

## 关于分层线性模型样本容量问题的研究

张璇 王嘉宇

2011-12-13 14:33:23 来源: 《统计与决策》(武汉)2010年15期第4~8页

**内容提要:** 文章运用Jackknife和Bootstrap的方法, 对参数估计的方差进行改进, 构造了合适的参数估计的置信区间。通过样本组数和组内个体数的变化, 利用数据模拟的方法进行研究, 表明参数估计的可靠性很大程度上依赖于组数; 对于固定效应参数, 组数取30就可以得到可靠的估计值。对于 $\sigma$ 和方差协方差成分T, 组数分别取50和70才能得到可靠的估计。

**关键词:** 分层线性模型 参数估计的覆盖率 Jackknife Bootstrap 数据模拟

**作者简介:** 张璇(1979-), 女, 湖南湘潭人, 中国人民大学统计学院博士研究生, 讲师, 研究方向: 统计模型及其计算、计量经济学(北京100084); 王嘉宇, 卡尔斯塔德大学国民经济与统计系, 乌普萨拉大学信息科学与统计系(瑞典65188)。

### 1 研究背景

很多社会研究都涉及分层数据结构, 例如, 经济学家探求在多个国家中经济政策是如何影响居民的消费行为, 研究采集的观测数据不仅包括以国家为层次的经济指标, 还包括以家庭为单位的信息, 因此整个观测的数据结构是分层的。此时, 同属一个层次的个体之间的相关性会大于来自不同层次的个体之间的相关性, 整个观测样本就不再具有独立同分布性质, 如果继续使用经典的线性回归模型, 就会得到有偏的参数估计和错误的统计推断结果。近年来, 随着分层线性模型统计理论的发展, 一套完整的应用于分层结构数据的统计推断方法已经建立起来, 并且能得到有效的参数估计。

分层线性模型(hierarchical linear models)的称谓最早由Lindley和Smith(1972) [1]提出。这个模型在不同的研究领域有不同的称呼, 在社会学研究中, 它经常被称为多层线性模型(multilevel linear model); 在生物统计研究中常用的名字是混合效应模型(mixed-effects models)和随机效应模型(random-effects models); 计量经济学文献称之为随机系数回归模型(random-coefficient regression models)等。

模型的一般形式为:



目前比较常用的分层线性模型的估计方法有极大似然法(ML), 约束极大似然法(REML) (Littell, 1996) [2]和完全贝叶斯法(Full Bayes)。理论研究表明, 大样本情况下ML和REML得到的参数估计都是一致最优估计量, 但是在样本较小的情况下, REML在估计方差协方差成分和T时, 考虑了固定效应系数 $\gamma$ 的不确定性, 因此REML比ML得到了更可靠的

和T的估计值。Full Bayes考虑了所有辅助参数的不确定性，因此理论上得到的参数估计较ML和REML更可靠，但是此估计方法需要运用MCMC算法，很多情况下，并非能得到一个收敛的Markov链，当分层线性模型形式较复杂时（如待定参数向量增多、层数增加等），Full Bayes方法相当复杂。因此很多统计软件采用ML和REML估计分层线性模型，本研究采用REML方法估计。

由于ML和REML方法估计的前提假设都是样本量要足够大，因此样本量较小时，这两种估计都是有偏的，由此得到的参数置信区间和假设检验都是不可信的。因此，分层线性模型样本量问题的研究是一个重大课题，近十几年来，很多的学者都致力于这方面的研究。Bryk和Raudenbush(1992) [3]很早就指出，可以凭借OLS回归的经验法则：增加一个解释变量至少需要增加10个观测样本，联系到分层线性模型的估计，增加一个层2的结果变量（层1模型中待定的随机参数）至少需要增加10个观测样本，这个准则只是OLS回归经验法则的平移，并没有清楚回答模型要得到较好估计所需的最小样本量问题。Kim(1990) [4]在研究分层线性模型斜率参数估计的时候，发现当组数较少而组内的观测值相对很大时，固定效应和随机效应的估计会产生很大的偏差，但是Kim的研究在相同的条件下，只进行了50次模拟计算，因此结果不太可靠。Mok(1995) [5]的研究也得出了类似Kim的结论，并且他进一步指出参数估计的偏差和效率更大程度取决于样本中的组数大小。Busing(1993) [6]指出当组数超过300时，随机效应的估计才是无偏的。Kreft(1996) [7]运用模拟技术探讨了分层线性模型的势，建议30/30准则，即30个组，每组30个观测值可以得到较可靠的估计。Hox(1998) [8]在Kreft(1996)之后更详尽阐述表明，50组，每组20个观测值可以使交互效应得到较可靠的假设检验；100组，每组10个观测值可以较可靠进行随机效应的假设检验。另外Bliese(1998) [9]明确指出组内相关系数(ICC)与样本量相互联系。Snijders和Bosker(1999) [10]阐明对于较大的ICC值，增加组数也不会得到一致的标准差的估计。Mass和Hox(2005) [11]针对不同的层1和层2样本量进行了模拟研究，结果表明层2的样本数小于50，会导致有偏的协方差成分估计。

