偏正态数据下半参数混合效应模型的贝叶斯估计

郑丛平,王 涛,谢有余
云南师范大学 数学学院,昆明 650091

摘 要:针对纵向数据服从非正态分布情况下混合效应模型的估计问题,提出偏正态分布半参数混合效应模型的贝叶斯估计方法;假定个体测量误差服从偏正态分布,纵向指标与时间的关系采用 B 样条方法建模,在共轭先验下考虑该模型的贝叶斯分析,基于 MH 算法与 Gibbs 抽样的混合算法获取未知参数、随机效应和非参数函数的贝叶斯估计;数值模拟中,数据非正态分布条件下将偏正态方法得到的估计与传统半参数混合效应模型估计方法进行对比,发现偏正态半参数混合效应模型在有限样本情况下表现更好,说明偏正态半参数混合效应模型与传统模型相比,可以更好地拟合偏态数据,获得更加精准的参数估计;最后将该方法应用于 ADNI 数据中,研究了神经评分与基线临床指标间的关系,得出了合理的结论,证明了方法的合理性。

关键词:偏正态分布;B 样条;混合效应模型;贝叶斯估计;ADNI 数据
中图分类号:O212.8 **文献标识码**:A doi:10.16055/j.issn.1672-058X.2023.0004.013

Bayesian Estimation of Semi-parametric Mixed Effect Model under Skew-normal Data

ZHENG Congping, WANG Tao, XIE Youyu
School of Mathematics, Yunnan Normal University, Kunming 650091, China

Abstract: Aiming at the estimation problem of the mixed effect model when longitudinal data obey non-normal distribution, a Bayesian estimation method of semi-parametric mixed effect model with skew-normal distribution was proposed. Individual measurement error obeys skew-normal distribution, and the relationship between longitudinal index and time was modeled by B spline method. Bayesian analysis of the model was considered under conjugate prior, and Bayesian estimation of unknown parameters, random effects and nonparametric functions was obtained based on the mixed algorithm of MH algorithm and Gibbs sampling. In the numerical simulation, under the condition of non-normal distribution of data, the estimation obtained by the skew-normal method was compared with that of the traditional semi-parametric mixed effect model. It is found that the skew-normal semiparametric mixed effect model performs better under the condition of limited samples, which indicates that the skew-normal semiparametric mixed effect model can better fit the skewed data than the traditional model, and the Bayesian method can effectively use prior information to obtain more accurate parameter estimation. Finally, the modified method was applied to ADNI data, and the relationship between neural score and baseline clinical indicators was studied. A reasonable conclusion was drawn, which proved the rationality of the method.

Keywords: skew-normal distribution; B spline; mixed effect model; Bayesian estimation; ADNI research

收稿日期:2022-06-08 **修回日期**:2022-07-02 **文章编号**:1672-058X(2023)04-0093-06
基金项目:国家自然科学基金(81360449);云南省教育厅科学研究基金项目(2022Y187);云南师范大学研究生科研创新基金(YJSJJ22-B95).
作者简介:郑丛平(1998—),男,江西赣州人,硕士研究生,从事生物与卫生统计研究.
作者简介:王涛(1964—),男,云南昆明人,副教授,从事生物医学统计、教育统计、社会统计研究. Email:wtaokm@263.net.
引用格式:郑丛平,王涛,谢有余. 偏正态数据下半参数混合效应模型的贝叶斯估计[J]. 重庆工商大学学报(自然科学版),2023,40(4):93—98.
ZHENG Congping, WANG Tao, XIE Youyu. Bayesian estimation of semi-parametric mixed effect model under skew-normal data [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(4): 93—98.

1 引 言

混合效应模型(HLM)最早由 Airy 在 1861 年提出,是重要的统计模型,该模型包含固定效应和随机效应,通过固定效应反映总体变化,随机效应反映个体间异质性,在社会学、经济学和生物医学等方面数据分析中有广泛的应用。假设 Y_{ij} 为响应变量,$\boldsymbol{X}_{ij}=(X_{ij1},\cdots,X_{ijp})$,$\boldsymbol{Z}_{ij}=(Z_{ij1},\cdots,Z_{ijq})$ 分别是固定效应 $\boldsymbol{\beta}$ 和随机效应 $\boldsymbol{b}_i$ 的设计向量,则混合效应模型有如下形式:

$$Y_{ij}=\boldsymbol{X}_{ij}\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{t}_{ij})\boldsymbol{b}_i+\varepsilon_{ij} \tag{1}$$

其中,$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$,随机效应向量 $\boldsymbol{b}_i$ 彼此间相互独立,假设服从多元正态分布 $N_q(0,D)$,ε_{ij} 服从正态分布 $N(0,\sigma^2)$。

近几十年来,对混合效应模型的研究已取得较大进展。Angelo 等[1]利用 B 样条对光滑样条模型进行改进,提出一种新的拟合非线性混合效应模型的方法,该方法既能够处理个体间的非同质性,又能够对时间效应进行有效刻画,不仅具有参数模型的可解释性,而且具有非参数模型的灵活性,在实际生活中具有更加广泛的适用性;阙烨等[2]采用广义最小二乘法对未知参数、随机效应和方差分量进行估计,并证明了估计量的渐进性质。最小二乘法计算简便,但不够稳健,对离群点敏感。为了提高估计效率,Lindley 等[3]将贝叶斯方法应用于线性混合模型,贝叶斯方法除了可以利用样本信息外,还可以结合先验信息,从而提高统计推断的效果;Goel[4]将经验贝叶斯方法和混合效应模型融合,研究了协变量的超参数行为;齐培艳等[5]研究了含变点的半参数非线性混合效应模型的多重估算法,该方法相较于朴素贝叶斯方法和两步法具有更加精准的参数估计;付英姿等[6]研究了一类含有不可忽略缺失数据的半参数广义线性混合效应模型,考虑了该模型贝叶斯分析及模型选择问题,提出的方法有更广的适应性。

