应 用 数 学
MATHEMATICA APPLICATA
2020, 33(3): 747-756

基于偏Laplace正态数据下位置、均值回归模型的参数估计

郑桂芬, 吴刘仓, 聂兴锋
(昆明理工大学理学院, 云南 昆明 650093)

摘要: 参数估计是一种基本的统计推断形式, 也是统计学的一个重要分支. 在分析偏态数据时, 我们比较关注数据的众数、中位数和均值, 但是偏Laplace正态数据的众数和中位数难以精确求出, 因此用位置参数来近似代替. 故本文提出偏Laplace正态数据下位置和均值回归模型, 并研究该模型的参数估计, 模拟和实例研究结果表明本文提出的模型和方法是科学合理的.

关键词: 偏Laplace正态分布; 位置回归模型; 均值回归模型; EM算法
中图分类号: O212.1 **AMS(2000)主题分类:** 62J05; 62G05
文献标识码: A **文章编号:** 1001-9847(2020)03-0747-10

1. 引言

在金融、医学和社会经济领域中, 存在大量偏斜且厚尾的数据. 如果用正态分布、偏正态和偏t分布去拟合, 得到的信息不精确, 利用偏Laplace正态分布对数据进行拟合获得的信息更具有准确性和可靠性, 因此研究偏Laplace正态分布具有重要意义.

在过去几十年里, 学者们提出很多方法分析处理偏态数据. Azzalini[1]提出了偏态指数幂分布同时处理偏态和重尾两种情况; Monti[2]对偏态指数幂分布性质和推断进行了研究; WU等[3]利用联合惩罚似然方法对偏正态分布下联合位置与尺度模型提出了一种可行有效的变量选择方法; 吴刘仓等[4]研究了偏正态数据下联合位置与尺度混合专家回归模型的参数估计; 马婷等[5], 吴刘仓等[6]分别基于SN, StN分布下研究了联合位置、尺度与偏度模型的极大似然估计. 偏正态分布的概率密度由差的平方进行刻画, 为了能使估计的结果更加具有稳健性, 把偏正态分布进行扩展, 从而引入偏Laplace正态分布, 其概率密度用差的绝对值来表示. 因此, 分布的尾部比正态分布更加平坦. 由于偏Laplace正态分布受异常点数据的影响不大、得到的结果比较稳健, 吸引了很多学者的研究兴趣. Dogru和Arslan[7]在偏Laplace正态分布下研究了混合回归模型的参数估计. Garay等[8]研究了偏正态分布混合尺度的非线性回归模型的统计诊断. 张舒宇等[9]研究了基于Laplace分布下混合联合位置与尺度模型的参数估计.

综上所述, 虽然偏Laplace正态分布的回归模型已经有很多研究成果, 但在偏Laplace正态分布下对位置和均值回归模型建模的涉及较少, 考虑到位置和均值建模的重要性, 本文详细介绍了利用EM算法对这两个模型的参数进行极大似然估计, 并通过实例结果表明本文所提出来的模型和方法的实用性和可行性.

本文结构安排如下: 第二部分给出了偏Laplace正态分布的一些性质; 第三部分给出了偏Laplace正态分布下位置和均值回归模型; 第四部分利用EM算法对位置和均值回归模型的

* **收稿日期:** 2019-08-22
基金项目: 国家自然科学基金项目 (11861041, 11261025)
通讯作者: 吴刘仓, 男, 汉族, 云南人, 教授, 研究方向: 应用统计.

参数进行极大似然估计; 第五部分通过Monte Carlo随机模拟实验证实了本文提出方法的有效性; 最后, 实例研究结果表明, 本文所提出的模型和方法是科学合理的.

2. 偏Laplace正态分布

对于服从偏Laplace正态分布的随机变量Y可以表示为$Y\sim \mathrm{SLN}(\mu,\sigma^2,\lambda)$, 其中$\mu$为位置参数, σ^2为尺度参数, λ为偏度参数. 则其概率密度函数可表示为

$$f(y)=2f_L(y;\mu,\sigma)\Phi(\lambda\frac{y-\mu}{\sigma}), \tag{2.1}$$

其中Φ为标准正态分布的分布函数, $f_L(y;\mu,\sigma)$为Laplace分布的概率密度函数, 且

$$f_L(y;\mu,\sigma)=\frac{1}{2\sigma}\exp(-\frac{|y-\mu|}{\sigma}).$$

$$E(y)=\frac{\sigma(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma}{2\pi\lambda^2}+\mu. \tag{2.2}$$

Ⅰ 偏Laplace正态分布下的随机表示

设$Z\sim \mathrm{SN}(0,1,\lambda)$, V的概率密度函数为$f_V(v)=v^{-3}\exp(-(2v^2)^{-1})$, $v>0$是两个独立随机变量, 随机变量$Y\sim \mathrm{SLN}(\mu,\sigma^2,\lambda)$表达式为

$$Y=\mu+\sigma\frac{Z}{V},$$

然后, 利用文[1]中的201页和文[10]中的定理1, 分布随机变量Z的随机表示得随机变量Y的以下随机表示

$$Y=\mu+\sigma(\frac{\lambda|Z_1|}{\sqrt{V^2(V^2+\lambda^2)}}+\frac{Z_2}{\sqrt{V^2+\lambda^2}}),$$

其中$Z_1\sim N(0,1)$, $Z_2\sim N(0,1)$是独立随机变量, 从而得到偏Laplace正态分布的层次表示

$$\begin{aligned} &Y|u,v\sim N(\mu+\frac{\sigma\lambda u}{v^2+\lambda^2},\frac{\sigma^2}{v^2+\lambda^2}),\\ &U|v\sim \mathrm{TN}((0,\frac{v^2+\lambda^2}{v^2}),(0,\infty)),\\ &V\sim f_V(v)=v^{-3}\exp(-(2v^2)^{-1}), \end{aligned} \tag{2.3}$$

