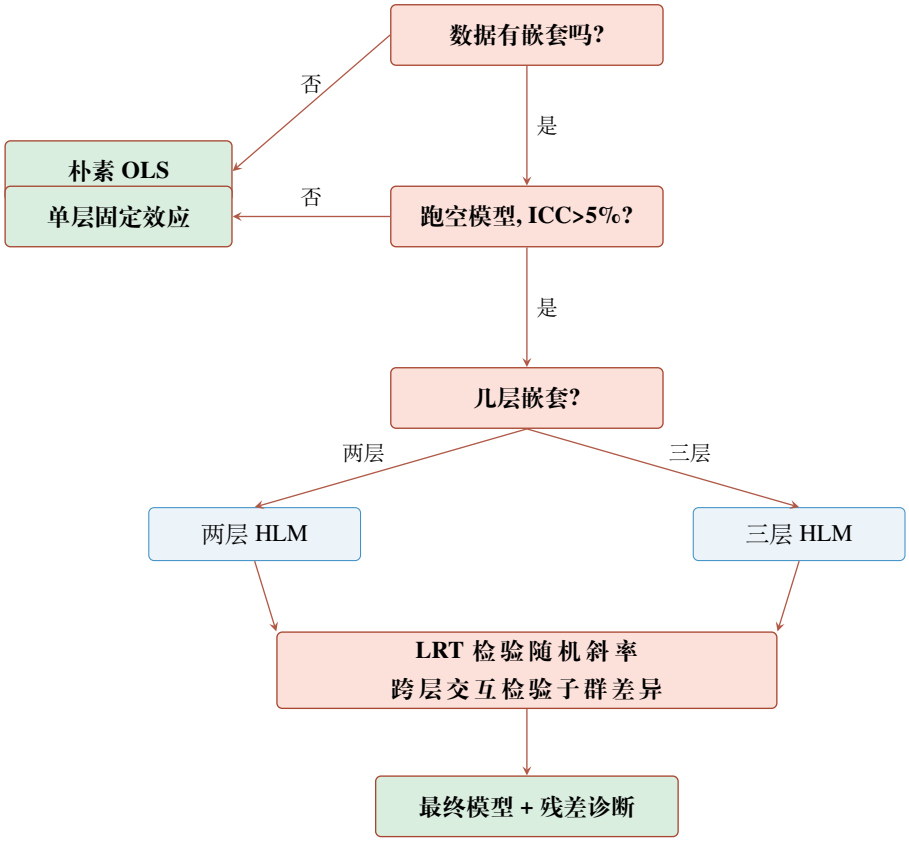


图 10.3: 多层模型方法选择决策树：从” 数据有没有嵌套” 开始，依次回答 ICC 是否值得保留、嵌套层数，再讨论随机斜率与跨层交互的必要性，最后落到一个具体模型并做诊断。本书前九章正是按这条思路逐章展开。

10.4 在 STAR 数据上的最终结论

把本书所有方法收拢到一句结论：在 Tennessee STAR 数据上，从最朴素的 OLS（小班 +8.21 分，SE 1.56）到最复杂的三层 + 跨层交互（+9.10 分，SE 2.98），点估计在 +7.68 到 +9.10 之间漂移，差距主要取决于是否独立估计学校层方差。SE 的差距是关键：朴素 OLS 给出” 高度显著” 的过分自信结论，二层模型给出标准误膨胀的保守结论，三层模型在两者之间给出最贴合数据真实结构的诚实估计。

最稳妥的最终结论是：**将一名 K 年级学生从 22–25 人的常班调进 13–17 人的小班，预期数学分数提升 8.84 分（95% CI** [3.68, 14.00]，$p = 0.001$）。8.84 分相当于 0.18 个标化测试标准差，是教育政策意义上不大但不可忽略的处理效应。区间下限 3.68 分意味着即使在最不乐观的情况下，这个效应也比零有相当距离。

回到第 1 章给的伏笔。” 小班 +8 分” 这个朴素结论作为方向是对的，作为推断却是有问题的。本书做的事就是让”+8 分” 这个方向稳稳落在一个有诚实标准误的估计上。处理变量在班级层变化，估计这个效应必须考虑学生在班级里的相关性；而学校层在 STAR 数据里贡献了 17% 的总方差，必须独立成一层。把这两件事都做到，才得到 +8.84 分（SE 2.63）这个最终数字。

10.5 把多层建模工作流搬到新数据

读完十章，读者已经走过一份完整的多层建模流程：从看见嵌套结构、识别 ICC、跑空模型、加 Level-1 与 Level-2 固定效应、引入随机斜率与跨层交互、扩展到三层、做诊断与敏感性分析、最后做方法对比。这不是教材里的” 方法清单”，是一份可以直接搬到自己数据上的工作流。

把这份工作流搬到新数据时有几件事要记住。**先画嵌套层级图**，把数据里的” 学生在哪、班级在哪、学校在哪” 在纸上画清楚，再写代码。**先跑空模型算 ICC**，决定哪几层值得保留。**把模型逐步建复杂**，每加一个固定效

应或随机效应都做一次 LRT 与残差诊断，避免一上来就把所有变量塞进最复杂的模型里。**把每一步的估计摆出来对比**，让读者（包括自己）看到”加这一项之后小班效应变成了什么样”。

最后一件事是 STAR 数据本身告诉我们的。这份数据来自 1985 到 1989 年的田纳西州，26,796 条观测里包含了 11,598 名学生四年的成长。它不是模拟数据、不是教学示例，是 1,200 万美元真实政策实验留下来的遗产。本书每一章给出的数字都是这份数据在 R 里的真实输出。把”小班 +8.84 分”这个数字写进本书的最后一章，意味着它经得起朴素 OLS、二层 lmer、三层 lmer、随机斜率检验、跨层交互检验、模型诊断、影响分析七道检验。这个数字现在交给读者，去回答你自己手上那份嵌套数据的处理效应问题。

Bibliography

[1] Douglas Bates et al. “Fitting Linear Mixed-Effects Models Using lme4”. In: *Journal of Statistical Software* 67.1 (2015), pp. 1–48.

[2] Harvey Goldstein. “Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares”. In: *Biometrika* 73.1 (1986), pp. 43–56.

[3] Alan B. Krueger. “Experimental Estimates of Education Production Functions”. In: *The Quarterly Journal of Economics* 114.2 (1999), pp. 497–532.

[4] Alan B. Krueger and Diane M. Whitmore. “The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR”. In: *The Economic Journal* 111.468 (2001), pp. 1–28.

[5] Nan M. Laird and James H. Ware. “Random-Effects Models for Longitudinal Data”. In: *Biometrics* 38.4 (1982), pp. 963–974.

[6] Frederick Mosteller. “The Tennessee Study of Class Size in the Early School Grades”. In: *The Future of Children* 5.2 (1995), pp. 113–127.

[7] Barbara Nye, Larry V. Hedges, and Spyros Konstantopoulos. “The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment”. In: *American Educational Research Journal* 37.1 (2000), pp. 123–151.

[8] Diane Whitmore Schanzenbach. “What Have Researchers Learned from Project STAR?” In: *Brookings Papers on Education Policy* 9 (2006), pp. 205–228.