在纵向数据分析中,由于测量误差的原因,数据不服从正态分布的情况时有发生。针对此类问题,最简单的方法是直接假设测量误差服从正态分布进行估计,这样的假定可以带来计算简便以及良好的统计性质,但出现异常点时,稳健性会被破坏。另一种常用方法是对数据进行变换,使得变换后的数据呈正态或者近似正态分布,如 Log 变换、平方根变换和 Box-Cox 变换等,但转换后的正态性假定仍需考察验证,适用范围有限。为此,学者们对测量误差非正态情况下半参数混合效应模型进行了广泛的研究。Huang 等[7]采用学生 t 分布建模个体内的测量误差,该方法对厚尾分布的数据具有更强的稳健性;Matos 等[8]研究了一类纵向截尾数据,对正态分布进行了修正处理;Sahu 等[9]首次在贝叶斯框架下,对偏正态数据进行回归分析;Lachos 等[10]在随机效应服从偏正态分布的情形下,研究模型参数的极大似然估计;叶仁道[11]研究了偏正态混合效应模型的固定效应和偏度参数的经验贝叶斯估计问题。然而,测量误差服从偏正态分布下利用贝叶斯方法对半参数混合效应模型的研究还未见报道。

本文针对纵向响应变量服从偏正态分布,研究了半参数混合效应模型的贝叶斯估计问题,其中,个体测量误差服从偏正态分布,纵向指标与时间的关系采用 B 样条建模。为结合先验信息,考虑该模型的贝叶斯分析,基于 MH 算法与 Gibbs 抽样的混合算法获取未知参数、随机效应和非参数函数的贝叶斯估计。通过数值模拟证明了研究方法的有效性,实例分析进一步说明方法的合理性。

2 模型建立

2.1 半参数混合效应模型

考虑 n 个不相关的个体,每个个体有 m_i 次观测,采用如下半参数混合效应模型:

$$Y_{ij}=\boldsymbol{X}_{ij}\boldsymbol{\beta}+\boldsymbol{Z}(\boldsymbol{t}_{ij})\boldsymbol{b}_i+f(\boldsymbol{t}_{ij})+\varepsilon_{ij} \tag{2}$$

其中,$i=1,\cdots,n;j=1,\cdots,m_i;\sum_{i=1}^{n}m_i=M$;第 i 个个体在第 j 个观测时间 t_{ij} 观察到的响应变量是 Y_{ij};$f(t_{ij})$ 是一个未知的光滑函数;$\boldsymbol{X}_{ij}$,$\boldsymbol{Z}(\boldsymbol{t}_{ij})$,$\boldsymbol{\beta}$,$\boldsymbol{b}_i$ 的定义与式(1)相同;ε_{ij} 是第 i 个个体在第 j 个观察时间点 t_{ij} 的测量误差,假定它服从偏正态分布 $SN(\mu,\sigma^2,\rho)$,μ 是位置参数,σ^2 是协方差,ρ 是偏度参数。如果所有的分量函数都是线性的,那么模型退化成线性混合效应模型。

2.2 偏正态分布

参考 Sahu[9]提出的偏正态分布 $SN(\mu,\sigma^2,\rho)$,它的概率密度函数为

$$f(\varepsilon;\mu,\sigma^2,\rho)=2\varphi(\varepsilon;\mu,\sigma^2)\Phi\left(\frac{\rho(\varepsilon-\mu)}{\sigma}\right)$$

$\varphi(\cdot)$ 和 $\Phi(\cdot)$ 分别是正态密度函数和正态分布函数,限制 $\mu=\sqrt{2/\pi}\rho$ 可使均值为零,减少待估参数,考虑如下的分层模型:

$$\varepsilon=\rho|X_0|+X_1,X_0\sim N_1(0,1),X_1\sim N_1(\mu,\Pi)$$

其中,X_0 与 X_1 相互独立,基于上述分层模型,纵向数据 Y_{ij} 在给定潜变量 x_{ij} 的条件下,服从正态分布:

$$Y_{ij}|x_{ij},\boldsymbol{b}_i\sim N(\boldsymbol{X}_{ij}\boldsymbol{\beta}+\boldsymbol{Z}_{ij}\boldsymbol{b}_i+f(t_{ij})+(|x_{ij}|-\sqrt{2/\pi})\rho,\ \sigma^2),x_{ij}\sim N(0,1),x_{ij}\perp\boldsymbol{b}_i$$

2.3 B 样条

根据 Rupper 等[12]的建议,采用贝叶斯框架下的 B 样条逼近未知光滑函数 $f(t)$。考虑如下形式的光滑函数:

$$f(t)=\delta_0^{(1)}+\delta_1^{(1)}t+\cdots+\delta_s^{(1)}t^s+\sum_{l=1}^{L}\delta_l^{(2)}(t-\tau_l)_+^s \quad (3)$$