这里$U=\sqrt{V^{-2}(V^2+\lambda^2)}|Z_1|$, $\mathrm{TN}(\cdot)$为截尾正态分布, 用(2.3)中的层次表示法可以得到如下的条件期望

$$\mathrm{E}(V^2|y)=\frac{\sigma}{|y-\mu|}, \tag{2.4}$$

$$\mathrm{E}(U|y)=\lambda s+\frac{\phi(\lambda s)}{\Phi(\lambda s)}, \tag{2.5}$$

$$\mathrm{E}(U^2|y)=1+\lambda s\mathrm{E}(U|y), \tag{2.6}$$

其中$s=\frac{y-\mu}{\sigma}$.

3. 位置和均值回归模型

Ⅰ 位置回归模型

$$\begin{cases} y_i\sim \mathrm{SLN}(\mu_i,\sigma^2,\lambda),\\ \mu_i=x_i^{\mathrm{T}}\beta,\\ i=1,2,\cdots,n. \end{cases} \tag{3.1}$$

由概率密度函数式(2.1)及位置参数回归模型(3.1)可以得到

$$f(y_i;\beta,\sigma^2,\lambda)=\frac{1}{\sigma}\exp(-\frac{|y_i-x_i^{\mathrm{T}}\beta|}{\sigma})\Phi(k_{i1}), \tag{3.2}$$

其中$k_{i1}=\lambda\frac{(y_i-x_i^{\mathrm{T}}\beta)}{\sigma}$.

Ⅱ 均值回归模型

$$\begin{cases} y_i\sim \mathrm{SLN}(\mu_i,\sigma^2,\lambda),\\ \mathrm{E}\,(y_i)=x_i^{\mathrm{T}}\alpha,\\ i=1,2,\cdots,n. \end{cases} \tag{3.3}$$

从式(2.2)有$\mathrm{E}(y_i)=\frac{\sigma(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma}{2\pi\lambda^2}+\mu_i=x_i^{\mathrm{T}}\alpha$, 从而有$\mu_i=x_i^{\mathrm{T}}\alpha-\frac{\sigma(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma}{2\pi\lambda^2}$, 故由式(2.1)及式(3.3)可以得到相应的密度函数

$$f(y_i;\alpha,\sigma^2,\lambda)=\frac{\Phi(k_{i2})}{\sigma}\exp(-\frac{|y_i-x_i^{\mathrm{T}}\alpha+\frac{\sigma(4\lambda^2-1)\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma}{2\pi\lambda^2}|}{\sigma}), \tag{3.4}$$

其中

$$k_{i2}=\frac{\lambda y_i-\lambda x_i^{\mathrm{T}}\alpha}{\sigma}+\frac{(4\lambda^2-1)\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]}{2\lambda}+\frac{\sqrt{2\pi}}{\pi}.$$

这里y_i为第i个响应变量, 服从位置参数为μ_i, 尺度参数为σ^2, 偏度参数为λ的偏Laplace正态分布, $x_i=(x_{i1},\cdots,x_{ip})^{\mathrm{T}}$是解释变量, $\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$是维数为$p\times 1$的位置回归模型的未知参数, $\alpha=(\alpha_1,\cdots,\alpha_p)^{\mathrm{T}}$是维数为$p\times 1$的均值回归模型的未知参数.

本文采用极大似然估计的方法对参数进行估计, 但因有潜变量存在, 所以直接估计参数比较困难. EM算法作为解决潜变量问题参数估计的有效方法, 因此接下来本文介绍所提出模型参数的极大似然估计的EM算法.

4. 极大似然估计的EM算法

Ⅰ 位置回归模型下极大似然估计的EM算法

由式(3.2)可得似然函数为:

$$l(y_i;\beta,\sigma^2,\lambda)=\prod_{i=1}^{n}[\frac{1}{\sigma}\exp(-\frac{|y_i-x_i^{\mathrm{T}}\beta|}{\sigma})\Phi(k_{i1})]. \tag{4.1}$$

令$\theta_1=(\beta^{\mathrm{T}},\sigma^2,\lambda)^{\mathrm{T}}$, 则$L(\beta,\sigma^2,\lambda)=L(\theta_1)$.

两边取自然对数, 得到对数似然函数为:

$$\begin{aligned} L(\theta_1;y,u,v)&=\sum_{i=1}^{n}[-\frac{1}{2}\log\sigma^2-\frac{|y_i-x_i^{\mathrm{T}}\beta|}{\sigma}+\log\Phi(k_{i1})]\\ &=-\frac{n}{2}\log\sigma^2-\sum_{i=1}^{n}\frac{|y_i-x_i^{\mathrm{T}}\beta|}{\sigma}+\sum_{i=1}^{n}\log\Phi(k_{i1}). \end{aligned} \tag{4.2}$$

设U和V为潜变量, 利用位置回归模型与偏Laplace正态分布的层次表示法得

$$Y_i|u_i,v_i\sim N(x_i^{\mathrm{T}}\beta+\frac{\sigma\lambda u_i}{v_i^2+\lambda^2},\frac{\sigma^2}{v_i^2+\lambda^2}),\qquad U_i|v_i\sim \mathrm{TN}((0,\frac{v_i^2+\lambda^2}{v_i^2}),(0,\infty)),$$
$$V_i\sim f_V(v_i)=v_i^{-3}\exp(-(2v_i^2)^{-1}).$$

