doi:10.16112/j.cnki.53-1223/n.2020.01.018

带有不可忽略缺失数据的联合均值与方差模型的贝叶斯估计

赵远英[1],吴刘仓[2],徐登可[3]
(1. 贵阳学院 数学与信息科学学院,贵州 贵阳 550005; 2. 昆明理工大学 理学院,云南 昆明 650500;
3. 浙江农林大学 统计系,浙江 杭州 311300)

摘要:对响应变量带有不可忽略缺失数据的联合均值与方差模型的贝叶斯估计问题进行了研究.缺失数据机制通过 logistic 回归模型来指定,模型参数和缺失数据机制参数的联合贝叶斯估计通过运用 MH 算法及 Gibbs 抽样获得,并用数值计算阐明上述方法的可行性.

关键词:贝叶斯估计; Gibbs 抽样; 联合均值与方差模型; MH 算法; 不可忽略缺失数据
中图分类号:O212.8 **文献标志码**:A **文章编号**:1007-855X(2020)01-0125-08

Bayesian Estimation for Joint Mean and Variance Models with Nonignorable Missing Data

ZHAO Yuanying[1], WU Liucang[2], XU Dengke[3]
(1. College of Mathematics and Information Science, Guiyang University, Guiyang 550005, China;
2. Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China;
3. Department of Statistics, Zhejiang Agriculture and Forestry University, Hangzhou 311300, China)

Abstract: A Bayesian method is proposed to analyze joint mean and variance models in which the responses are missing with non-ignorable missingness mechanism. The missingness mechanism is expressed by a logistic model, and Metropolis-Hastings algorithm and Gibbs sampler are employed to obtain the joint Bayesian estimation for model parameters and missingness mechanism parameters. The proposed methodology is demonstrated by numerical examples.

Key words: Bayesian estimation; Gibbs sampler; joint mean and variance models; Metropolis-Hastings algorithm; nonignorable missing data

0 引 言

随机误差项的同方差性是古典回归问题中一个最简单最基本的假设条件,但是对实际问题进行统计模型时由于各种原因方差齐性未必满足,因此对方差建模显得尤其重要,一方面是因为方差建模自身具备科学的理论意义,并且对均值模型有效的参数估计和准确的统计推断起到十分关键的作用. 另一方面,要达到很好地控制方差的目标,一个自然的想法就是期望更多更好地剖析方差的来源. 这些年以来对均值方差联合建模的研究已经取得了大量成果. 例如,Harvey [1] 对异方差回归模型的两步估计方法和似然估计方法做出比较; 在正态回归模型的框架下, Aitkin [2] 用对数线性模型对方差进行建模,并用两个实例说明其提出来的方法; Verbyla [3] 在正态分布联合均值与方差模型的框架下,研究其似然推断和统计诊断问题; 吴

收稿日期:2018-11-20. 基金项目:国家自然科学基金项目(11861041,11761016,11801514); 浙江省自然科学基金项目(LY17A010026); 贵州省贵阳市科技局贵阳学院科技专项资金项目(GYU-KYZ(2019-2020)PT04-04); 贵州省高等学校大学生创新创业训练计划项目(20195201381).
作者简介:赵远英(1981-),男,博士,副教授. 主要研究方向: 应用统计. E-mail: zhaoyuanying_@126.com
通信作者:吴刘仓(1976-),男,博士,教授. 主要研究方向: 应用统计. E-mail: wuliucang@163.com

刘仓等[4]基于惩罚似然的方法对均值模型和方差模型进行同时变量选择研究;关于均值方差联合建模更多的理论方法,请查阅文献[5-6]等.

在统计学和相关学科研究过程中经常会遇到缺失数据问题,若直接运用完全数据的统计理论和方法处理缺失数据,可能会出现较大偏差的参数估计的结果,甚至导致产生偏离事实的结论或得出不符合实际情况的解释等.对这个问题的研究已经得到了众多统计学学者的关注.比如,Little等[7]在其专著中对缺失数据的历史发展、分析方法和应用领域进行比较系统介绍;Lee等[8]基于贝叶斯统计框架和不可忽略缺失数据机制研究结构方程模型的统计推断问题;Zhou等[9]将估计方程、经验似然理论以及基于核估计的缺失数据方法相结合研究了缺失数据模型的参数估计问题;Zhang等[10]研究了响应变量存在缺失的条件下测量误差模型的限制经验似然置限区间;针对非线性模型,Tang等[11]探讨了随机缺失数据下非线性模型的经验似然估计;Guo等[12]讨论了协变量随机缺失条件下单指标模型的经验似然估计问题,等等.但是在贝叶斯统计框架下带有缺失数据的联合均值与方差模型的统计推断却未见相关报道.本文研究响应变量带有不可忽略缺失数据的联合均值与方差模型的贝叶斯估计问题.