其中,s 是样条自由度,L 是光滑函数的节点数,将光滑函数的定义域划分为 $L+1$ 个回归区间,节点和样条采用如下规则进行选取:τ_l 为第 l 个节点,通常取样本的 $l+1/L+2$ 分位数,并满足 $\tau_0=t_{\min}<\tau_1<\cdots<\tau_l=t_{\max}$,$(t-\tau_l)_+^s=(\max\{0,t-\tau_l\})^s$,令回归系数向量 $\boldsymbol{\delta}=(\boldsymbol{\delta}^{(1)\mathrm{T}},\boldsymbol{\delta}^{(2)\mathrm{T}})^{\mathrm{T}}=(\delta_0^{(1)},\delta_1^{(1)},\cdots,\delta_k^{(1)},\delta_1^{(2)},\cdots,\delta_L^{(2)})^{\mathrm{T}}$ 截断的幂基函数 $\boldsymbol{\varphi}(t)=(\boldsymbol{\varphi}_1(t)^{\mathrm{T}},\boldsymbol{\varphi}_2(t)^{\mathrm{T}})^{\mathrm{T}}=(1,t,\cdots,t^s,(t-\tau_1)_+^s,\cdots,(t-\tau_L)_+^s)^{\mathrm{T}}$,故式(3)可写成

$$f(t)=\boldsymbol{\varphi}_1(t)^{\mathrm{T}}\boldsymbol{\delta}^{(1)}+\boldsymbol{\varphi}_2(t)^{\mathrm{T}}\boldsymbol{\delta}^{(2)}=\boldsymbol{\varphi}(t)^{\mathrm{T}}\boldsymbol{\delta}$$

记 $\boldsymbol{Y}_i=(Y_{i1},\cdots,Y_{im_i})^{\mathrm{T}}$,$\boldsymbol{Y}=(\boldsymbol{Y}_1^{\mathrm{T}},\cdots,\boldsymbol{Y}_n^{\mathrm{T}})^{\mathrm{T}}$,同样定义 $\boldsymbol{X},\boldsymbol{Z},\boldsymbol{x},\boldsymbol{\varepsilon}$。令 $\boldsymbol{Z}=\mathrm{diag}(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n)$ 表示对角元素为 $\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_n$ 的块对角矩阵,则式(2)又可以表示成

$$\boldsymbol{Y}=\boldsymbol{X}_*\boldsymbol{\beta}_*+\boldsymbol{Z}\boldsymbol{b}+\boldsymbol{W}\boldsymbol{a}+\boldsymbol{\varepsilon} \quad (4)$$

其中,$\boldsymbol{X}_*=(\boldsymbol{X},\boldsymbol{\varphi}_1^{\mathrm{T}})$,$\boldsymbol{W}=\boldsymbol{\varphi}_2^{\mathrm{T}}(t_{ij})_{M\times L}$,$\boldsymbol{a}=\boldsymbol{\delta}^{(2)}$,$\boldsymbol{\beta}_*=(\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\delta}^{(1)\mathrm{T}})^{\mathrm{T}}$,$\boldsymbol{\delta}=(\boldsymbol{\delta}^{(1)\mathrm{T}},\boldsymbol{\delta}^{(1)\mathrm{T}})^{\mathrm{T}}\sim N_{s+1+L}(0,\boldsymbol{\zeta})$,其中,$\boldsymbol{\zeta}=\mathrm{diag}(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)$,$\boldsymbol{\zeta}_1=\mathrm{diag}(\sigma_{\zeta_1}^2,\cdots,\sigma_{\zeta_1}^2)$,$\boldsymbol{\zeta}_2=\mathrm{diag}(\sigma_{\zeta_2}^2,\cdots,\sigma_{\zeta_2}^2)$,$\sigma_{\zeta_1}^2$ 是未知参数,$\sigma_{\zeta_2}^2$ 为额外的方差成分。在贝叶斯框架下,样条的个数由对应节点的系数方差所控制,选择合适的节点数能够确保样条的柔韧性。

3　半参数混合效应模型的贝叶斯分析

为了获得 $\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{a},\rho$ 和 σ^2 中所有未知参数的贝叶斯估计并进行贝叶斯推断,需要从 $\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{a},\rho,\sigma^2$ 的联合后验分布中获取样本观测,由式(4)得似然方程的函数表达式如下:

$$\begin{aligned}l(\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{a},\rho,\sigma^2|\boldsymbol{Y})=&(2\pi\sigma^2)^{-\frac{n}{2}}(\exp\{-(2\sigma)^{-2}\times\\&(\boldsymbol{Y}-\boldsymbol{X}_*\boldsymbol{\beta}_*-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{W}\boldsymbol{a}-(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})\rho)^{\mathrm{T}}\times\\&(\boldsymbol{Y}-\boldsymbol{X}_*\boldsymbol{\beta}_*-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{W}\boldsymbol{a}-(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})\rho)\}\end{aligned} \quad (5)$$

考虑 $\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{D},\boldsymbol{a},\boldsymbol{\zeta},\rho,\sigma^2$ 的共轭先验分布如下:

$$p(\boldsymbol{\beta}_*|\boldsymbol{\beta}_0,\boldsymbol{H}_0)\sim N_{p+s}(\boldsymbol{\beta}_0,\boldsymbol{H}_0),p(\boldsymbol{b}_i|\boldsymbol{D})\sim N_q(0,\boldsymbol{D})$$

$$p(\boldsymbol{D}|\alpha_0,R_0)\sim IW_q(\alpha_0,\boldsymbol{R}_0)$$

$$p(\boldsymbol{a}|\boldsymbol{\zeta})\sim N_L(0,\boldsymbol{\zeta}),p(\boldsymbol{\zeta}|r_0,\boldsymbol{U}_0)\sim IW(r_0,\boldsymbol{U}_0)$$

$$p(\rho|\rho_0,h_0)\sim N(\rho_0,h_0),P(\sigma^2)\sim 1/\sigma^2$$

其中,$\boldsymbol{e}$ 是对应单位向量,IW 表示逆威沙特分布,$\boldsymbol{\beta}_0,\boldsymbol{H}_0,\boldsymbol{D},\alpha_0,\boldsymbol{R}_0,\boldsymbol{\zeta},r_0,\boldsymbol{U}_0,\rho_0,h_0$ 是超参数。