纵观以上研究，其方法主要是运用数据模拟的方法，针对不同层1、层2样本量的组合，估计分层线性模型，比较固定效应、随机效应和方差协方差成分的估计值。在计算机发展日新月异的今天，进行模拟计算简单易行，应用广泛。但是对于此类问题，以往研究存在许多问题：①大部分研究考虑的是参数估计的准确性，只有很少的一部分重视估计的标准差的准确性。②许多研究在讨论估计参数假设检验或构造置信区间时，想当然的运用了大样本的近似理论，即参数估计近似服从正态分布，但是实际上，在样本量较小时，分层线性模型的参数估计值，尤其是方差协方差成分的估计值并非服从正态分布(Raudenbush(1984))。

## 2研究方法

本研究仍采用数据模拟的方法，用R语言编译计算程序。另外，研究采用简单的两层模型，每层模型各有一个解释变量：

在数据模拟过程中，研究分别改变组数，组内个体数和组间相关系数ICC，在样本变化的条件下，构造参数估计的置信区间，计算置信区间覆盖真实参数的覆盖率来考查估计值的分布情况。①组内个体数取值为5, 30, 50, ICC取值为1, 2, 3, 考察组数从5至100改变情况下，估计值的分布如何变化；②组数取值为30, 50, 100, ICC取值为1, 2, 3, 考察组内个体数从5至100改变情况下，估计值的分布如何变化。

在构造置信区间时，以往研究直接将方差用近似标准差代替，构造正态分布的置信区间，而本研究用另外一种方法计算参数估计的标准差。

R程序lme4程序包中的lmer命令可以提供分层线性模型的计算，它提供固定效应系数和方差协方差成分的估计值，以及固定效应系数估计的标准差，但是不提供方差协方差成分估计的标准差，lmer命令的编写者Doug Bates(2008)表示，第一，方差协方差成分估计的标准差的计算是非常复杂的，很多情况下，估计值已经在参数空间以外，即便是估计出来，也是无意义的，因此有些软件（如SAS）提供标准差也并不可靠。第二，当参数分布差不多是对称的情况下，提供参数估计的标准差才是有意义的，因为可以由此构造置信区间。

Harvey Goldstein(1999)提到, 样本不是很大的情况下, 如果随机扰动项不再服从正态分布, 分层线性模型的固定效应和随机效应的估计仍然是一致的, 但是其估计的标准差不能用来构建置信区间和进行显著性检验。一种替代的方法是运用密集计算中的Jackknife和Bootstrap计算估计的标准差, 由Jackknife和Bootstrap理论可知估计的标准差可以利用正态分布构建置信区间和进行显著性检验。

### 3研究步骤

#### 3.1模拟数据的生成

 F104U604. jpg

 F104U605. jpg

(4) 讨论覆盖真实值情况: 如果以上的置信区间确实覆盖了真实值, 则取值为1; 否则取0。

(5) 计算覆盖率: 重复第一步至第四步1000次, 统计覆盖次数, 计算覆盖率。

(6) 在不同样本量情况下, 考察覆盖率随样本量如何变化。固定 $n$ (=30, 50, 100),  $p$ (=1, 2, 3),  $N$ 从5到100变化; 固定 $N$ (=5, 30, 50),  $p$ (=1, 2, 3),  $n$ 从5到100变化。

#### 3.2.2运用参数Bootstrap方法

和Jackknife方法相比, 参数Bootstrap方法步骤(2)中有差异, 它将(2)分成以下几步。

 F104U606. jpg

 F104U607. jpg

构造置信区间和覆盖率的计算和Jackknife的过程一样。

### 4研究结果

#### 4.1固定效应和方差协方差成分的置信区间覆盖情况

由于不论是Jackknife还是Bootstrap方法, 运算量都是非常大的, 因此我们先考察在组成样本三个条件 $N$ ,  $n$ ,  $p$ 取不同值的组合下置信区间的覆盖情况。

##### 4.1.1Jackknife方法

表1的结果表明, 当 $N=5$ 时, 计算的覆盖率都离95%较远, 这时若 $n$ 的取值较大, 如 $n=100$ , 覆盖率会相对提高, 但是方差协方差成分中  F104U608. jpg 的覆盖率还是很低, 这说明分层线性模型的估计很大程度上依赖于 $N$ 的个数。如果仅仅考虑参数点估计, 当 $N$ 的取值很小, 即使 $n$ 取值很大,  F104U608. jpg 的估计值也不能收敛, 因此存在某个整体较大的样本量(较大的 $n$ 和较小的 $N$ ), 方差协方差成分的估计是相当不可靠的。另外从表1中, 还可以看出组内相关系数(ICC)的改变对置信区间的覆盖率没