本文的框架结构见下:第1部分定义NMAR机制下的联合均值与方差模型;在文章的第2部分,缺失数据机制参数和模型参数的联合贝叶斯估计通过运用Metropolis-Hastings(MH)算法[14-15]与Gibbs抽样[13]获得;第3部分,随机模拟研究和实例分析的数值例子说明本文所提出的贝叶斯估计方法的可行性.第4部分给出本文的结论.

1 模型与假设

1.1 联合均值与方差模型

本文对联合均值与方差模型进行如下定义:

$$\begin{cases} y_i \sim N(\mu_i,\sigma_i^2) \\ \mu_i = x_i^{\mathrm{T}}\beta = \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1,\cdots,n \\ \log(\sigma_i^2) = z_i^{\mathrm{T}}\alpha = \sum_{k=1}^{q} z_{ik}\alpha_k \end{cases} \tag{1}$$

其中:符号$y_i \sim N(\mu_i,\sigma_i^2)$表示响应变量$y_i$服从均值为$\mu_i$,方差为$\sigma_i^2$的正态分布,$x_i = (x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$以及$z_i = (z_{i1},x_{i2},\cdots,z_{iq})^{\mathrm{T}}$各自代表均值模型以及方差模型中的协变量,向量$\beta = (\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$以及向量$\alpha = (\alpha_1,\alpha_2,\cdots,\alpha_q)^{\mathrm{T}}$分别是$p \times 1$以及$q \times 1$的未知参数.

1.2 带有不可忽略缺失数据的联合均值与方差模型

对数据$\{(y_i,x_i,z_i): i = 1,\cdots,n\}$,本文假设$\{(x_i,z_i): i = 1,\cdots,n\}$是可以完全观测的,而$\{y_i: i = 1,\cdots,n\}$可能存在缺失的情况.为此引进响应变量$y_i$是否缺失的示性变量$r_i$,当$y_i$缺失时(记$y_i = y_{im}$)令$r_i = 1$,当$y_i$可观测时(记$y_i = y_{io}$)令$r_i = 0$.类似于Ibrahim,Chen等[16]的方法,如下的logistic模型被用以建立变量y_i的缺失机制

$$\mathrm{log}it\{p(r_i = 1 \mid y_i,x_i,z_i,\varphi)\} = \log\frac{p(r_i = 1 \mid y_i,x_i,z_i,\varphi)}{1 - p(r_i = 1 \mid y_i,x_i,z_i,\varphi)} = \varphi_1 + \varphi_2 y_i + \varphi_3 x_{i1}\cdots + \varphi_{p+2}x_{ip} \tag{2}$$

这里:$\varphi = (\varphi_1,\varphi_2,\cdots,\varphi_{p+2})^{\mathrm{T}}$为缺失机制模型中的未知参数.在模型(2)中若$\varphi_2 \neq 0$,则缺失数据的缺失机制即为非随机缺失(NMAR),亦即不可忽略的缺失数据;若$\varphi_2 = 0$,但$\varphi_3,\cdots,\varphi_{p+2}$中至少有一个不为0,则缺失数据的缺失机制即为随机缺失(MAR);若$\varphi_2 = \cdots = \varphi_{p+2} = 0$,则缺失数据的缺失机制即为完全随机缺失(MCAR).

记示性向量$r = (r_1,\cdots,r_n)^{\mathrm{T}}$,$y = \{y_1,\cdots,y_n\}$,$x = \{x_1,\cdots,x_n\}$和$z = \{z_1,\cdots,z_n\}$,则在上述假设条件下有:

$$p(r \mid y,x,z,\varphi) = \prod_{i=1}^{n}\{p(r_i = 1 \mid y_i,x_i,z_i,\varphi)\}^{r_i}\{1 - p(r_i = 1 \mid y_i,x_i,z_i,\varphi)\}^{1-r_i} \tag{3}$$

由于响应变量构成的集合可以表示为 $y = \{y_{obs}, y_{mis}\}$,其中 y_{obs} 和 y_{mis} 分别表示响应变量可观测部分和缺失部分构成的集合,故 $D = \{y, x, z, r\}$ 是完全数据集,$D_{obs} = \{y_{obs}, x, z, r\}$ 是观测数据集. 本文的主要工作就是在带有不可忽略数据(非随机缺失数据) y_{mis} 联合均值与方差模型的框架下,基于观测数据集 D_{obs} 讨论模型参数 $\theta = (\beta^T, \alpha^T)^T$ 和缺失机制参数 φ 的联合贝叶斯估计问题.

