文章编号: 1000-5862(2012)05-0461-05

贝叶斯理论框架下的 2 种纵向缺失数据处理方法的比较
—— 以潜在变量增长曲线模型为例

杨林山[1], 曹亦薇[2]

(1.深圳市海云天教育测评公司, 广东 深圳 518067; 2.深圳大学师范学院, 广东 深圳 518060)

摘要: 在贝叶斯估计框架下, 通过模拟研究比较完全贝叶斯和部分贝叶斯方法对参数估计的影响. 研究结果表明: 随着缺失比例的增加, 2 种方法得到的均方误差(RMSE)都会增大; 完全贝叶斯方法和部分贝叶斯方法在缺失比例较小时几乎相同, 只在缺失比例为 0.5 时, 前者明显优于后者.

关键词: 缺失数据; 完全贝叶斯; 部分贝叶斯; 纵向模型; WinBUGS
中图分类号: B 841.2　　文献标志码: A

0　引言

纵向研究是心理学研究的基本方法之一, 它也叫追踪研究, 指在一个相对长的时间段里对同一个或同一批个体进行重复测试研究. 与横断研究相比, 纵向研究的最大优势在于可以分离时间点上个体的增长趋势、个体间差异以及变量之间的因果关系[1]. 但在纵向研究中经常遇到数据缺失的问题: 例如被试中途流失而造成前后观测数据样本量不一致, 或者被试由于种种原因在不同的测试中回答问题不完全等. 早在 1987 年, R.J.A.Little 等[2]将数据缺失机制分为 3 类: 完全随机缺失、随机缺失和非随机缺失. 所谓完全随机缺失, 是指数据不论是否缺失, 其缺失概率独立于观测数据或缺失数据本身; 随机缺失则指数据缺失概率只依赖于观测数据而与缺失数据无关; 而非随机缺失强调数据缺失概率同时依赖于观察与缺失数据[3].

数据缺失即是测试信息的缺失, 直接的影响就是给纵向分析结果, 特别是给模型参数估计带来偏差. 因此在实际分析中往往要采用一些方法来处理缺失数据. J. W. Graham 通俗地将这些方法分为传统方法与现代方法 2 大类[4].

传统方法是指简单的删除处理, 如列举删除和配对删除等[4]. 列举删除也称为个案删除, 即删除在分析变量中有缺失反应的个体; 而配对删除是指在统计处理时只删掉在需要分析变量上缺失反应的个体, 如在相关分析时, 遇到缺失数据时采用这种处理方法. J. W. Graham 强调传统方法并不是过时、失效, 但是在数据缺失率过大时, 会造成统计检验力的下降[4].

现代方法中常见的是极大似然法、多重借补法[4]. 它们共同的特征是依存所选用的统计模型来估计模型参数与缺失数据.

例如 R.J.A.Little 在数据随机缺失的情况下, 发现可以用极大似然估计多层线性模型中的参数[5], Hox 用模拟研究验证了在非随机缺失的情况下, 多层线性模型也可以得到参数无偏估计的结果[6]; D.B.Rubin 发现在特殊缺失数据模型(包括潜变量曲线增长模型)的定义下, 用极大似然估计方法也可以得到缺失数据的估计值[7].

多重借补法是指在模型参数估计前产生多个借补值来替代缺失值, 这些数据反映了缺失值的不确定性, 同时也产生多个完整数据集. 在此基础上用分析完整数据集的统计方法对这些数据集分别进行统计分析, 然后对所得结果进行综合推断, 最终得到所需变量参数的估计[2, 8]. 如 D. A. Newman 为了研究组织管理中个体的变化, 比较多重借补、回归

收稿日期: 2012-07-30
作者简介: 杨林山(1983-), 男, 江西彭泽人, 硕士, 主要从事教育统计与测量方面的研究.

借补等缺失数据处理方法对参数估计误差的影响后, 发现多重借补方法优于回归借补[9].

20 世纪 60 年代 W. Edwards 等[10]把贝叶斯理论引入心理学研究中, 有别于样本理论的贝叶斯方法为研究者们带来了新的视野.