设$u=(u_1,\cdots,u_n)$, $v=(v_1,\cdots,v_n)$为缺失数据, (y,u,v)为完全数据. 用层次表示法得完全数据下对数似然函数为

$$\begin{aligned} L_c(\theta_1;y,u,v)=&-n\log\pi-\frac{n}{2}\log\sigma^2-\sum_{i=1}^{n}[2\log v_i+(2v_i^2)^{-1}]\\ &-\frac{1}{2}\sum_{i=1}^{n}[\frac{v_i^2(y_i-x_i^{\mathrm{T}}\beta)^2}{\sigma^2}+u_i^2-\frac{2u_i\lambda(y_i-x_i^{\mathrm{T}}\beta)}{\sigma}+\frac{\lambda^2(y_i-x_i^{\mathrm{T}}\beta)^2}{\sigma^2}]. \end{aligned} \tag{4.3}$$

为了得到θ_1的极大似然估计, 本文把(4.3)式极大化. 然而, 这种极大化得到的估计依赖于潜变量. 因此, 为了处理潜变量, 必须由给出的观测数据y_i求出完全数据下对数似然函数的条件期望. 设$\theta_1^{(t)}$为极大似然估计$\hat{\theta}_1$的第t次近似, 从而有

$$\begin{aligned}\mathrm{E}(L_c(\theta_1;y,u,v)|y_i,\theta_1^{(t)}) =& -n\log\pi-\frac{n}{2}\log\sigma^2-\sum_{i=1}^{n}[2\mathrm{E}(\log V_i|y_i,\theta_1^{(t)})+\mathrm{E}((2v_i^2)^{-1})|y_i,\theta_1^{(t)})]\\ &-\sum_{i=1}^{n}\{\frac{1}{2}[(\frac{\mathrm{E}(V_i^2|y_i,\theta_1^{(t)})(y_i-x_i^{\mathrm{T}}\beta)^2}{\sigma^2}+\mathrm{E}(U_i^2|y_i,\theta_1^{(t)})]\}\\ &+\sum_{i=1}^{n}[\frac{\lambda(y_i-x_i^{\mathrm{T}}\beta)}{\sigma}\mathrm{E}(U_i|y_i,\theta_1^{(t)})-\frac{\lambda^2(y_i-x_i^{\mathrm{T}}\beta)^2}{2\sigma^2}]. \end{aligned}\tag{4.4}$$

利用(2.4)-(2.6)式给出的条件期望, 可以计算出式(4.4)的条件期望$\mathrm{E}(V_i^2|y_i,\theta_1^{(t)})$, $\mathrm{E}(U_i|y_i,\theta_1^{(t)})$和$\mathrm{E}(U_i^2|y_i,\theta_1^{(t)})$, 故

$$v_l^{(t)}=\mathrm{E}(V_i^2|y_i,\theta_1^{(t)})=\frac{\sigma^{(t)}}{|y_i-x_i^{\mathrm{T}}\beta^{(t)}|},\tag{4.5}$$

$$u_{1i}^{(t)}=\mathrm{E}(U_i|y_i,\theta_1^{(t)})=k_{i1}^{(t)}+\frac{\phi(k_{i1}^{(t)})}{\Phi(k_{i1}^{(t)})},\tag{4.6}$$

$$u_{2i}^{(t)}=\mathrm{E}(U_i^2|y_i,\theta_1^{(t)})=1+k_{i1}^{(t)}u_{1i}^{(t)},\tag{4.7}$$

其中$k_{i1}^{(t)}=\lambda^{(t)}\frac{(y_i-x_i^{\mathrm{T}}\beta^{(t)})}{\sigma^{(t)}}$. 即使用式(4.4)中的条件期望, 从而得到关于θ_1的目标函数的最大值.

$$\begin{aligned}Q(\theta_1|\theta_1^{(t)}) =& -n\log\pi-\frac{n}{2}\log\sigma^2-\sum_{i=1}^{n}\Big\{\frac{1}{2}\Big[\frac{v_l^{(t)}(y_i-x_i^{\mathrm{T}}\beta)^2}{\sigma^2}+u_{2i}^{(t)}\\ &-\frac{2u_{1i}^{(t)}\lambda(y_i-x_i^{\mathrm{T}}\beta)}{\sigma}+\frac{\lambda^2(y_i-x_i^{\mathrm{T}}\beta)^2}{\sigma^2}\Big]\Big\}. \end{aligned}\tag{4.8}$$

EM算法(Expectation Maximization Algorithm)是一种迭代算法, 其具体流程分为两个步骤进行: E-step是根据参数初始值或上一次迭代所得结果来计算对数似然函数的期望值; M-step是将对数似然函数最大化以获得新的参数值, 用新得到的参数值代替初始值或上一次迭代所得结果使得对数似然函数最大化. 重复执行以上两步骤, 直至收敛. 下面给出EM算法在偏Laplace正态数据下位置回归模型的参数估计中的计算步骤:

E-step: 给定观测数和当前参数值, 求出(4.3)式中给出的完全数据似然函数的条件期望, 即计算(4.5)-(4.7)式中的条件期望.