2 模型的贝叶斯分析

为了在贝叶斯统计理论框架下进行统计分析,首先得假定未知参数 θ 和 φ 的联合先验分布. 本文对参数 θ 和 φ 的先验分布作以下假定:

$$p(\theta, \varphi) = p(\theta)\, p(\varphi) = p(\beta)\, p(\alpha)\, p(\varphi) \tag{4}$$

其中:

$$p(\beta) \sim N(\beta_0, \sum\nolimits_{\beta}), p(\alpha) \sim N(\alpha_0, \sum\nolimits_{\alpha}), p(\varphi) \sim N(\varphi_0, \sum\nolimits_{\varphi}) \tag{5}$$

这里: $N(\beta, \sum)$ 表示数学期望为 β,协方差阵为 $\sum$ 的多元正态分布,β_0, $\sum_{\beta}$, α_0, $\sum_{\alpha}$, φ_0 和 $\sum_{\varphi}$ 是事先给定的超参数.

根据联合均值与方差模型的模型假设, $\{\theta, \varphi\}$ 基于观测数据集 D_{obs} 的贝叶斯统计推断可以根据下面的式子

$$p(\theta, \varphi \mid D_{obs}) \propto \left\{\int \prod_{i=1}^{n} p(y_i \mid x_i, z_i, \theta)\, p(r_i \mid y_i, x_i, z_i, \varphi)\, dy_{mis}\right\} p(\theta, \varphi) \tag{6}$$

进行. 易见条件分布 $p(\theta, \varphi \mid D_{obs})$ 中包含非常棘手的高维积分问题,通常情况下都没有显式解. 为了解决上述困难,类似于唐年胜等[17], $\{\theta, \varphi, y_{mis}\}$ 基于 D_{obs} 的后验分布为:

$$p(\theta, \varphi, y_{mis} \mid D_{obs}) \propto p(\theta, \varphi, D) \propto \left\{\prod_{i=1}^{n} p(y_i \mid x_i, z_i, \theta)\, p(r_i \mid y_i, x_i, z_i, \varphi)\right\} p(\theta, \varphi) \tag{7}$$

同(6) 对比,(7) 没有涉及高维积分问题. 然而由于分层模型的复杂性,容易看出直接从后验分布 $p(\theta, \varphi, y_{mis} \mid D_{obs})$ 中抽样仍然是几乎不可能的,因此本文借助 Gibbs 抽样[13] 算法获得联合后验分布 $p(\theta, \varphi, y_{mis} \mid D_{obs})$ 的随机观测序列 $\{(\theta^{(t)}, \varphi^{(t)}, y_{mis}^{(t)}): t = 1, 2, \cdots, T\}$,之后依据这个随机观测序列进行相关的贝叶斯统计分析. 从 $p(\theta, \varphi, y_{mis} \mid D_{obs})$ 中抽取 $\{(\theta^{(t)}, \varphi^{(t)}, y_{mis}^{(t)}): t = 1, 2, \cdots, T\}$ 的 Gibbs 抽样[13] 详细步骤为: (1) 选取参数 θ, φ 和 y_{mis} 的初始值,并令 $t = 0$; (2) 从条件分布 $p(\theta \mid y_{obs}, x, z, y_{mis}^{(t)})$ 中抽取 $\theta^{(t+1)}$; (3) 从条件分布 $p(\varphi \mid y_{obs}, x, z, r, y_{mis}^{(t)})$ 中抽取 $\varphi^{(t+1)}$; (4) 从条件分布 $p(y_{mis} \mid y_{obs}, x, z, r, \theta^{(t+1)}, \varphi^{(t+1)})$ 中抽取 $y_{mis}^{(t+1)}$; (5) 令 $t = t + 1$,重复步骤(2) ~ (4) 直至算法收敛.

2.1 条件分布

下面具体给出 Gibbs 抽样[13] 算法中所需要的条件分布. 首先给出步骤(2) 中 θ 的条件分布:

$$p(\theta \mid y_{obs}, x, z, y_{mis}) = p(\theta \mid y, x, z) \propto p(y \mid x, z, \theta)\, p(\theta) \propto$$

$$\exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left((y_i - \mu_i)^2 \sigma_i^{-2} + \log(\sigma_i^2)\right) + (\beta - \beta_0)^T \sum\nolimits_{\beta}^{-1} (\beta - \beta_0) + (\alpha - \alpha_0)^T \sum\nolimits_{\alpha}^{-1} (\alpha - \alpha_0)\right]\right\}. \tag{8}$$

由于 $\theta = (\beta^T, \alpha^T)^T$,下面分别给出 β 和 α 的条件分布. 对 β 的条件分布有:

$$p(\beta \mid y, x, z, \alpha) \propto p(y \mid x, z, \theta)\, p(\beta) \propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(y_i - \mu_i)^2 \sigma_i^{-2} + (\beta - \beta_0)^T \sum\nolimits_{\beta}^{-1} (\beta - \beta_0)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}(\beta - \beta^*)^T \sum\nolimits^{*-1} (\beta - \beta^*)\right\}$$

即

$$\beta \sim N(\beta^*, \sum\nolimits^{*}) \tag{9}$$

其中:

$$\beta^{*} = \sum^{*} \left[\sum\nolimits_{\beta}^{-1}\beta_0 + \sum_{i=1}^{n}\sigma_i^{-2}x_i y_i\right], \sum^{*} = \left(\sum\nolimits_{\beta}^{-1} + \sum_{i=1}^{n}\sigma_i^{-2}x_i x_i^{\mathrm{T}}\right)^{-1}.$$