在贝叶斯估计的框架下, 对于纵向研究中缺失数据的处理, 研究者逐渐青睐使用完全贝叶斯(Fully Bayesian, FB)方法[2, 8, 11-13]. 完全贝叶斯方法的基本思想是把缺失数据视为新增未知参数, 通过模拟全体变量与缺失值的联合后验分布来估计模型参数与缺失值[2], 因此被 G. Carrigan 等[11]称为扩张的多重借补. Li Jinhui[12]于 2006 年研究发现: 完全贝叶斯方法与极大似然相比, 在估计含有缺失数据的纵向模型时, 具有较高的功效和一致性. Zhang Zhiyong 等[13]成功将此方法用于全美青年纵向调查中的学业数据.

在频率统计学派里, 许多研究者曾通过 EM 算法获得基于观测数据的极大似然估计, 这种方法一般忽略缺失数据[8]. 设想在贝叶斯统计理论的框架下, 是否也可以忽略缺失数据, 只在观测样本似然的基础上结合参数的先验信息以求得参数的后验分布? 由于这种处理方法只利用观测数据估计模型参数而不估计缺失数据, 本文把这种方法拟称为部分贝叶斯(Partially Bayesian, PB)方法.

目前对于处理纵向缺失数据的贝叶斯方法的讨论, 多集中于完全贝叶斯方法, 很少关注部分贝叶斯方法的适用性, 更缺少这 2 种方法应用于纵向模型后的结果比较. 本研究将选用潜变量增长曲线模型, 模拟不同缺失比例的数据集, 分别应用完全贝叶斯方法和部分贝叶斯方法对模型参数进行估计并比较其精确性.

1 贝叶斯方法

1.1 贝叶斯估计简介

在贝叶斯估计中, 全部的未知参数(包括模型参数 θ)均为随机变量, 它们的不确定性用概率分布来量化. 根据贝叶斯定理公式, 指定参数 θ 的一个先验分布 $p(\theta)$ 以及它的似然函数 $L(Y\mid\theta)$, 则可得到 θ 的后验分布

$$p(\theta\mid Y)=p(\theta)L(Y\mid\theta)/p(Y), \tag{1}$$

其中 $p(Y)$ 为全概率分布. 根据后验分布 $p(\theta\mid Y)$, 即可求得参数 θ 的估计统计量, 如后验期望(均值)、中数或众数或方差等. 例如后验期望的计算公式为

$$\bar{\theta}=\int\theta p(\theta\mid Y)\mathrm{d}\theta. \tag{2}$$

同样, 与此相关的方差也可以得到

$$Var(\theta)=\int(\theta-\bar{\theta})p(\theta\mid Y)(\theta-\bar{\theta})^t\mathrm{d}\theta. \tag{3}$$

在贝叶斯统计中, 可靠区间的应用目的和频率学派的目的是相同的. 形式上一个对参数 θ $100\times(1-\alpha)\%$ 的置信区间(L, U)可以这样得到

$$1-\alpha\leqslant\int_L^U p(\theta\mid y)\mathrm{d}\theta. \tag{4}$$

这里 L 和 U 分别是其下界和上界.

1.2 MCMC 算法及 Gibbs sampling 法

在具体的实践中, 参数 θ 的后验均值需要经过大量的高维积分才能得到. 随着MCMC算法(Markov chain Monte Carlo Algorithm)的产生, 为高维积分提供了有效的工具. 在MCMC 算法中的Gibbs sampling应用最为广泛[14]. 下面以2个参数为例对Gibbs sampling方法作一简单说明.