M-step: 把条件期望带入$Q(\theta_1|\theta_1^{(t)})$中求出$\theta_1$的估计值. 再将对数似然函数极大化并得到一个新的参数值, 直到第$t+1$次. 最后重复E步和M步直到收敛. 利用式(4.8)给出的目标函数, 得到位置参数的Score函数为

$$\begin{aligned}S(\theta_1)=&\frac{\partial Q(\theta_1|\theta_1^{(t)})}{\partial\theta_1}=(S_1{}^{\mathrm{T}}(\beta),S_2(\sigma^2),S_3(\lambda))^{\mathrm{T}},\\ S_1(\beta)=&\sum_{i=1}^{n}\frac{(y_i-x_i^{\mathrm{T}}\beta)x_i{}^{\mathrm{T}}}{\sigma^2}(v_l^{(t)}+\lambda^2)-\sum_{i=1}^{n}\frac{\lambda x_i^{\mathrm{T}}u_{1i}^{(t)}}{\sigma},\\ S_2(\sigma^2)=&\sum_{i=1}^{n}\frac{(y_i-x_i^{\mathrm{T}}\beta)^2}{2\sigma^4}(v_l^{(t)}+\lambda^2)-\sum_{i=1}^{n}\frac{\lambda u_{1i}^{(t)}(y_i-x_i^{\mathrm{T}}\beta)}{2\sigma^3}-\frac{n}{2\sigma^2},\\ S_3(\lambda)=&\sum_{i=1}^{n}\frac{(y_i-x_i^{\mathrm{T}}\beta)u_{1i}^{(t)}}{\sigma}-\sum_{i=1}^{n}\frac{(y_i-x_i^{\mathrm{T}}\beta)^2\lambda}{\sigma^2}.\end{aligned}$$

设计如下迭代:

$$\theta_1^{(t+1)}=\theta_1^{(t)}+(-H(\theta_1^{(t)}))^{-1}S(\theta_1^{(t)}),$$

其中

$$S(\theta_1)=\frac{\partial Q(\theta_1|\theta_1^{(t)})}{\partial\theta_1},H(\theta_1)=\frac{\partial^2Q(\theta_1|\theta_1^{(t)})}{\partial\theta_1\partial{\theta_1}^{\mathrm{T}}},\frac{\partial^2Q(\theta_1|\theta_1^{(t)})}{\partial\beta\partial\beta^{\mathrm{T}}}=-\sum_{i=1}^{n}\frac{x_ix_i^{\mathrm{T}}(\lambda^2+v_l^{(t)})}{\sigma^2},$$

$$\frac{\partial^2Q(\theta_1|\theta_1^{(t)})}{\partial(\sigma^2)^2}=-\sum_{i=1}^{n}\frac{(y_i-x_i^{\mathrm{T}}\beta)^2}{\sigma^6}(\lambda^2+v_l^{(t)})+\sum_{i=1}^{n}\frac{3(y_i-x_i^{\mathrm{T}}\beta)\lambda u_{1i}^{(t)}}{4\sigma^5}+\frac{n}{2\sigma^4},$$

$$\frac{\partial^2Q(\theta_1|\theta_1^{(t)})}{\partial\lambda^2}=-\sum_{i=1}^{n}\frac{(y_i-x_i^{\mathrm{T}}\beta)^2}{\sigma^2}.$$

Ⅱ 均值回归模型下极大似然估计的EM算法

由式(3.4)可得似然函数为

$$l(y_i;\beta,\sigma^2,\lambda)$$
$$=\prod_{i=1}^{n}\{\frac{\Phi(k_{i2})}{\sigma}\exp(\frac{-|2\pi\lambda^2(y_i-x_i^{\mathrm{T}}\alpha)+\sigma(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma|}{2\pi\lambda^2\sigma})\}. \tag{4.9}$$

令$\theta_2=(\alpha^{\mathrm{T}},\sigma^2,\lambda)^{\mathrm{T}}$, 则$L(\alpha,\sigma^2,\lambda)=L(\theta_2)$, 两边取自然对数, 得到对数似然函数为

$$L(\theta_2;y,u,v)=\sum_{i=1}^{n}\{\frac{-|2\pi\lambda^2(y_i-x_i^{\mathrm{T}}\alpha)+\sigma(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma|}{2\pi\lambda^2\sigma}\}$$
$$+\sum_{i=1}^{n}[\log\Phi(k_{i2})-\log\sigma]. \tag{4.10}$$

设U和V为潜变量, 利用均值回归模型与偏Laplace正态分布的层次表示法得

$$Y_i|u_i,v_i\sim N(x_i^{\mathrm{T}}\alpha-\frac{\sigma(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma}{2\pi\lambda^2}+\frac{\sigma\lambda u_i}{v_i^2+\lambda^2},\frac{\sigma^2}{v_i^2+\lambda^2}),$$

$$U_i|v_i\sim\mathrm{TN}((0,\frac{v_i^2+\lambda^2}{v_i^2}),(0,\infty)),V_i\sim f_V(v_i)=v_i^{-3}\exp(-(2v_i^2)^{-1}).$$

设$u=(u_1,\cdots,u_n)$, $v=(v_1,\cdots,v_n)$为缺失数据, (y,u,v)为完全数据. 然后用层次表示法得完全数据下对数似然函数为

$$L_c(\theta_2;y,u,v)=-n\log\pi-\frac{n}{2}\log\sigma^2-\sum_{i=1}^{n}[2\log v_i+(2v_i^2)^{-1}]$$
$$-\frac{1}{2}\sum_{i=1}^{n}\frac{v_i^2A^2}{\sigma^2}-\frac{1}{2}\sum_{i=1}^{n}(u_i^2-\frac{2\lambda u_iA}{\sigma})-\frac{1}{2}\sum_{i=1}^{n}\frac{\lambda^2A^2}{\sigma^2}, \tag{4.11}$$

其中

$$A=y_i-x_i^{\mathrm{T}}\alpha+\frac{\sigma(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda\sigma}{2\pi\lambda^2}.$$