3.1　条件分布及算法

3.1.1　固定效应 $\boldsymbol{\beta}_*$ 的条件后验分布

固定效应 $\boldsymbol{\beta}_*$ 的先验分布为 $\pi(\boldsymbol{\beta}_*)\sim N_{p+s}(\boldsymbol{\beta}_0,\boldsymbol{H}_0)$,结合似然方程式(5)和 Leonard[13] 的多元配方方法,$\boldsymbol{\beta}_*$ 的条件后验分布为

$$\begin{aligned}\pi(\boldsymbol{b},\boldsymbol{a},\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\propto&\exp(\boldsymbol{\beta}_*-(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*))^{\mathrm{T}}\times\\&(\frac{1}{\sigma^2}\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{X}_*+\boldsymbol{H}_0^{-1})(\boldsymbol{\beta}_*-(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*))\end{aligned}$$

其中:

$$\begin{aligned}\boldsymbol{\beta}^*=&(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\sigma^2\boldsymbol{H}_0^{-1})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{\beta}_0-(\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{X}_*)^{-1}\times\\&\boldsymbol{X}_*^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{W}\boldsymbol{a}-(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})\rho))\end{aligned}$$

根据核函数可得 $\boldsymbol{\beta}_*$ 的后验分布为

$$\begin{aligned}p(\boldsymbol{\beta}_*|\boldsymbol{b},\boldsymbol{a},\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\sim&N_{p+s}(\boldsymbol{\beta}_0-\boldsymbol{\beta}^*,\\&(\sigma^{-2}+\boldsymbol{X}_*^{\mathrm{T}}\boldsymbol{X}_*+\boldsymbol{H}_0^{-1})^{-1})\end{aligned} \quad (6)$$

3.1.2　随机效应 $\boldsymbol{b}_i$ 的条件后验分布

随机效应 $\boldsymbol{b}_i$ 的先验分布为 $p(\boldsymbol{b}_i|\boldsymbol{D})\sim N_q(0,\boldsymbol{D})$,则 $\boldsymbol{b}_i$ 的条件后验分布:

$$\begin{aligned}\pi(\boldsymbol{b}_i|\boldsymbol{\beta}_{*i},\boldsymbol{D},\boldsymbol{\zeta},\rho,\boldsymbol{x}_i,\sigma^2,\boldsymbol{Y}_i)\propto&\\\exp\{-\frac{1}{2\sigma^2}(\boldsymbol{b}_i-\boldsymbol{b}_i^*)^{\mathrm{T}}(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{Z}_i+\boldsymbol{D}^{-1})(\boldsymbol{b}_i-\boldsymbol{b}_i^*)\}&\end{aligned}$$

其中:$\boldsymbol{b}_i^*=(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{Z}_i+\sigma^2\boldsymbol{D}^{-1})^{-1}(\boldsymbol{Z}_i)^{\mathrm{T}}(\boldsymbol{Y}_i-\boldsymbol{X}_{*i}\boldsymbol{\beta}_{*i}-\boldsymbol{W}_i\boldsymbol{a}_i-(\boldsymbol{x}_i-\sqrt{2/\pi}\boldsymbol{e})\rho)$。

根据核函数,可得随机效应 $\boldsymbol{b}_i$ 的后验分布:

$$\begin{aligned}p(\boldsymbol{b}_i|\boldsymbol{\beta}_*,\boldsymbol{D},\boldsymbol{\zeta},\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\sim&\\N_q(\boldsymbol{b}_i^*,(\sigma^{-2}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{Z}_i+\boldsymbol{D}^{-1})^{-1})&\end{aligned} \quad (7)$$

3.1.3　随机效应 $\boldsymbol{b}_i$ 的方差项 $\boldsymbol{D}$ 的条件后验分布

考虑先验分布为 $p(\boldsymbol{D}|\alpha_0,\boldsymbol{R}_0)\sim IW_q(\alpha_0,\boldsymbol{R}_0)$,则 $\boldsymbol{D}$ 的条件后验分布:

$$\begin{aligned}\pi(\boldsymbol{D}|\boldsymbol{\beta}_*,\boldsymbol{b}_i,\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\propto&|\boldsymbol{D}|^{-(n+\alpha_0+q+1)/2}\times\\&\exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{D}^{-1}(\boldsymbol{R}_0+\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}}))\}\end{aligned}$$

所以 $\boldsymbol{D}$ 的后验分布为

$$p(\boldsymbol{D}|\boldsymbol{\beta}_*,\boldsymbol{b}_i,\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\sim IW_q(n+\alpha_0,\boldsymbol{R}_0+\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}}) \quad (8)$$