有太大影响。

当N达到30, n达到30时, 固定效应和的估计是比较可靠的, 其覆盖率基本上达到93%, 只有三个值小于93%; 但是此时对于

而言, 大部分的覆盖率仍然偏低, 只有个别能达到93%以上。

表1N, n, p不同取值的组合下, Jackknife和Bootstrap方法计算参数95%置信区间覆盖率(%)



注: “J”表示Jackknife法得到的覆盖率; “B”表示Bootstrap法得到的覆盖率。

#### 4.1.2 Bootstrap方法

和Jackknife相比, Bootstrap方法得出的结论很相似, 但是对于N较小时, 可以清楚地看到ICC的值越大, 其95%的置信区间的覆盖率越小; 另外对于 $\sigma$ 的估计和Jackknife不一样, 当N=5时, 只要提高组内的样本量n, 例如当n=100时, 其95%的置信区间的覆盖率的均值为94.5%; 当n=30, N=5, 95%的置信区间的覆盖率的均值也达到93%, 这说明 $\sigma$ 的估计依赖的是整个样本容量, 若组数稍小时, 可以通过增加组内的个体数, 提高估计的可靠性。

从以上的分析可以看出, Jackknife和Bootstrap方法只是在样本组数较小时有差异, 而且其差异主要表现在 $\sigma$ 的估计上, 这种差异来源于这两种方法计算时再抽样的过程不同, 当然这种Parametric Bootstrap的方法在样本组数较小时优于Jackknife的方法, 但当样本数增加时区别不大, 而Parametric Bootstrap的计算量远远超过Jackknife的计算, 因此本研究下面的分析, 当N和N连续的变化时, 我们考虑Jackknife的计算方法。

#### 4.2N从5到100变化时参数估计置信区间的覆盖率

现在考虑n取30, 50, 100; ICC取1, 2, 3的不同组合, N从5到100变化。因为ICC的变化对覆盖率的影响不大, 因此以下的覆盖率均为ICC三种取值下的平均覆盖率。另外, 为了考察覆盖率的可接受性, 我们利用随机模拟的方法建立相应样本量下的覆盖率的置信区间(利用随机数均值覆盖率的置信区间), 构成了相应的置信带。

#### 4.2.1 固定效应

以为例, 其他三个固定效应的变化情况相似。从图1清楚地看到, 当N取值很小时(如N=5, 10), 固定效应的覆盖率是很低的; 当N达到35时, 三种情况下的覆盖率都进入了置信带。

#### 4.2.2 方差协方差

考虑 $\sigma$ 和的变化参照。① $\sigma$ 的估计情况; 当N很小时(如N=5, 10), 标准差 $\sigma$ 覆盖率远低于置信带的下界; 随着n的增加, 有更多的覆盖率进入置信带, 平均当N达到50时, 三种情况下的覆盖率完全进入置信带。②的估计情况: 的覆盖率在n的三种取值情况下, 都比固定效应 $\gamma$ 和标准差 $\sigma$ 相对应的覆盖率低, 因此N从5变化至100时, 只有较少的部分进入置信带, 对于n的三种取值, N达到70时, 覆盖率几乎完全进入置信带。对于较小的N, 随着n的增加, 参数覆盖率更靠近置信带。因此对于 $\tau$ 的估计, 可以通过n的增加得以改善。

#### 4.3n从5到100变化时参数估计置信区间的覆盖率

现在考虑N取5, 30, 50; ICC取1, 2, 3的不同组合, N从5到100变化。因为ICC的变化对覆盖率的影响不大, 因此以下的覆盖率为ICC三种取值下的平均覆盖率值。另外, 为了考察覆盖率的可接受性, 我们利用随机模拟的方法建立相应样本量下的覆盖率的置信区间(利用随机数均值覆盖率的置信区间), 构成了相应的置信带。

### 4.3.1 固定效应

以  为例, 其他三个固定效应的变化情况相似。图4(见下页)清楚地表明, N为5时, 不论n如何增加, 参数置信区间的覆盖率都远离置信带; 当N取30时, 只有两点在置信带外; 当N取50时, 整条覆盖率连线全部进入了置信带。