同理, 为了得到θ_2的极大似然估计, 必须把(4.11)式极大化, 从而得到完全数据下对数似然函数的条件期望, 设$\theta_2^{(t)}$为极大似然估计$\hat{\theta}_2$的第t次近似, 从而有

$$\mathrm{E}(L_c(\theta_2;y,u,v)|y_i,\theta_2^{(t)})=-n\log\pi-\frac{n}{2}\log\sigma^2-\sum_{i=1}^{n}[2\mathrm{E}(\log V_i|y_i,\theta_2^{(t)})+\mathrm{E}((2v_i^2)^{-1})|y_i,\theta_2^{(t)})]$$
$$-\frac{1}{2}\sum_{i=1}^{n}\frac{\mathrm{E}(V_i^2|y_i,\theta_2^{(t)})A^2}{\sigma^2}-\frac{1}{2}\sum_{i=1}^{n}[\mathrm{E}(U_i^2|y_i,\theta_2^{(t)})-\frac{2\lambda\mathrm{E}(U_i|y_i,\theta_2^{(t)})A}{\sigma}]$$

$$-\frac{\lambda^2}{2\sigma^2}\sum_{i=1}^{n}A^2, \tag{4.12}$$

其中

$$v_p^{(t)}=\mathrm{E}(V_i^2|y_i,\theta_2^{(t)})=\frac{2\pi\sigma^{(t)}\lambda^{2(t)}}{|2\pi\lambda^{2(t)}y_i-2\pi\lambda^{2(t)}\alpha^{2(t)}x_i^{\mathrm{T}}+m^{(t)}+2\sqrt{2\pi}\lambda^{(t)}\sigma^{(t)}|}, \tag{4.13}$$

$$u_{3i}^{(t)}=\mathrm{E}(U_i|y_i,\theta_2^{(t)})=k_{i2}^{(t)}+\frac{\phi(k_{i2}^{(t)})}{\Phi(k_{i2}^{(t)})}, \tag{4.14}$$

$$u_{4i}^{(t)}=\mathrm{E}(U_i^2|y_i,\theta_2^{(t)})=1+k_{i2}^{(t)}u_{3i}^{(t)}, \tag{4.15}$$

其中

$$k_{i2}^{(t)}=\lambda^{(t)}\frac{y_i-x_i^{\mathrm{T}}\alpha^{(t)}}{\sigma^{(t)}}+\frac{(4\lambda^{2(t)}-1)\exp(\frac{1}{8\lambda^{2(t)}})[1-2\Phi(\frac{1}{2\lambda^{(t)}})+\mathrm{sign}(\lambda^{(t)})]}{2\lambda^{(t)}}+\frac{\sqrt{2\pi}}{\pi},$$

$$m^{(t)}=\sigma^{2(t)}(4\lambda^{2(t)}-1)\exp(\frac{1}{8\lambda^{2(t)}})[1-2\Phi(\frac{1}{2\lambda^{(t)}})+\mathrm{sign}(\lambda^{(t)})].$$

通过式(4.12)中的条件期望, 得到了关于θ_2的目标函数的最大值

$$Q(\theta_2|\theta_2^{(t)})=-n\log\pi-\frac{n}{2}\log\sigma^2-\frac{1}{2}\sum_{i=1}^{n}(u_{4i}^{(t)}-\frac{2\lambda Au_{3i}^{(t)}}{\sigma})-\frac{1}{2}\sum_{i=1}^{n}\frac{\lambda^2+v_p^{(t)}}{\sigma^2}A^2. \tag{4.16}$$

下面给出EM算法在偏Laplace正态数据下均值回归模型的参数估计中的计算步骤：

E-step: 给定观测数据和当前参数值, 求出式(4.11)中给出的完全数据似然函数的条件期望, 即计算(4.13)-(4.15)中的条件期望.

M-step: 把条件期望带入$Q(\theta_2|\theta_2^{(t)})$中求出$\theta_2$的估计值. 再将对数似然函数极大化并得到一个新的参数值, 直到第$t+1$次. 最后重复E步和M步直到收敛. 利用式(4.16)给出的目标函数得到均值参数的Score函数为

$$S(\theta_2)=\frac{\partial Q(\theta_2|\theta_2^{(t)})}{\partial\theta_2}=(S_4{}^{\mathrm{T}}(\alpha),S_5(\sigma^2),S_6(\lambda))^{\mathrm{T}},$$

$$S_4(\alpha)=\sum_{i=1}^{n}\frac{v_p^{(t)}+\lambda^2}{\sigma^2}x_i^{\mathrm{T}}A-\sum_{i=1}^{n}\frac{\lambda u_{3i}^{(t)}x_i^{\mathrm{T}}}{\sigma},$$

$$S_5(\sigma^2)=\sum_{i=1}^{n}\frac{v_p^{(t)}+\lambda^2}{2\sigma^4}A^2-\sum_{i=1}^{n}\frac{v_p^{(t)}+\lambda^2}{4\pi\lambda^2\sigma^3}Ah-\sum_{i=1}^{n}\frac{\lambda u_{3i}^{(t)}}{2\sigma^3}A+\sum_{i=1}^{n}\frac{\lambda u_{3i}^{(t)}}{4\pi\lambda\sigma^2}h-\frac{n}{2\sigma^2},$$

$$S_6(\lambda)=-\sum_{i=1}^{n}\frac{\lambda}{\sigma^2}A^2-\sum_{i=1}^{n}\frac{v_p^{(t)}+\lambda^2}{\sigma^2}Ab+\sum_{i=1}^{n}\frac{u_{3i}^{(t)}}{\sigma}A+\sum_{i=1}^{n}\frac{\lambda u_{3i}^{(t)}}{\sigma}b.$$