对 α 的条件分布有:

$$\begin{aligned} p(\alpha \mid y,x,z,\beta) &\propto p(y \mid x,z,\theta)\,p(\alpha) \\ &\propto \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left((y_i-\mu_i)^2\sigma_i^{-2}+\log(\sigma_i^2)\right)+(\alpha-\alpha_0)^{\mathrm{T}}\sum\nolimits_{\alpha}^{-1}(\alpha-\alpha_0)\right]\right\} \end{aligned} \tag{10}$$

其次给出步骤(3)中 φ 的条件分布:

$$\begin{aligned} p(\varphi \mid y_{\mathrm{obs}},x,z,r,y_{\mathrm{mis}}) &= p(\varphi \mid y,x,z,r) \propto p(r \mid y,x,z,\varphi)\,p(\varphi) \propto \\ &\frac{\exp\left\{\sum_{i=1}^{n}r_i\varphi^{\mathrm{T}}\omega_i-\frac{1}{2}(\varphi-\varphi_0)^{\mathrm{T}}\sum\nolimits_{\varphi}^{-1}(\varphi-\varphi_0)\right\}}{\prod_{i=1}^{n}(1+\exp(\varphi^{\mathrm{T}}\omega_i))^{r_i}} \end{aligned} \tag{11}$$

其中: $\omega_i = (1,y_i,x_i^{\mathrm{T}})^{\mathrm{T}}$.

最后给出步骤(4)中 y_{mis} 的条件分布,由于 y_{mis} 是响应变量缺失部分构成的集合,即 $y_{\mathrm{mis}} = \{y_{mi}\}$, 下面给出 y_{mi} 的条件分布:

$$\begin{aligned} p(y_{mi} \mid x_i,z_i,r_i,\theta,\varphi) &\propto p(y_{mi} \mid x_i,z_i,\theta) \times p(r_i \mid y_{mi},x_i,z_i,\varphi) \\ &\propto \exp\left\{-\frac{1}{2}(y_{mi}-\mu_i)^2\sigma_i^{-2}+r_i\varphi^{\mathrm{T}}\omega_i-\log(1+\exp(\varphi^{\mathrm{T}}\omega_i))\right\} \end{aligned} \tag{12}$$

其中: $\omega_i = (1,y_{mi},x_i^{\mathrm{T}})^{\mathrm{T}}$.

2.2 完成抽样

由上面的推导可知,条件分布(9)是常见的标准分布,能够直接抽样得到该分布的样本. 但是(10) ~ (12)中的后验分布都是非标准分布,不能从这些分布中直接抽样获得其样本. 我们使用 *MH* 算法[14-15]来解决这一困难. 这里分别选取多元正态分布 $N(0,\sigma_\alpha^2\Omega_\alpha)$, $N(0,\sigma_\varphi^2\Omega_\varphi)$ 和 $N(0,\sigma_{y_{\mathrm{mis}}}^2\Omega_{y_{\mathrm{mis}}})$ 为(10) ~ (12)中后验分布的建议分布,其中:

$$\Omega_\alpha^{-1} = -\frac{\partial^2\log\{p(y \mid x,z,\theta)\}}{\partial\alpha\partial\alpha^{\mathrm{T}}}\Big|_{\alpha=0} + \sum\nolimits_{\alpha}^{-1}, \Omega_\varphi^{-1} = \frac{1}{4}r_i\omega_i\omega_i^{\mathrm{T}} + \sum\nolimits_{\varphi}^{-1},$$

$$\Omega_{y_{\mathrm{mis}}}^{-1} = -\frac{\partial^2\log\{p(y_{mi} \mid x_i,z_i,r_i,\theta,\varphi)\}}{\partial y_{mi}^2}\Big|_{y_{mi}=0}.$$

从后验分布 $p(\alpha \mid y,x,z,\beta)$ 产生样本 α 的 MH 算法[14-15]的步骤如下: 已知 α 在第 t 次迭代时的迭代值为 $\alpha^{(t)}$,从 $N(\alpha^{(t)},\sigma_\alpha^2\Omega_\alpha)$ 中随机产生样本 α^{*},之后从均匀分布 $U(0,1)$ 中独立地产生随机样本 u,如果 $u \leqslant \min\left\{1,\frac{p(\alpha^{*} \mid y,x,z,\beta)}{p(\alpha^{(t)} \mid y,x,z,\beta)}\right\}$,则令 $\alpha^{(t+1)} = \alpha^{*}$;否则令 $\alpha^{(t+1)} = \alpha^{(t)}$. 在具体应用时,经常选择 σ_α^2 使得接受 α 的平均概率在 [0.25,0.45] 之间. 从分布 $p(\varphi \mid y,x,r)$ 中产生样本 φ 以及从分布 $p(y_{mi} \mid x_i,z_i,r_i,\theta,\varphi)$ 中产生样本 y_{mi} 的 MH 算法[14-15]也可以类似得到.