图1n的不同取值下  的95%的置信区间覆盖率

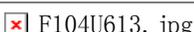


图2n的不同取值下  $\sigma$  的95%的置信区间覆盖率



图3n的不同取值下  的95%的置信区间覆盖率



图4N取5, 30, 50时,  的95%的置信区间覆盖率

### 4.3.2 方差协方差

考虑  $\sigma$  和 ,  的变化参照 。①  $\sigma$  的估计情况。图5显示出, N为5时, 不论n如何增加, 参数  $\sigma$  置信区间的覆盖率都远离置信带; 当N取30时, 大部分的覆盖率都进入了置信带; 当N取50时, 除一点外, 整条覆盖率连线全部进入了置信带。②  的估计情况。图6表明  的估计和固定效应参数与  $\sigma$  有很大差别, 当N取50时, 还有很多覆盖率在置信带以外, 这表明对于方差协方差成分的估计N为50是不够的。结合图3, 我们可以看到对于方差协方差成分的估计, N达到70或者更多才可靠。另外, 图6还显示了一个有趣的现象, 当N取30和50时, 较小的n(比如  $n < 20$ ) 的覆盖率稳定地进入了置信带, 而较大的n(例如:  $30 < n < 90$ ) 的覆盖率反而在置信带以外, 这表明这时较小的n的参数估计比较大的n更可靠。

## 5 研究结论

从以上的深入分析中, 我们可以看到分层线性模型估计的可靠性与样本中的组数, 组内个体样本数的取值有密切关系, 总结如下:

(1) 样本中的组数N。分层线性模型估计的可靠性很大程度上依赖于N的取值, 随着N的增加, 固定效应, 方差协方差成分的参数估计的95%的置信区间的真实覆盖率能够进入置信带。当N较小时(如N=5), 无论n如何增加, 所有参数估计都不可靠, 其覆盖率远远低于95%的置信带。对于固定效应参数  $\gamma$  的估计, N取30就可以得到可靠的参数估计。对于参数  $\sigma$  的估计, N取50, 才能使

其得到可靠的估计，方差协方差成分T的可靠估计需要N的取值达到70以上。

 F104U617. jpg

图5N取5, 30, 50时,  $\sigma$  的95%的置信区间覆盖率

 F104U618. jpg

图6N取5, 30, 50时,  F104U611. jpg 的95%的置信区间覆盖率

(2)组内个体样本数n。在N固定的情况下,随着n的增加,所有参数估计的可靠性都没有明显得到改善,即覆盖率连线没有明显的上升趋势。另外当N取值为30和70之间时,对于较小的n( $n < 20$ ),比较大的n的参数估计更可靠。

参考文献:

[1]Lindley, D.V., Smith, A.F.M. Bayes Estimates for Linear Model[J]. Journal of the Royal Statistical Society, Series B, 1972, (34).

[2]Littell, R., Milliken, G., Stroup, W., Wolfinger, R. SAS System for Mixed Models[M]. Cary, NC: SAS Institute Incorporated, 1996.

[3]Bryk, A.S., Raudenbush, S.W. Hierarchical Linear Models[M]. Newbury Park, CA: Sage, 1992.

[4]Kim, K.S. Multilevel Data Analysis: A Comparison of Analytical Alternatives[M]. Los Angeles: University of California, 1990.

[5]Mok, M. Sample Size Requirements for 2-level Designs in Educational Research[C]. Working Paper, 1995.

[6]Busing, F.M.T.A. Distribution Characteristics of Variance Estimates in Two-level Models[M]. Netherlands: Psychometric and Research Methodology, Leiden, 1993.

[7]Kreft, I.G.G. Are Multilevel Techniques Necessary? An Overview, Including Simulation Studies[C]. Working Paper, 1996.

[8]Hox, J. Multilevel Modeling: When and Why. In R. Mathar & M. Schader, Classification, Data Analysis, and Data Highways[M]. Berlin, Germany: Springer-Verlag, 1998.

[9]Bliese, P.D. Group Size, ICC Values, and Group-Level Correlations: A Simulation[J]. Organizational Research Methods, 1998, 1(4).

[10]Snijders, T.A.B., Bosker, R.J. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling[M]. Thousand Oaks, CA: Sage, 1999.

[11]Mass, C.J.M., Hox, J.J. Sufficient Sample Sizes for Multilevel Medeling[J]. Methodology, 2005, 1(3). ^

文档附件：

---

隐藏评论

用户昵称： (您填写的昵称将出现在评论列表中)  匿名

请遵守守法并注意语言文明。发言最多为2000字符（每个汉字相当于两个字符）

5654

发表

中国社会科学网电话：010-84177688；84177878；84177879      Email: [skw01@cass.org.cn](mailto:skw01@cass.org.cn)

投稿邮箱：[skw01@cass.org.cn](mailto:skw01@cass.org.cn)      网友之声信箱：[skw02@cass.org.cn](mailto:skw02@cass.org.cn)      地址：北京市朝阳区望京中环南路1号>      邮编：100102

版权所有：中国社会科学院 [版权声明](#)      京ICP备05072735号