设计如下迭代:

$$\theta_2^{(t+1)}=\theta_2^{(t)}+(-H(\theta_2^{(t)}))^{-1}S(\theta_2^{(t)}),$$

其中

$$S(\theta_2)=\frac{\partial Q(\theta_2|\theta_2^{(t)})}{\partial\theta_2},H(\theta_2)=\frac{\partial^2Q(\theta_2|\theta_2^{(t)})}{\partial\theta_2\partial\theta_2{}^{\mathrm{T}}},\frac{\partial^2Q(\theta_2|\theta_2^{(t)})}{\partial\alpha\partial\alpha^{\mathrm{T}}}=-\sum_{i=1}^{n}\frac{v_p^{(t)}+\lambda^2}{\sigma^2}x_ix_i^{\mathrm{T}}$$

$$\frac{\partial^2Q(\theta_2|\theta_2^{(t)})}{\partial(\sigma^2)^2}=\sum_{i=1}^{n}[\frac{5(v_p^{(t)}+\lambda^2)Ah}{8\pi\lambda^2\sigma^5}+\frac{3\lambda u_{3i}^{(t)}A}{4\sigma^5}-\frac{v_p^{(t)}+\lambda^2}{16\pi^2\lambda^4\sigma^4}h^2-\frac{3u_{3i}^{(t)}h}{8\pi\lambda\sigma^4}-\frac{v_p^{(t)}+\lambda^2}{\sigma^6}A^2]+\frac{n}{2\sigma^4},$$

$$\frac{\partial^2Q(\theta_2|\theta_2^{(t)})}{\partial\lambda^2}=\sum_{i=1}^{n}(\frac{2u_{3i}^{(t)}b}{\sigma}+\frac{\lambda u_{3i}^{(t)}e}{\sigma}-\frac{v_p^{(t)}+\lambda^2}{\sigma^2}Ae-\frac{v_p^{(t)}+\lambda^2}{\sigma^2}b^2-\frac{4\lambda Ab}{\sigma^2}-\frac{A^2}{\sigma^2}).$$

其中

$$h=(4\lambda^2-1)\pi\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]+2\sqrt{2\pi}\lambda.$$

$$b=(4\lambda^2+1)\frac{\sigma\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]}{8\lambda^2}+\frac{\sigma\exp(\frac{1}{8\lambda^2})\phi(\frac{1}{2\lambda})}{2\lambda^4}(4\lambda^2-1)-\frac{\sqrt{2\pi}\sigma}{\pi\lambda^2}.$$

$$\begin{aligned}e=&(-1.5-\frac{3}{4\lambda^2}-\frac{1}{32\lambda^4})\frac{\sigma\exp(\frac{1}{8\lambda^2})[1-2\Phi(\frac{1}{2\lambda})+\mathrm{sign}(\lambda)]}{\lambda^4}\\&+\frac{\sigma\exp(\frac{1}{8\lambda^2})\phi(\frac{1}{2\lambda})}{\lambda^3}(-4+\frac{2}{\lambda^2}+\frac{1}{4\lambda^4})+\frac{\sigma(0.5-\frac{1}{8\lambda^2})}{\sqrt{2\pi}\lambda^5}+\frac{2\sqrt{2\pi\sigma}}{\pi\lambda^3},\end{aligned}$$

其中, $-H(\theta_1^{(t)})$与$-H(\theta_2^{(t)})$是观测信息阵, $S(\theta_1^{(t)})$和$S(\theta_2^{(t)})$是Score函数. E-step和M-step反复迭代, 直至收敛.

5. Monte Carlo模拟

Ⅰ 位置回归模型参数估计的Monte Carlo模拟

为评价位置回归模型参数估计方法的有效性, 本文对有限样本进行模拟研究, 参数估计的精确度使用均方误差(MSE)来评价和衡量, 其定义如下:

$$\mathrm{MSE}(\hat{\beta})=\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\beta}_i-\beta_0)^{\mathrm{T}}(\hat{\beta}_i-\beta_0),$$

$$\mathrm{MSE}(\hat{\sigma^2})=\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\sigma_i}^2-{\sigma_0}^2)^2,\quad \mathrm{MSE}(\hat{\lambda})=\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\lambda}_i-\lambda_0)^2,$$

其中$\beta_0,\sigma_0^2,\lambda_0$是参数$\beta,\sigma^2,\lambda$的真值. 根据模型(3.1), 考虑偏Laplace正态数据下位置参数的线性模型.

$$\begin{cases}y_i\sim\mathrm{SLN}(\mu_i,\sigma^2,\lambda),\\ \mu_i=x_i^{\mathrm{T}}\beta,\\ i=1,2,\cdots,n.\end{cases}\tag{5.1}$$

根据模型(5.1)产生模拟数据, 其中$x_i\sim U(-1,1)$. $y_i(i=1,2,\cdots,n)$是根据偏Laplace正态分布产生的响应变量, 且y_i服从偏Laplace正态分布, y_i的产生过程如下:

1) 样本U来自均匀分布$(0,1)$并设$V=\sqrt{-\frac{1}{2\log U}}$;

2) 样本Z_1和Z_2独立于标准正态分布$N(0,1)$;

3) 用适当的参数值给出偏Laplace正态分布样本.

$$Y_i=x_i^{\mathrm{T}}\beta+\sigma[\frac{\lambda|Z_1|}{\sqrt{V^2(V^2+\lambda^2)}}+\frac{Z_2}{\sqrt{V^2+\lambda^2}}].$$

Ⅱ 均值回归模型参数估计的Monte Carlo模拟

为评价均值回归模型参数估计方法的有效性, 参数估计的精确度使用均方误差(MSE)来评价和衡量, 其定义如下:

$$\mathrm{MSE}(\hat{\alpha})=\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\alpha}_i-\alpha_0)^{\mathrm{T}}(\hat{\alpha}_i-\alpha_0),$$

$$\mathrm{MSE}(\hat{\sigma^2})=\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\sigma_i}^2-{\sigma_0}^2)^2,\quad \mathrm{MSE}(\hat{\lambda})=\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\lambda}_i-\lambda_0)^2,$$