2.3 贝叶斯估计

根据上述 Gibbs 抽样[13]算法获得的后验分布 $p(\theta,\varphi,y_{\mathrm{mis}} \mid D_{\mathrm{obs}})$ 的随机观测序列 $\{(\theta^{(t)},\varphi^{(t)},y_{\mathrm{mis}}^{(t)}): t = 1,2,\cdots,T\}$,参数 θ,φ 和 y_{mis} 的联合贝叶斯估计可以表示为:

$$\hat{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}, \hat{\varphi} = \frac{1}{T}\sum_{t=1}^{T}\varphi^{(t)} \text{ 和 } \hat{y}_{\mathrm{mis}} = \frac{1}{T}\sum_{t=1}^{T}y_{\mathrm{mis}}^{(t)}, \tag{13}$$

和 Geyer[18]相似,上式中未知参数 θ 的贝叶斯估计是 θ 后验均值的相合估计; θ 的第 j 分量的 $1-\tau$(显著性水平 τ 通常为 $\tau = 0.05$ 或 0.1 等)后验置信区间为 $(\theta_{jl},\theta_{ju})$,这里 θ_{jl} 与 θ_{ju} 分别代表随机序列 $\{\theta^{(t)}: t =$

$1,2,\cdots,T\}$ 第 j 分量递增序列的 $\frac{\tau}{2}$ 与 $1-\frac{\tau}{2}$ 分位数.

3 数值例子

3.1 随机模拟

为了展示本文贝叶斯估计方法的有效性,我们做了如下的随机模拟. 对任意给定的 $i=1,\cdots,n$,假定 $y_i \sim N(\mu_i,\sigma_i^2)$,,μ_i 与 σ_i^2 的具体表达式分别如下:

$$\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 \tag{14}$$

$$\log(\sigma_i^2) = z_{i1}\alpha_1 + z_{i2}\alpha_2 \tag{15}$$

协变量 $x_i=(x_{i1},x_{i2},x_{i3},x_{i4})^{\mathrm{T}}$ 和 $z_i=(z_{i1},z_{i2})^{\mathrm{T}}$ 分别从标准正态分布 $N(0,I_4)$ 与 $N(0,I_2)$ 中独立地产生,符号 I_m 是 $m\times m$ 的单位阵. 设置参数的真值如下:$\beta^{\mathrm{T}}=(\beta_1,\beta_2,\beta_3,\beta_4)=(1,1,1,1)$ 和 $\alpha^{\mathrm{T}}=(\alpha_1,\alpha_2)=(1,1)$. 进一步,我们假设响应变量 y_i 存在部分数据缺失. 考虑使用以下 2 种缺失数据机制模型来产生模拟研究所需的数据样本.

Type 1(NMAR 机制):假设缺失机制模型为:

$$\mathrm{logit}\{p(r_i=1\mid y_i,x_i,z_i,\varphi)\} = \varphi_1+\varphi_2 y_i+\varphi_3 x_{i1}+\varphi_4 x_{i2}+\varphi_5 x_{i3}+\varphi_6 x_{i4} \tag{16}$$

其中:φ 的真值为 $\varphi=(\varphi_1,\varphi_2,\varphi_3,\varphi_4,\varphi_5,\varphi_6)^{\mathrm{T}}=(-1,0.3,0.3,0.3,0.3,0.3)^{\mathrm{T}}$.

Type 2(MAR 机制):记 $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$,符号 $[m]$ 表示实数 m 的整数部分. 假设缺失数据的缺失机制模型为:(1) 对任意的 $i=1,\cdots,[\frac{n}{4}]$,从正态分布 $N(0,1)$ 中独立地产生 ζ_i,当 $\zeta_i+1.0\leqslant y_i$ 时响应变量 y_i 为缺失数据;(2) 对任意的 $i=[\frac{n}{4}]+1,\cdots,[\frac{n}{2}]$,独立地从标准正态分布 $N(0,1)$ 中产生 ζ_i,当 $\zeta_i-1.0\leqslant y_i$ 时响应变量 y_i 为缺失数据;(3) 对任意的 $i=[\frac{n}{2}]+1,\cdots,[\frac{3n}{4}]$,当 $\bar{y}\leqslant y_i$ 时响应变量 y_i 为缺失数据;(4) 对任意的 $i=[\frac{3n}{4}]+1,\cdots,n$,当 $\bar{y}\geqslant y_i$ 时响应变量 y_i 为缺失数据. 显然,该缺失数据的缺失机制是 MAR 机制.