在已知数据 Y 的条件下, 2个随机变量参数ξ、η的后验联合分布为 $P(\xi,\eta\mid Y)$, 它们各自的完全条件分布为 $P(\xi\mid\eta,Y)$, $P(\eta\mid\xi,Y)$. 任意设 $[\xi^{(0)},\eta^{(0)}]$ 为初始值, 代入各自的条件分布 $P(\xi\mid\eta^{(0)},Y)$ 和 $P(\eta\mid\xi^{[0]},Y)$ 后模拟得到一组新的值: $[\xi^{(1)},\eta^{(1)}]$; 再将这组值分别代入 $P(\xi\mid\eta^{(1)},Y)$ 和 $P(\eta\mid\xi^{[1]},Y)$, 模拟得到 $[\xi^{(2)},\eta^{(2)}]$, …, 这样重复迭代 t 次($t\to\infty$)后, 便得到一系列的 $[\xi^{(t)},\eta^{(t)}]$(t =1, 2, …, T). 在此过程中, 观察第 m 次迭代后的结果, 若可以判断 $[\xi^{(m)},\eta^{(m)}]$ 的边际分布为平稳分布 $P(\xi,\eta\mid Y)$ 时, 就称数列收敛了. 但是在 m 次迭代前的各个状态的边际分布还不能认为其是 $P(\xi,\eta\mid Y)$. 因此删除最初的 m 次迭代值, 保留($T-m$)次的迭代值 $[\xi^{(m+1)},\eta^{(m+1)}]\cdots\ [\xi^{(T)},\eta^{(T)}]$ 来进行估计. ξ或η的后验分布可通过计算遍历平均得到

$$P(\xi\mid Y)=\frac{1}{T-m}\sum_{t=m+1}^{T}p(\xi\mid\eta^t,Y),$$

$$P(\eta\mid Y)=\frac{1}{T-m}\sum_{t=m+1}^{T}p(\eta\mid\xi^t,Y). \tag{5}$$

应用 Gibbs 抽样有 2 个关键: (1)如何获得条件后验分布, (2)确定生成 Markov 链的收敛点[13]. 如果采用共轭先验分布, 条件后验分布就很容易获得. 此外, 如果使用新的软件(如 WinBUGS[15])去获得其收敛, 研究者不需要明确指定其条件后验分布. 在实证中, 经常用到一种叫做“眼球”的方法. 即通过

视觉来观察生成序列的点轨迹来判断其收敛的情况. 通常，如果在点轨迹中没有变动的点或倾向，就能接受生产序列的收敛，如图 1 所示.

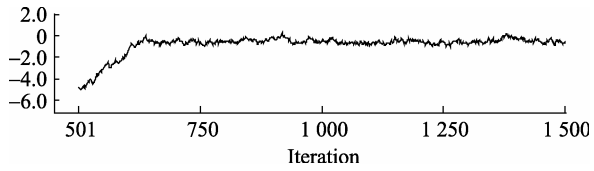

图 1　待估参数的后验分布轨迹图

1.3　完全贝叶斯及部分贝叶斯方法简介

完全贝叶斯方法是通过 MCMC 模拟得到缺失数据下的联合后验分布 $p(\theta, Y_{mis} \mid Y_{obs}, M)$，基于这个后验分布再同时估计参数和缺失数据.

设 Y_{obs} 表示观测数据，Y_{mis} 表示缺失数据，$Y=(Y_{obs}, Y_{mis})$ 表示完全数据，M 表示缺失机制[5]. 根据贝叶斯定理，可以得到联合后验分布

$$p(\theta, Y_{mis} \mid Y_{obs}, M) \propto p(M \mid \theta, Y_{mis}, Y_{obs}),$$
$$p(Y_{obs} \mid \theta, Y_{mis}) p(Y_{mis} \mid \theta) p(\theta), \tag{6}$$

这里 M 通过公式 $p(M \mid \theta, Y_{mis}, Y_{obs})$ 模型提供信息，如果数据缺失模式为完全随机缺失，则 M 独立与 θ、Y_{obs} 和 Y_{mis}，那么公式可以化简为

$$p(\theta, Y_{mis} \mid Y_{obs}, M) \propto p(Y_{obs} \mid \theta, Y_{mis}) p(Y_{mis} \mid \theta) p(\theta), \tag{7}$$

获得联合后验分布 $p(\theta, Y_{mis} \mid Y_{obs}, M)$ 之后，即可继续求得模型参数 θ 或者缺失值 Y_{mis} 的边缘分布 $p(\theta \mid Y_{obs}, M)$ 或 $p(Y_{mis} \mid Y_{obs}, M)$. 而在部分贝叶斯方法中，我们需要忽略 Y_{mis}，这时只有观测数据和先验信息发挥作用.

$$p(\theta \mid Y_{obs}, M) \propto p(Y_{obs}, M \mid \theta) p(\theta). \tag{8}$$

1.4　贝叶斯方法与 WinBUGS

WinBUGS 是一款针对贝叶斯分析使用非常广泛的免费软件，不管是简单模型或复杂模型，它处理起来都十分灵活. 一个完整的 WinBUGS 程序包含 3 个部分：模型说明、初始值和数据. 在指定分析数据时，如果含缺失数据，可用“NA”表示. 在使用完全贝叶斯方法时，当缺失数据为“NA”时，WinBUGS 就会视缺失数据为未知参数，跟模型参数没有本质上的区别.