3.1.4　随机效应 $\boldsymbol{a}$ 的条件后验分布

随机效应 $\boldsymbol{a}$ 的先验分布为 $p(\boldsymbol{a}|\boldsymbol{\zeta})\sim N_L(0,\boldsymbol{\zeta})$,则 $\boldsymbol{a}$ 的条件后验分布:

$$\begin{aligned}\pi(\boldsymbol{a}|\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{\zeta},\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\propto&\exp\{-\frac{1}{2}(\boldsymbol{a}-\boldsymbol{a}^*)^{\mathrm{T}}\times\\&\frac{1}{\sigma^2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\boldsymbol{\zeta}^{-1})(\boldsymbol{a}-\boldsymbol{a}^*)\}\end{aligned}$$

其中:$\boldsymbol{a}^*=(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\sigma^2\boldsymbol{\zeta}^{-1}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}_*\boldsymbol{\beta}_*-\boldsymbol{Z}\boldsymbol{b}-(\boldsymbol{x}-\sqrt{2/\pi}\boldsymbol{e})\rho))$。

根据核函数可知 $\boldsymbol{a}$ 的后验分布为

$$p(\boldsymbol{a}|\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{\zeta},\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\sim N_L(\boldsymbol{a}^*,\sigma^{-2}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}+\boldsymbol{\zeta}^{-1}) \quad (9)$$

3.1.5　随机效应 $\boldsymbol{a}$ 的方差项 $\boldsymbol{\zeta}$ 的条件后验分布

这里考虑随机效应 $\boldsymbol{a}$ 的方差项 $\boldsymbol{\zeta}$ 的先验分布为 $p(\boldsymbol{\zeta}|r_0,\boldsymbol{U}_0)\sim IW(r_0,\boldsymbol{U}_0)$,则 $\boldsymbol{\zeta}$ 的条件后验分布:

$$\begin{aligned}\pi(\boldsymbol{\zeta}|\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{D},\boldsymbol{a},\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\propto&|\boldsymbol{\zeta}|^{-(L+r_0+1+1)/2}\times\end{aligned}$$

$$\exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\zeta}^{-1}(\boldsymbol{U}_0+\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}))\right\}$$

因此,$\boldsymbol{\zeta}$ 的条件后验分布为

$$p(\boldsymbol{\zeta}|\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{D},\boldsymbol{a},\rho,\boldsymbol{x},\sigma^2,\boldsymbol{Y})\sim IW(L+r_0,\boldsymbol{U}_0+\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}) \tag{10}$$

3.1.6 随机误差偏度参数 ρ 的条件后验分布

假定随机误差偏度参数 ρ 的先验分布 $p(\rho|\rho_0,h_0)\sim N(\rho_0,h_0)$,条件后验分布:

$$p(\rho|\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{D},\boldsymbol{a},\boldsymbol{\zeta},\boldsymbol{x},\sigma^2,\boldsymbol{Y})\sim N(\mu_\rho,\Sigma_\rho) \tag{11}$$

其中:$\Theta_\rho=h_0^{-1}+\frac{1}{\sigma^2}(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})^{\mathrm{T}}(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})$

$$\mu_\rho=\Theta_\rho\left\{\frac{1}{\sigma^2}(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})(\boldsymbol{Y}-\boldsymbol{X}_*\boldsymbol{\beta}_*-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{W}\boldsymbol{a})+h_0^{-1}\rho_0\right\}$$

3.1.7 随机误差方差项 σ^2 的条件后验分布

随机误差方差项 σ^2 的先验分布为 $\pi(\sigma^2)\sim 1/\sigma^2$,则 σ^2 的条件后验分布:

$$p(\sigma^2|\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{a},\rho,\boldsymbol{x},\boldsymbol{Y})\sim \mathrm{IG}\left(\frac{n}{2},\frac{\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}}{2}\right) \tag{12}$$

其中:$\boldsymbol{\omega}=\boldsymbol{Y}-\boldsymbol{X}_*\boldsymbol{\beta}_*-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{W}\boldsymbol{a}-(\boldsymbol{x}-\sqrt{2/\pi}\boldsymbol{e})\rho$。

3.1.8 随机误差潜变量 $\boldsymbol{x}$ 的条件后验分布

考虑先验 $\boldsymbol{x}\sim N_M(0,\boldsymbol{e})$,则有

$$p(\boldsymbol{x}|\boldsymbol{\beta}_*,\boldsymbol{b},\boldsymbol{D},\boldsymbol{a},\rho,\sigma^2,\boldsymbol{Y})\propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{X}_*\boldsymbol{\beta}_*-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{W}\boldsymbol{a}-(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})\rho)^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}_*\boldsymbol{\beta}_*-\boldsymbol{Z}\boldsymbol{b}-\boldsymbol{W}\boldsymbol{a})-(|\boldsymbol{x}|-\sqrt{2/\pi}\boldsymbol{e})\rho-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right\} \tag{13}$$

3.2 MH 算法

式(6)—式(12)为正态、逆威沙特或逆伽马分布,可直接运用 Gibbs 抽样获取样本。随机误差潜变量 $\boldsymbol{x}$ 的条件后验分布式(13)形式复杂,为了获得潜变量 $\boldsymbol{x}$ 的样本,参考 Tang 等[14],采用 MH 抽样算法来实现从它的后验分布中进行抽样。

Step1 迭代 k 步后,x_{ij} 取为 $x_{ij}^{(k)}$。

Step2 从正态分布 $\mathrm{N}_1(x_{ij}^{(k)},\sigma_x^2\rho)$ 中产生潜在转移点 x_{ij},其中 $\rho=\sqrt{1/(\rho^2\sigma^2+1)}$,$\sigma_x^2$ 是方差调节系数,用于调节潜在转移点平均接受率于区间[0.25,0.35]内。