基于上述模型假设和指定的参数真值,产生样本量为 $n=80$ 的模拟数据样本,在上述两种缺失数据机制模型下,产生的模拟数据样本中响应变量 y_i 的缺失比例分别约为 36.38% 和 43.16%. 我们都用模型(1) 和(14) ~ (16) 拟合产生的样本. 为了评估贝叶斯估计对先验分布是否敏感,本文考虑以下两种类型的超参数:

Case I:$\beta_0=\beta^0,\alpha_0=\alpha^0$ 和 $\varphi_0=\varphi^0$,$\sum_\beta=0.25I_4$,$\sum_\alpha=0.25I_2$ 和 $\sum_\varphi=0.25I_6$,其中 β^0,α^0 和 φ^0 分别是 β,α 和 φ 的真值,此类型被视为具有良好的先验信息;

Case II:$\beta_0=(0,0,0,0)^{\mathrm{T}},\alpha_0=(0,0)^{\mathrm{T}}$ 和 $\varphi_0=\varphi^0$,$\sum_\beta=10I_4$,$\sum_\alpha=10I_2$ 和 $\sum_\varphi=10I_6$,此类型被视为无信息先验.

本文从 Gibbs 抽样算法收敛[19] 后的样本中,收集 5 000 个随机观察序列计算式子(13) 中的联合贝叶斯估计. 本文做了 100 次重复试验来评估参数估计的效果. 表 1 给出了带缺失数据的联合均值与方差模型未知参数的贝叶斯估计结果,这里 'RMS' 代表参数真值与 100 次重复试验的贝叶斯估计的偏差的平方的平均值的算术平方根,'SD' 代表 100 次重复试验参数贝叶斯估计的样本标准差,'BIAS' 代表参数真值与 100 次重复试验的贝叶斯估计的平均值的偏差的绝对值. 根据表 1 不难得到以下的结论:(1) 贝叶斯估计对先验信息并不太敏感,且第一种类型的贝叶斯估计略优于第二种类型的贝叶斯估计;(2) 两种先验类型所有参数的 'BIAS' 值都小于 0.07,表明贝叶斯估计在不同的先验类型下具有很高的估计精度;(3) 'SD' 的值与 'RMS' 的值十分接近,表明贝叶斯估计的标准差非常准确;(4) 当缺失数据机制是 MAR 机制时,用 NMAR 缺失机制模型来拟合数据样本得到贝叶斯估计也是非常理想的.

表1 随机模拟未知参数的贝叶斯估计

Tab. 1 Bayesian estimation of unknown parameters in stochastic simulation

		NMAR			MAR		
Case	参数	BIAS	SD	RMS	BIAS	SD	RMS
I	β_1	0.009	0.102	0.102	0.022	0.110	0.112
	β_2	0.019	0.105	0.107	0.016	0.100	0.100
	β_3	0.003	0.107	0.106	0.021	0.103	0.104
	β_4	0.010	0.105	0.105	0.008	0.103	0.103
	α_1	0.002	0.206	0.205	0.035	0.190	0.193
	α_1	0.012	0.192	0.191	0.035	0.191	0.193
	Sum *RMS*			0.816			0.805
II	β_1	0.014	0.112	0.112	0.027	0.123	0.125
	β_2	0.026	0.116	0.118	0.016	0.107	0.107
	β_3	0.007	0.113	0.113	0.026	0.112	0.115
	β_4	0.014	0.114	0.114	0.008	0.115	0.114
	α_1	0.009	0.260	0.259	0.061	0.243	0.249
	α_2	0.032	0.248	0.248	0.064	0.256	0.263
	Sum *RMS*			0.964			0.973

3.2 实例分析

本文采用的实际案例是MINITAB树数据[2]. Aiktin[2]曾使用下面联合均值与方差模型(17)对该数据进行拟合,

$$\begin{cases} y_i \sim N(\mu_i, \sigma_i^2) \\ \mu_i = \beta_1 + x_{i1}\beta_2 + x_{i2}\beta_3 \\ \log(\sigma_i^2) = \alpha_1 + z_{i1}\alpha_2 + z_{i2}\alpha_3 \end{cases} \qquad i = 1,2,\cdots,31, \tag{17}$$

结果显示此模型能较理想地反映树的体积V对树高H与直径D的关系. 在模型(17)中响应变量$y_i = \log V_i$,均值模型与方差模型中的协变量分别是$x_{i1} = \log H_i, x_{i2} = \log D_i, z_{i1} = \log D_i$,和$z_{i2} = \log D_i \times \log D_i$,相关的估计结果为$\hat{\beta} = (-6.390, 1.080, 1.955)^{\mathrm{T}}$和$\hat{\alpha} = (-179.8, 133.9, -25.52)^{\mathrm{T}}$.