其中$\alpha_0, \sigma_0^2, \lambda_0$是参数$\alpha, \sigma^2, \lambda$的真值. 根据模型(3.3), 考虑偏Laplace正态数据下均值参数的线性模型.

$$\begin{cases} y_i \sim \mathrm{SLN}(\mu_i, \sigma^2, \lambda), \\ \mathrm{E}\,(y_i) = x_i^{\mathrm{T}}\alpha, \\ i = 1, 2, \cdots, n. \end{cases} \tag{5.2}$$

根据模型(5.2)产生模拟数据, 其中$x_i \sim U(-1, 1)$. $y_i (i = 1, 2, \cdots, n)$是根据偏Laplace正态分布产生的响应变量, 且y_i服从偏Laplace正态分布, y_i的产生过程如下:

1) 样本U来自均匀分布$(0, 1)$并设$V = \sqrt{-\frac{1}{2\log U}}$;

2) 样本Z_1和Z_2独立于标准正态分布$N(0, 1)$;

3) 用适当的参数值给出偏Laplace正态分布样本.

$$Y_i = x_i^{\mathrm{T}}\alpha - A + \sigma[\frac{\lambda|Z_1|}{\sqrt{V^2(V^2+\lambda^2)}} + \frac{Z_2}{\sqrt{V^2+\lambda^2}}].$$

为了进行模拟研究, 我们分别给了如下的真实参数值

$$\beta_0 = (1,1,1)^{\mathrm{T}}, \sigma_0^2 = 2, \lambda_0 = 0.5; \beta_0 = (1,1,1)^{\mathrm{T}}, \sigma_0^2 = 2, \lambda_0 = 0; \beta_0 = (1,1,1)^{\mathrm{T}}, \sigma_0^2 = 2, \lambda_0 = -0.5.$$

$$\alpha_0 = (1,2,3)^{\mathrm{T}}, \sigma_0^2 = 1, \lambda_0 = 0.5; \alpha_0 = (1,2,3)^{\mathrm{T}}, \sigma_0^2 = 1, \lambda_0 = -0.5.$$

均取样本量$n = 50, 100, 150, 200$, 重复模拟1000次. 模拟结果见表1、表2.

表 1　位置回归模型的参数估计模拟结果

λ	n	$\hat{\boldsymbol{\beta}}^{\mathrm{T}}$	$\hat{\sigma}^2$	$\hat{\lambda}$	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\sigma}^2$)	MSE($\hat{\lambda}$)
	50	(0.9874, 0.9939, 0.9941)	1.9980	0.5049	0.0738	0.0000	0.0038
0.5	100	(0.9999, 1.0002, 0.9970)	1.9991	0.5042	0.0249	0.0000	0.0017
	150	(1.0008, 0.9991, 0.9974)	1.9994	0.5015	0.0140	0.0000	0.0011
	200	(0.9997, 0.9996, 1.0003)	1.9996	0.5024	0.0084	0.0000	0.0009
	50	(0.9894, 0.9974, 0.9901)	1.9981	0.0027	0.0844	0.0000	0.0075
0	100	(0.9996, 0.9966, 1.0025)	1.9991	-0.0051	0.0287	0.0000	0.0033
	150	(1.0002, 0.9976, 0.9991)	1.9994	0.0020	0.0146	0.0000	0.0023
	200	(0.9997, 0.9981, 0.9999)	1.9996	0.0025	0.0083	0.0000	0.0017
	50	(1.0010, 0.9985, 1.0041)	1.9979	-0.5062	0.0768	0.0000	0.0041
-0.5	100	(0.9984, 0.9952, 0.9962)	1.9991	-0.5015	0.0249	0.0000	0.0018
	150	(1.0012, 0.9991, 0.9971)	1.9994	-0.5008	0.0129	0.0000	0.0011
	200	(0.9973, 1.0035, 1.0004)	1.9996	-0.5001	0.0083	0.0000	0.0008

表 2　均值回归模型的参数估计模拟结果(由于λ在分母上, 故$\lambda \neq 0$)

λ	n	$\hat{\boldsymbol{\alpha}}^{\mathrm{T}}$	$\hat{\sigma}^2$	$\hat{\lambda}$	MSE($\hat{\boldsymbol{\alpha}}$)	MSE($\hat{\sigma}^2$)	MSE($\hat{\lambda}$)
	50	(1.0026, 1.9936, 2.9957)	0.9975	0.4831	0.0409	0.0026	0.0027
0.5	100	(0.9971, 1.9987, 3.0030)	1.0011	0.4911	0.0157	0.0009	0.0009
	150	(0.9992, 2.0002, 2.9996)	1.0000	0.4952	0.0093	0.0005	0.0006
	200	(0.9993, 1.9978, 2.9980)	1.0000	0.4968	0.0061	0.0003	0.0004
	50	(1.0017, 2.0020, 2.9987)	1.0000	-0.4809	0.0436	0.0024	0.0024
-0.5	100	(1.0006, 2.0011, 2.9963)	0.9996	-0.4911	0.0144	0.0008	0.0010
	150	(0.9999, 1.9977, 3.0014)	0.9995	-0.4929	0.0092	0.0005	0.0006
	200	(0.9995, 2.0012, 3.0011)	1.0000	-0.4958	0.0061	0.0003	0.0004

从表1和表2可以得到, 随着样本量n的增大, 所有数的估计值越来越接近真值, 而且估计的均方误差(MSE)也越来越小. 以上结论表明, 本文提出的偏Laplace正态数据下位置和均值回归模型及所使用的EM算法对参数的极大似然估计取得了较理想的效果.