Step3 潜在转移点 x_{ij} 被接受的概率:

$$\min\left\{1,\frac{p(x_{ij}|\boldsymbol{\beta}_*,\boldsymbol{b}_i,\boldsymbol{a},\rho,\sigma^2,Y_{ij})}{p(x_{ij}^{(k)}|\boldsymbol{\beta}_*,\boldsymbol{b}_i,\boldsymbol{a},\rho,\sigma^2,Y_{ij})}\right\}$$

3.3 贝叶斯推断

$\{(\boldsymbol{\beta}_*^{(k)},\boldsymbol{b}_i^{(k)},\boldsymbol{D}^{(k)},\boldsymbol{a}^{(k)},\boldsymbol{\zeta}^{(k)},\rho^{(k)},\sigma^{2(k)}):k=1,\cdots,\lambda\}$ 是通过 Gibbs 抽样以及 MH 算法所产生的 $\{\boldsymbol{\beta}_*,\boldsymbol{b}_i,\boldsymbol{D},\boldsymbol{a},\boldsymbol{\zeta},\rho,\sigma^2\}$ 的第 k 步迭代后的抽样观察值,则 $\{\boldsymbol{\beta}_*,\boldsymbol{b}_i,\boldsymbol{D},\boldsymbol{a},\boldsymbol{\zeta},\rho,\sigma^2\}$ 的贝叶斯估计为

$$\widehat{\boldsymbol{\beta}}_*=\frac{1}{\lambda}\sum_{k=1}^{\lambda}\boldsymbol{\beta}_*^{(k)},\ \widehat{\boldsymbol{b}}_i=\frac{1}{\lambda}\sum_{k=1}^{\lambda}\boldsymbol{b}_i^{(k)},\ \widehat{\boldsymbol{D}}=\frac{1}{\lambda}\sum_{k=1}^{\lambda}\boldsymbol{D}^{(k)}$$

$$\widehat{\boldsymbol{a}}=\frac{1}{\lambda}\sum_{k=1}^{\lambda}\boldsymbol{a}^{(k)},\ \widehat{\boldsymbol{\zeta}}=\frac{1}{\lambda}\sum_{k=1}^{\lambda}\boldsymbol{\zeta}^{(k)},\ \widehat{\rho}=\frac{1}{\lambda}\sum_{k=1}^{\lambda}\rho^{(k)}$$

Geyer[15]指出,上述贝叶斯估计均是其后验均值的相合估计,同理,后验样本的协方差阵亦可作为未知参数协方差阵的贝叶斯估计,如

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_*|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z})=\frac{1}{\lambda-1}\sum_{k=1}^{\lambda}(\boldsymbol{\beta}_*^{(k)}-\widehat{\boldsymbol{\beta}}_*)(\boldsymbol{\beta}_*^{(k)}-\widehat{\boldsymbol{\beta}}_*)^{\mathrm{T}}$$

4 数值模拟

本节采用 MCMC 模拟说明偏正态分布半参数混合效应模型的有效性。基于式(2)产生纵向数据集 $\{y_{ij};i=1,\cdots,n;j=1,\cdots,m_i\}$,第 i 个个体的第 j 个观察时间 $t_{ij}=0.5(j-1)$,协变量 $X_{ij}=(1,X_{ij1},X_{ij2})$ 的两个分量分别由正态分布 $N(0,1)$ 和均匀分布 $U(0,1)$ 独立产生,随机误差 b_{i0} 服从正态分布 $N(0,\delta^2)$。定义真实的非参数函数为 $f(t)=\sin(2\pi t/2.5)$,其中 $t\in[0,2.5]$。误差项 ε_{ij} 从偏正态分布 $SN(\mu,\sigma^2,\rho)$ 中产生。参数的真实值分别定义为 $\boldsymbol{\beta}=(1,0.5,1)^{\mathrm{T}}$,$\delta^2=1$,$\rho=1$,$\mu=-\sqrt{2/\pi}\rho$,$\sigma^2=1$。

模拟 1 先验分布的超参数 $\boldsymbol{\beta}_0$,ρ_0 分别取相应的真实参数,$\boldsymbol{H}_0=0.1\boldsymbol{I}_2$,$h_0=0.1$。模拟计算过程,独立重复试验为 50 次,每次独立重复试验中所有参数都抽取 15 000 个样本,舍弃前 5 000 个样本,采用后 10 000 个样本计算贝叶斯估计值,样本量分别为 $n=30$ 和 $n=300$。"Bias"表示真实值与 50 次独立重复试验贝叶斯估计值平均值的偏差绝对值,"RMS"表示 50 次独立重复试验贝叶斯估计值与真实值的偏差均方根(表 1)。

模拟 2 对比测量误差服从偏正态分布和正态分布的情况,样本量为 $n=30$,除 ε_{ij} 的分布不同外,其他条件与模拟 1 相同。SN 表示偏正态分布,NA 表示正态分布(表 2)。

表 1 模拟 1 中贝叶斯估计结果

Table 1 Bayesian estimation results in simulation 1

参 数	$n=30$		$n=300$	
	Bias	RMS	Bias	RMS
$\boldsymbol{\beta}_0$	0.012	0.132	0.011	0.139
$\boldsymbol{\beta}_1$	0.018	0.141	0.018	0.144
$\boldsymbol{\beta}_2$	0.008	0.043	0.006	0.072
$\boldsymbol{\delta}$	0.035	0.183	0.040	0.056
σ	0.007	0.166	0.008	0.139
ρ	0.047	0.167	0.022	0.118