由于MINITAB树数据是完全观察数据,为了进行缺失数据研究,本文采用人工数据缺失的方式,致使响应变量$y_i = \log V_i$并未完全观测到,表2给出人工数据缺失的详细状况,缺失比例约为35.48%. 运用联合均值与方差模型(17)以及如下的logistic回归模型:

$$\mathrm{logit}\{p(r_i = 1 \mid y_i, x_i, \varphi)\} = \log \frac{p(r_i = 1 \mid y_i, x_i, \varphi)}{1 - p(r_i = 1 \mid y_i, x_i, \varphi)} = \varphi_1 + \varphi_2 y_i + \varphi_3 x_{i1} + \varphi_4 x_{i2}, \tag{18}$$

并采用本文第2部分介绍的贝叶斯估计方法对该不完全数据进行拟合,其中模型参数$\beta = (\beta_1, \beta_2, \beta_3)^{\mathrm{T}}$和$\alpha = (\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$为感兴趣参数,缺失机制参数$\varphi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)^{\mathrm{T}}$为讨厌参数. 类似于模拟研究部分,由于未知参数先验分布中的超参数对贝叶斯估计结果的影响并不大,因此本文在先验分布中分别取以下超参数:$\beta_0 = (-6.5, 1.0, 2.0)^{\mathrm{T}}, \alpha_0 = (-180.0, 135.0, -26.0)^{\mathrm{T}}$和$\varphi_0 = (-1.0, 0.3, 0.3, 0.3)^{\mathrm{T}}$,$\sum_\beta = 0.25 I_3, \sum_\alpha = 0.25 I_3$和$\sum_\varphi = 10 I_4$. 并设置参数的三组初始值:

$(\beta^{\mathrm{T}}, \alpha^{\mathrm{T}}, \varphi^{\mathrm{T}})^{\mathrm{T}} = (-6.5, 1.0, 2.0, -180.0, 135.0, -26.0, -1.0, 0.3, 0.3, 0.3)$,

$(\beta^{\mathrm{T}}, \alpha^{\mathrm{T}}, \varphi^{\mathrm{T}})^{\mathrm{T}} = (-6.0, 1.5, 2.5, -178.0, 133.0, -25.0, 1.0, 0.0, 0.0, 0.0)$,

$(\beta^{\mathrm{T}}, \alpha^{\mathrm{T}}, \varphi^{\mathrm{T}})^{\mathrm{T}} = (-6.6, 0.8, 1.9, -182.0, 136.0, -24.0, -1.0, 0.5, 0.5, 0.5)$.

经过简单计算表明,当迭代次数$t = 4\,000$时Gibbs抽样算法已经收敛[19]. 本文遗弃参数前6 000次迭代值,采集6 000次迭代之后的5 000个随机样本来计算缺失数据模型下未知参数的贝叶斯估计,表3给出

最后的计算结果. 表 3 表明在 NMAR 缺失数据模型机制下参数的贝叶斯估计为 $\hat{\beta} = (-6.51, 1.01, 2.01)^T$ 和 $\hat{\alpha} = (-179.96, 130.41, -27.73)^T$, 这和 Aiktin [2] 的估计结果是基本一致的.

表 2　部分响应变量缺失的 MINITAB 树数据

Tab. 2　Minitab tree data missing from some response variables

个体编号	Girth	Height	Volume	个体编号	Girth	Height	Volume
1	8.3	70	10.3*	17	12.9	85	33.8*
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2*	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9*
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7*	22	14.2	80	31.7*
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2*	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4*
11	11.3	79	24.2*	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5*
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1*	31	20.6	87	77.0
16	12.9	74	22.2				

注: 带* 表示人工数据缺失.

表 3　缺失数据模型下 MINITAB 树数据的贝叶斯估计

Tab. 3　Bayesian estimation of Minitab tree data under missing data model

均值模型			方差模型		
参数	估 计	90% 后验置信区间	参数	估计	90% 后验置信区间
β_1	-6.51	(-7.34, -5.67)	α_1	-179.96	(-180.77, -179.10)
β_2	1.01	(0.18, 1.82)	α_2	130.41	(130.39, 130.43)
β_3	2.01	(1.19, 2.85)	α_3	-27.73	(-27.77, -27.69)

4 结 论

本文应用 Gibbs 抽样与 MH 算法讨论非随机缺失机制下联合均值与方差模型的贝叶斯估计问题. 数值结果表明,所提出来的方法能有效地计算模型参数和缺失机制参数的联合贝叶斯估计.

参考文献:

[1] Harvey A C. Estimating Regression Models with Multiplicative Heteroscedasticity [J]. Econometrica, 1976, 44(3): 461-465.

[2] Aitkin M. Modelling Variance Heterogeneity in Normal Regression Using GLIM [J]. Applied Statistics, 1987, 36(3): 332-339.

[3] Verbyla A P. Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics [J]. Journal of the Royal Statistical Society, Series B, 1993, 55(2): 493-508.

[4] 吴刘仓,张忠占,徐登可. 联合均值与方差模型的变量选择 [J]. 系统工程理论与实践, 2012, 32(8): 1754-1760.