6. 实例分析

近年来, 随着人们收入的增长和生活水平的提高, 观看电影逐渐成为人们消遣娱乐的一种方式, 故电影行业发展迅速, 下面利用本文提出的偏Laplace正态分布的位置和均值回归模型及其方法, 对电影票房数据进行参数估计. 本文对收集到的各类电影总票房和首映票房进行统计分析, 该数据中包含一个响应变量Y-总票房和一个解释变量X-首映票房, 计算可得电影总票房的偏度系数$\hat{\lambda} = 2.2100$, 结果表明是右偏的, 直方图如图1所示.

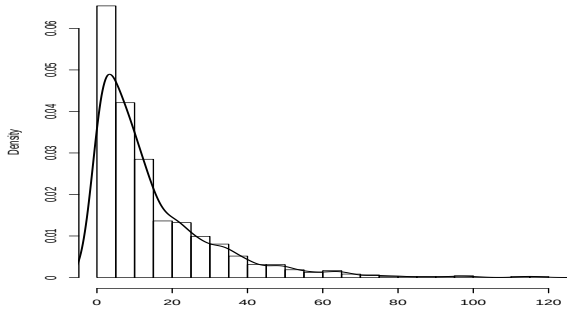

图1　电影票房数据分布直方图

图1和偏度系数说明电影票房数据近似的服从偏Laplace正态分布, 所以可以利用该数据对偏Laplace正态分布的位置和均值回归模型做参数估计, 考虑Y与X之间的模型如下:

$$\begin{cases} y_i \sim \mathrm{SLN}(\mu_i, \sigma^2, \lambda), \\ \mu_i = \beta_0 + x_i\beta_1, \\ \mathrm{E}\,(y_i) = \alpha_0 + x_i\alpha_1, \\ i = 1, 2, \cdots, 969. \end{cases} \tag{6.1}$$

利用第四部分提出的参数估计方法, 得到下表3.

表 3　电影票房数据的位置、均值回归模型参数估计结果

位置回归模型	β_0	β_1	σ^2	λ
	1.0002	-0.6420	1.0002	2.2100
均值回归模型	α_0	α_1	σ^2	λ
	1.0001	2.0235	1.0000	2.2100

由于在同一组电影票房数据中, 尺度和偏度参数是一样的. 从表中可以看出在两个模型中的σ^2和λ大体相同, 但β和α在模型中代表了不同的位置, 所以存在较大差异, 与实际相符合, 表明我们提出的模型和方法是科学合理的.

7. 结论

本文利用EM算法研究了偏Laplace正态数据下位置和均值回归模型的参数估计. 从Monte Carlo模拟结果来看, 本文提出的EM算法对位置和均值回归模型的参数估计取得了较好的效果, 并且在实例分析中, 对电影票房实际数据的应用研究也表明了本文提出的模型和方法是科学合理的.

参考文献:

[1] AZZALINI A. Further results on a class of distributions which includes the normal ones[J]. Statistica, 1986, 46(2): 199-208.

[2] MONTI D C C. Inferential aspects of the skew exponential power distribution[J]. Journal of the American Statistical Association, 2004, 99(466): 439-450.

[3] WU L C , ZHANG Z Z, XU D K. Variable selection in joint location and scale models of the skew-normal distribution[J]. Journal of Statistical Computation and Simulation, 2013, 83(7): 1266-1278.

[4] 吴刘仓, 杨松琴, 戴琳. 基于偏正态数据下联合位置与尺度混合专家回归模型的参数估计[J]. 高校应用数学学报, 2018, 33(1): 36-44.

[5] 马婷, 吴刘仓, 黄丽. 基于偏正态分布联合位置、尺度与偏度模型的极大似然估计[J]. 数理统计与管理, 2013, 32(3): 433-439.

[6] 吴刘仓, 马婷, 詹金龙. 基于StN分布联合位置, 尺度与偏度模型的极大似然估计[J]. 高校应用数学学报, 2013, 28(4): 431-438.

[7] DOGRU F Z, ARSLAN O. Parameter estimation for mixtures of skew Laplace normal distributions and application in mixture regression modeling[J]. Communications in Statistics-Theory and Methods, 2017, 46(21): 10879-10896.

[8] GARAY A M, LACHOS V H, LABRA F V, et al. Statistical diagnostics for nonlinear regression models based on scale mixtures of skew-normal distributions[J]. Journal of Statistical Computation and Simulation, 2014, 84(8): 1761-1778.

[9] 张舒宇, 吴刘仓, 詹金龙. 基于Laplace分布下混合联合位置与尺度模型的参数估计[J]. 应用概率统计, 2017, 33(5): 487-496.

[10] DOGRU, FATMA ZEHRA, ARSLAN O. Joint modelling of the location, scale and skewness parameters of the skew Laplace normal distribution[J]. Iranian Journal of Science and Technology, Transactions A: Science, 2019, 43(3): 1249-1257.

Parameter Estimation of Location and Mean Regression Model Based on Skew Laplace Normal Data

ZHENG Guifen, WU Liucang, NIE Xingfeng

(*Faculty of Science, Kunming University of Science and Technology, Kunming 650093, China*)

Abstract: Parameter estimation is a fundamental form of statistical inference and an important branch of statistics. In the analysis of skewed data, we pay more attention to the mode, median and mean of the data, but the mode and median of the Skew Laplace normal distribution are difficult to accurately calculate, so it is approximate to substitute location parameters for them. Therefore, the location and mean regression model under Skew Laplace normal distribution is proposed, and the parameter estimation of the model is studied. The results of simulation and case study show that the proposed model and method are scientific and reasonable.

Key words: Skew Laplace normal distribution; Location regression model; Mean regression model; EM algorithm