表 2　模拟 2 中贝叶斯估计结果

Table 2　Bayesian estimation results in simulation 2

参　数	SN		NA	
	Bias	RMS	Bias	RMS
$\boldsymbol{\beta}_0$	0.012	0.132	0.112	0.339
$\boldsymbol{\beta}_1$	0.018	0.141	0.009	0.053
$\boldsymbol{\beta}_2$	0.008	0.043	0.008	0.049
$\boldsymbol{\delta}$	0.035	0.183	0.105	0.159
σ	0.007	0.166	0.072	0.142
ρ	0.047	0.167	—	—

图 1 表示非参数函数 $f(t)$ 的估计函数。其中黑色实线代表真实曲线,虚线代表估计曲线,灰色区域为 95%可信区间。由图 1 可知:用 B 样条建模得到的多项式样条近似非参数函数,除在拐点附近有轻微差异外,整体拟合效果良好。

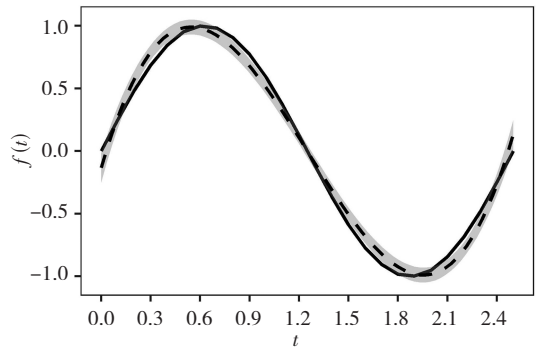

图 1　非参数函数的估计函数

Fig. 1　Estimation function of nonparametric functions

由表 1、表 2 模拟结果可以得到以下结论:

(1) 所有感兴趣参数的 Bias 值都小于 0.05,RMS 值小于 0.2,说明由上述方法得到的贝叶斯估计值比较准确,不易受样本值的影响。

(2) 针对偏正态数据,采用偏正态半参数混合效应模型得到参数的 Bias 值和 RMS 值普遍比正态模型的更小,说明偏正态半参数混合效应模型更稳健,适用范围更广。

5　实例分析

本文分析使用的数据来自阿尔兹海默病神经影像学倡议(ADNI)(http://adni.loni.ucla.edu),其中包括 81 名基线时患有遗忘型轻度认知障碍(MCI,一种介于正常和阿尔兹海默症之间的转换风险状态)的受试者,他们至少进行了一次诊断,最后部分受试者转换为阿尔兹海默病(AD)。评估定于基线、6、12、18 、24 和 36 个月进行,实际诊断时间可能会有所不同。本文取神经心理学评估 MMSE 作为响应变量,确定"g"(性别,取值为 0 或 1,分别代表女性和男性,40.74%为女性),"e"(受教育年限,平均值 16.22,标准差 2.72,范围 8~20),"a"(年龄,平均值 77.35,标准差 7.14,范围 60.84~89.86),"$N_{\mathrm{APOE}\varepsilon 4}$"(载脂蛋白 E$\varepsilon$4 等位基因数量,61.73%的受试者≥1)作为协变量。从图 2 直方图和概率密度曲线可以看出:个体内误差不支持正态假设,是非对称左偏分布。因此,我们采用偏正态半参数混合效应模型进行建模。

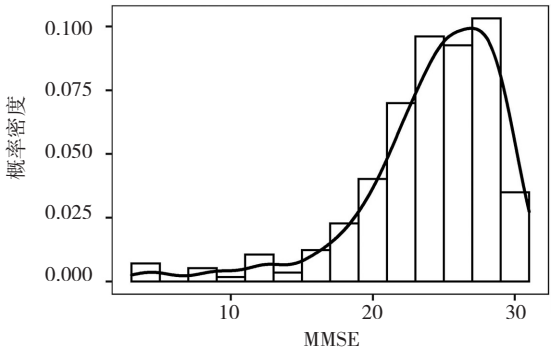

图 2　直方图和概率密度曲线

Fig. 2　Histogram and probability density curve

具体地,建立如下半参数混合效应模型:

$$\mathrm{MMSE}_{ij}=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 g_{ij}+\boldsymbol{\beta}_2 e_{ij}+\boldsymbol{\beta}_3 a_{ij}+\boldsymbol{\beta}_4 N_{\mathrm{APOE}\varepsilon 4\,ij}+f(t_{ij})+b_{i0}+\varepsilon_{ij}$$

其中,随机误差 ε_{ij} 服从偏正态分布 $SN(\mu,\sigma^2,\rho)$,随机截距 b_{i0} 服从正态分布 $N(0,\delta^2)$,$f(t)$ 是一个光滑函数。将数据代入模型,通过贝叶斯分析和 MCMC 方法得出模型的参数估计,具体结果见表 3:

表 3　贝叶斯估计结果

Table 3　Bayesian estimation results

参　数	Est	SE	95%CI
$\boldsymbol{\beta}_0$	22.99	6.43	(10.64,35.40)
$\boldsymbol{\beta}_1$	2.38	1.08	(0.21,4.54)
$\boldsymbol{\beta}_2$	0.13	0.21	(−0.28,0.54)
$\boldsymbol{\beta}_3$	−0.01	0.06	(−0.13,0.12)
$\boldsymbol{\beta}_4$	−1.27	0.84	(−2.88,0.37)
$\boldsymbol{\delta}$	4.67	0.43	(3.90,5.57)
σ	2.77	0.14	(2.51,3.07)
ρ	−2.72	0.10	(−2.92,−2.53)