[5] Xu D K and Zhang Z Z. A Semiparametric Bayesian approach to joint mean and variance models [J]. Statistics and Probability Letters, 2013, 83(7): 1624-1631.

[6] 赵远英,徐登可,庞一成. 联合均值与方差模型的 Bayes 分析 [J]. 高校应用数学学报, 2018, 33(2): 157-166.

[7] Little R J A, Rubin D B. Statistical Analysis With Missing Data [M]. New York: Wiley, 2002.

[8] Lee S Y, Tang N S. Bayesian analysis of nonlinear structural equation models with nonignoralble missing data [J]. Psychometri-

ka, 2006, 71: 541 – 564.

[9] Zhou Y, Wan A T K, Wang X J. Estimating equations inference with missing data [J]. Journal of the American Statistical Association, 2008, 103: 1187 – 1199.

[10] Zhang J, Cui H J. Empirical likelihood confidence region for parameters in linear EV model with missing data [J]. Acta Mathematica Scientia, 2009, 29A(6): 1465 – 1476.

[11] Tang N S, Zhao P Y. Empirical likelihood – based inference in nonlinear regression models with missing responses at random [J]. Statistics, 2013, 47(6): 1141 – 1159.

[12] Guo X, Niu C Z, Yang Y P, et al. Empirical likelihood for single index model with missing covariates at random [J]. Statistics, 2015, 49(3): 1 – 14.

[13] Geman S, Geman D. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images [J]. IEEE Transactions On Pattern Analysis and Machine Intelligence, 1984, 6(6): 721 – 741.

[14] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. Equations of state calculations by fast computing machine [J]. Journal of Chemical Physics, 1953, 21(6): 1087 – 1091.

[15] Hastings W K. Monte Carlo sampling methods using Markov chains and their applications [J]. Biometrika, 1970, 57(1): 97 – 109.

[16] Ibrahim J G, Chen M H, Lipsitz S R. Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable [J]. Biometrika, 2001, 88(2): 551 – 564.

[17] 唐年胜,韦博成. 非线性再生散度模型 [M]. 北京: 科学出版社,2007.

[18] Geyer C J. Practical Markov Chain Monte Carlo [J]. Statistical Science, 1992, 7(4): 473 – 483.

[19] Gelman A. Inference and monitoring convergence. In Markov chain Monte Carlo in practice [M]. London: Chapman and Hall, 1996.

(上接第124页)

[3] Garey M R, Johnson D S. Crossing number is NP – complete [J]. SIAM Journal on Algebraic Discrete Methods, 1983, 4(3): 312 – 316.

[4] 王晶, 黄元秋. 完全3 – 部图 $K_{1,10,n}$ 的交叉数 [J]. 高校应用数学学报, 2008, 23(3): 000349 – 356.

[5] 周志东, 吕胜祥, 周志东, 等. 关于一个特殊六阶图与路和圈的联图的交叉数 [J]. 数学进展, 2014, 43(1).

[6] Asano K. The crossing number of $K_{1,3,n}$ and $K_{2,3,n}$ [J]. Journal of graph theory, 1986, 10(1): 1 – 8.

[7] 吕胜祥, 黄元秋. $K_{2,4} \times S_n$ 的交叉数 [J]. 系统科学与数学, 2010, 30(7): 929 – 935.

[8] Zarankiewicz K. On a problem of P. Turán concerning graphs [J]. Fundamenta Mathematicae, 1955, 1(41): 137 – 145.

[9] Kleitman D J. The crossing number of $K_{5,n}$ [J]. Journal of Combinatorial Theory, 1970, 9(4): 315 – 323.

[10] Klešč M. The crossing numbers of join of the special graph on six vertices with path and cycle [J]. Discrete Mathematics, 2010, 310(9): 1475 – 1481.

[11] 王淑, 吕胜祥. 一个特殊六阶图与 n 个孤立点联图的交叉数 [J]. 昆明理工大学学报(自然科学版), 2017(2): 122 – 126.

[12] Černy J, Kynčl J, Tóth G. Improvement on the decay of crossing numbers [C]//International Symposium on Graph Drawing. Springer, Berlin, Heidelberg, 2007: 25 – 30.

[13] Salazar G. On a crossing number result of Richter and Thomassen [J]. Journal of Combinatorial Theory, Series B, 2000, 79(1): 98 – 99.

[14] Fox J, Tóth C D. On the decay of crossing numbers [J]. Journal of Combinatorial Theory, Series B, 2008, 98(1): 33 – 42.

[15] Pach J, Tóth G. Thirteen problems on crossing numbers [J]. Geombinatorics, 2000, 9(4): 194 – 207.

[16] Richter R B, Thomassen C. Minimal graphs with crossing number at least k [J]. Journal of Combinatorial Theory, Series B, 1993, 58(2): 217 – 224.

[17] Kleitman D J. A note on the parity of the number of crossings of a graph [J]. Journal of Combinatorial Theory, Series B, 1976, 21(1): 88 – 89.