对比完全贝叶斯方法，WinBUGS 中的部分贝叶斯方法更加简单，在指定的分析数据只要保留观测数据即可，此时只有模型参数参与估计时，而缺失数据相当于没有提供任何信息.

2　数据模拟及处理方法

2.1　模型选择

本文采用的模型是潜变量增长曲线模型[13]，它适用于在某几个固定时间点观测得来的纵向研究资料. 在潜变量增长曲线模型中，是用潜变量来描述总体的平均增长趋势和依时间变化的情况，并且可以分析总体之间存在的差异.

该模型可以用截距和斜率 2 个潜变量分别来描述初始水平和增长趋势水平，数学表达式为

$$\begin{aligned} y_{it} &= l_i + a_t s_i + e_{it}, \\ l_i &= \mu_l + \nu l_i, \\ s_i &= \mu_s + \nu s_i, \end{aligned} \tag{9}$$

其中 y_{it} 代表第 i 个个体在 t 个时间点上的观测值，l_i 表个体潜在初始水平，即潜变量截距，μ_l 表示所有个体初始水平的均值；s_i 表示个体增长斜率，是模型中的另外一个潜变量，μ_s 是所有个体增长斜率的均值；并通过调整 a_t 的不同因素载荷可以得到不同形态的增长曲线模型.

2.2　模拟数据

本文用蒙特卡洛方法产生 1 000 个被试在 4 次重复测量上的反应矩阵，分别有 4 种缺失比例（$p=0, p=0.1, p=0.3, p=0.5$）. 先根据参数分布模拟产生完整反应矩阵，再在完全随机缺失下模拟缺失反应矩阵.

对于 4 种缺失比例模拟含有缺失反应的反应矩阵，再分别运用两种缺失数据处理方法估计模型参数，每种组合情况均重复模拟 50 次，即建立 50 个含不同缺失反应的数据集. 所有模拟和缺失数据处理过程均在 R 软件[16]上实现，潜变量曲线增长模型的 WinBUGS 程序是参考 E. Y. Zhang 等的程序[13].

2.3　处理方法

参考 Zhang Zhiyong 等研究的先验分布，本文所有缺失数据处理方法都采用无信息先验[13]，模型参数的先验如表 1 所示.

表 1　潜变量增长曲线模型的参数先验

参数	无信息先验
初始水平（μ_l）	$dnorm(0,1,0E-6)$
总增长水平（μ_s）	$dnorm(0,1,0E-6)$
1、2 次增长比例（α_2）	$dnorm(0,1,0E-6)$
1、3 次增长比例（α_3）	$dnorm(0,1,0E-6)$
测量误差的倒数（$1/\sigma_e^2$）	$dgamma(.001..001)$
随机效应的协方差（$\begin{pmatrix} \sigma_L^2 & \sigma_{LS} \\ \sigma_{LS} & \sigma_S^2 \end{pmatrix}$）	$dwish\left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, 2\right]$

2.4 比较标准

为了比较 2 种缺失数据处理方法恢复真实参数的性能，这里运用指标均方根误差 RMSE (root mean square errors)[17].

$$RMSE(l)=\sqrt{\frac{\sum_{i=1}^{I}\sum_{m=1}^{M}(\hat{l_i}-l_i)^2}{MI}}, \tag{10}$$

$$RMSE(s)=\sqrt{\frac{\sum_{i=1}^{I}\sum_{m=1}^{M}(\hat{s_i}-s_i)^2}{MI}}, \tag{11}$$

其中 l 为初始水平，s 为增长水平，M 为模拟重复次数，I 为被试人数. *RMSE* 越小越好，说明估计参数与真实参数的差异越小.