表 3 给出了所有未知参数的贝叶斯估计、标准误以及 95%置信区间。由表 3 可以看出:偏度参数 $\rho=-2.72$ 很好地刻画了数据的偏离程度;$\hat{\beta}_1=2.38$ 表明"g"对 MMSE 具有一个较强的正向影响,$\hat{\beta}_2=0.13$ 表明"e"具有一个弱的正面影响,$\hat{\beta}_3=-0.01$ 表明"a"具有一

个弱的负面效应,$\hat{\beta}_4$ =-1. 27 表明协变量“$N_{\text{APOE}\varepsilon 4}$”对 MMSE 具有较强的负面影响,携带 APOEε4 等位基因个体更易出现认知障碍,患 AD 的风险更高。

6 结束语

本文研究了半参数混合效应模型,假设个体测量误差服从偏正态分布,未知光滑函数采用 B 样条逼近,在共轭先验下考虑该模型的贝叶斯分析,基于 MH 算法与 Gibbs 抽样混合算法获取未知参数、随机效应和非参数函数的贝叶斯估计。数值模拟中,响应变量服从偏正态分布,对误差偏正态和正态假设下的半参数混合效应模型结果进行对比,发现偏正态半参数混合效应模型在有限样本情况下表现更好,证明了方法的有效性。最后将该方法应用于 ADNI 数据中,研究神经评分与基线临床指标间的关系,结果与实际情况一致,证明了方法的合理性。

本文考虑的模型是混合效应模型,可以将模型推广至变系数模型、空间自回归模型等;针对误差非正态问题,除本文研究的偏正态分布,还可以考虑偏 t 分布、离散分布等情况;在实际研究中,不仅要关注 MMSE 的影响因素,也要关注患者由 MCI 向 AD 转化的事件时间,因此,对纵向过程和事件时间的联合建模也是本文后续进一步的研究方向。

参考文献(References):

[1] ANGELO E, RATCLIFFE S P, SAMUEL P, et al. A b-spline based semiparametric nonlinear mixed effects model [J]. Journal of Computational and Graphical Statistics, 2011, 20(2): 492—509.

[2] 阙烨, 黄振生. 部分线性混合效应模型的有效估计[J]. 应用概率统计, 2017, 33(5): 529—537.
QUE Ye, HUNAG Zhen-sheng. Efficient estimation for the partially linear models with random effects[J]. Chinese Journal of Applied Probability and Statistics, 2017, 33(5): 529—537.

[3] LINDLEY D V, SMITH A F M. Bayes estimates for the linear model[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(1): 1—41.

[4] GOEL P K. Information measures and Bayesian hierarchical models[J]. Journal of the American Statistical Association, 1983, 78(382): 408—410.

[5] 齐培艳, 华文杰, 段西发. 半参数非线性混合效应模型的多重估算法[J]. 统计与决策, 2021, 37(24): 24—28.
QI Pei-yan, HUA Wen-jie, DUAN Xi-fa. Multi-revaluation algorithm for semiparametric nonlinear mixed effect model[J]. Statistics & Decision, 2021, 37(24): 24—28.

[6] 付英姿, 陈异, 戴琳. 半参数广义线性混合效应模型的贝叶斯分析[J]. 统计与决策, 2014(8): 19—23.
FU Ying-zi, CHEN Yi, DAI Lin. Semi-parametric generalized linear mixing effect Bayesian analysis of model[J]. Statistics and Decision, 2014(8): 19—23.

[7] HUANG X, Li G, ELASHOFF R M. A joint model of longitudinal and competing risks survival data with heterogeneous random effects and outlying longitudinal measurements[J]. Statistics and Its interface, 2010, 3(2): 185—195.

[8] MATOS L A, CASTRO L M, LACHOS V H. Censored mixed-effects models for irregularly observed repeated measures with applications to HIV viral loads[J]. TEST, 2016, 25(4): 627—653.

[9] SAHU S K, DEY D K, Branco M D. A new class of multivariate skew distributions with applications to Bayesian regression models[J]. The Canadian Journal of Statistics, 2003, 31(2): 129—150.

[10] LACHOS V H, GHOSH P, ARELLANO-VALLE R B. Likelihood based inference for skew-normal independent liner mixed-models[J]. Statistica Sinica, 2010, 20(1): 303—322.

[11] 叶仁道, 张瑜. 偏正态混合效应模型参数的经验贝叶斯估计[J]. 系统科学与数学, 2019, 39 (11): 1895—1908.
YE Ren-dao, ZHANG Yu. Empirical Bayesian estimation of parameters of skew-normal mixed effect model[J]. Systems Science and Mathematics, 2019, 39 (11): 1895—1908.

[12] RUPPERT D, WAND M P, CARROLL R J. Semiparametric regression[M]. New York, USA: Cambridge University Press, 2003.

[13] LEONARD T, HSU J S J. Bayesian methods[M]. New York, USA: Cambridge University Press, 1999.

[14] TANG A M, TANG N S, ZHU H T. Influence analysis for skew-normal semiparametric joint models of multivariate longitudinal and multivariate survival data[J]. Statistics in Medicine, 2017, 36(9): 1476—1490.

[15] GEYER C J. Practical Markov chain Monte Carlo [J]. Statistical Science, 1992, 7(4): 473—483.

责任编辑:李翠薇