3 结果分析

对 4 种缺失比例的模拟反应矩阵分别运用 2 种缺失数据处理方法处理后进行参数估计(整个模拟、缺失处理及估计过程重复 50 次)，分别计算出 *RMSE*，结果见表 2.

表 **2** 4 种缺失比例下 2 种处理方法得到的 ***RMSE***

缺失比例 p	*RMSE*			
	l 初始水平		s 增长水平	
	部分贝叶斯	完全贝叶斯	部分贝叶斯	完全贝叶斯
p=0	0.373 1	0.373 1	0.000 1	0.000 1
p=0.1	0.391 1	0.373 8	0.029 9	0.029 3
p=0.3	0.392 1	0.392 1	0.133 2	0.133 3
p=0.5	0.617 4	0.468 9	0.472 9	0.261 3

为了更直观地不同缺失比例情况下，分别画出所有组合情况的初始水平 *RMSE* 散点图和增长水平 *RMSE* 散点图，如图 2 和图 3 所示，其中横坐标为缺失比例，纵坐标为 *RMSE* 大小.

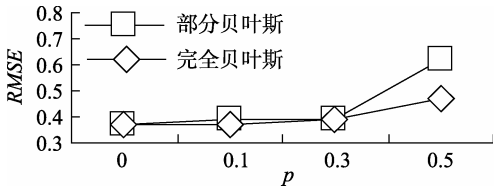

图 **2** 不同缺失比例下初始水平 *RMSE* 散点图

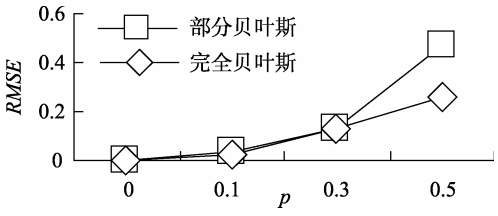

图 **3** 不同缺失比例下增长水平 *RMSE* 散点图

3.1 初始水平(l_i)***RMSE*** 变化情况

如图 2 所示，以缺失比例为 0 为基准线，随着缺失比例的增多，这 2 种处理方法的 *RMSE* 都会增大，其中它们的 *RMSE* 增长情况在 p=0.1 和 p=0.3 的情况下几乎相同，但在 p=0.5 时，完全贝叶斯方法得到的 RMSE 较部分贝叶斯的 *RMSE* 会小很多.

3.2 增长水平(s_i)***RMSE*** 变化情况

在图3 中，随着缺失比例的增多，完全贝叶斯方法的 *RMSE* 与部分贝叶斯方法的增长情况比较相似，但在缺失比例为 0.5 的时候，前者变化较小，而后者陡增.

综合初始水平和增长水平 *RMSE* 变化情况，不难发现在完全随机缺失的条件下，随着缺失比例的增加，模型参数估计的精度都会降低；完全贝叶斯方法和部分贝叶斯方法在 p=0.1 和 p=0.3 时几乎相同，只在缺失比例为 0.5 时，前者明显优于后者.

4 讨论

在纵向研究中，经常会遇到个体流失的情况，所以不管采用哪种模型来拟合纵向数据，都要考虑采用一种有效的缺失数据处理方法来减小估计误差. 完全贝叶斯方法作为一种基于模型的方法，逐渐受到研究者们的青睐[11-13]，但部分贝叶斯方法则较少用到纵向模型的缺失数据处理中，它的适用性未得到检验与比较.

本文选择数据完全随机缺失[2]的情况，在贝叶斯理论框架下用潜变量曲线增长模型拟合纵向数据. 本研究发现当数据缺失比例较大时(≥50%)，使用完全贝叶斯方法进行参数估计结果比用部分贝叶斯方法会得到较小的 *RMSE*，说明它是一种更为有效的处理方法. 但在缺失比例低于 50%时，2 种方法的参数估计效果则差异不大. 本研究认为在实际应用中，完全贝叶斯方法更容易操作. 例如 WinBUGS 软件中该方法只要用“NA”代替缺失数据，程序便能识别；而部分贝叶斯方法需要预处理来删除缺失反应.

另外本文关于上述 2 种方法的比较基于完全随机缺失数据的条件. 而在实际纵向数据中，缺失数据并不一定具有随机完全性，比如某一类型的个体可能比另一类型的个体更容易流失. 这就需要考虑缺失数据的类型，即随机缺失还是非随机缺失. 如何根据实测结果来判别缺失数据的类型，进而选择合适的处理方法将是本研究的今后课题.

5　参考文献

[1] 刘红云，孟庆茂. 纵向数据分析方法 [J]. 心理科学进展, 2003, 11(5): 586-92.

[2] 利特尔，鲁宾. 缺失数据统计分析 [M]. 2 版. 孙山泽，译. 北京：中国统计出版社, 2004.

[3] Nakai M, Ke W. Review of methods for handing missing data in longitudinal data analysis [J]. International Journal fo Mathematical Analysis, 2011, 5(1): 1-13.

[4] Graham J　W. Missing data analysis: making it work in the real world [J]. The Annual Review of Psychology, 2009, 60: 549-76.

[5] Little R J A. Modeling the drop-out mechanism in repeated-measures studies [J]. Journal of the American Statistical Association, 1995, 90: 1112-1121.

[6] In Litter T D, Schnabel K U, Baumert J. Modeling longitudinal and multilevel data: practice issues, applied approaches and specific examples [C]. New Jersey: Laerence Erlbaum Associates, 2000: 15-32.

[7] Rubin D B. Multiple imputation for nonresponse in survers [M]. New York: Wiley, 1987.

[8] Ibrahim J G, Chen M H, Lipsitz S R, et al. Missing-data methods for gerneralized linear models: a comparative review [J]. Journal of American Statistical Association, 2005, 469(100): 332-346.

[9] Newman D A. Longitudinal modeling with randomly and systematically missing data: a simulation of Ad hoc, maximum likelihood, and multiple impution techniques [J]. Organizational Research Methods , 2003, 6(3): 328-362.

[10] Edwards W, Lindman H, Savage L J. Bayesian statistical inference for psychological research [J]. Psychological Review, 1963, 70: 193-242.

[11] Carrigan G, Barnett A G, et al. Compensating missing data from longitudinal studies using WinBUGS [J]. Journal of Statistical Software, 2007, 19(7): .

[12] Li Jinhui. Analysis of longitudinal data with missing values [D]. California : University of　California, 2006.

[13] Zhang Zhiyong, Hamagami F, Wang L J, et al. Bayesian analysis of longitudinal data using growth curve models [J]. International Journal of Behavioral Development, 2007, 31 (4): 374-83.

[14] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian retoration of image [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6: 721-741.

[15] Spiegelhalter D J, Thomas A, Best NG, et al. WinBUGS Version 1.4 User Manual [EB/OL]. [2011-12-16]. http: //www. mrc-bsu. cam. ac. uk/bugs/.

[16] R Development Core Team. R: a language and environment for statistical computing[EB/OL]. [2012-12-16].　http: //www. R-project. org/.

[17] 漆书青，戴海琦，丁树良. 现代教育与心理测量学原理 [M]. 北京：高等教育出版社, 2002.

The Comparision of Two Approaches to Bayesian Method for Missing Data in Longitudinal Model
—— Growth Curve Model for Example

YANG Lin-shan[1], CAO Yi-wei[2]

(1.Shenzhen Seaskyland Education Assessment Co. Ltd., Shenzhen Guangdong 518067, China;
2.Normal College, Shenzhen University, Shenzhen Guangdong 518060, China)

Abstract: The research explored the relative performance of Fully Bayesian method and Partially method in the estimation of growth curve model parameters. Only Simulation studies were used in the compassion in which four missing rates (0, 0.10, 0.30, and 0.50) were investigated. In each situation, 50 matrixes with missing response were generated and the index *RMSE* (root mean square error) were compared the two approaches.

The results showed that: (1) the accuracy of parameter estimations of the two approaches were both affected by the missing rate, and as the increasing of missing rate, the bigger of *RMSE*. (2)When the missing rate is small, the *RMSEs* of the two approaches were almost same, however, Fully Bayesian method got better than Partially method when missing rate came to 0.50.

Key words: missing data; fully Bayesian; partially Bayesian; longitudinal models; WinBUGS

(责任编辑：冉小晓)