第26卷 第3期 气 象 科 学 Vol. 26, No. 3
2006年6月 SCIENTIA METEOROLOGICA SINICA June, 2006

用最小绝对偏差方法(LAD)估计极值分布参数的探讨

吴香华[1] 秦伟良[2] 王新蕾[1] 俞书平[1]
(1 南京信息工程大学数学系,南京 210044)(2 东南大学经管学院,南京 210096)

摘 要 数理统计问题中的最小绝对偏差方法(Least Absolute Deviation)由于具有良好的稳健性,近年来备受重视。本文所研究的是如何将最小绝对偏差法及非线性回归模型相结合,应用于极值分布的参数估计,并与经典的参数估计法相比较。通过对徐州市降雨量数据的研究表明,改进后的参数估计法不仅提高了模型拟和的精确度,而且有良好的稳定性,可以推广到相关气象要素的预测、预报研究中。其中,将LAD法运用于极值分布模型是一个新的尝试。

关键词 极值分布 LAD法 年最大日降雨量 Kolmogorov检验
分类号 P413 **文献标识码** A

引 言

最小绝对偏差方法最早由Roger Joseph Boscovich 于1957年提出,20世纪80年代以来,经过Huber,Basset等[1]的长期研究,理论方面得到了很大的发展,LAD对于大样本数据研究的稳健性得到了共识。但是,应用方面的研究却不是很多。

本文重点探讨用最小绝对偏差方法估计极值分布模型的参数,并应用于徐州的降雨量数据。以往这些气象数据方面的研究也不少:如金龙等[7]运用基于预报误差最小的回归模型处理了上海月平均温度序列,张敏等[5]运用经验正交函数分析南海夏季风的爆发,杨扬等[4]研究了1948~2001年全球陆地9~11月降水的变化趋势,并且都获得了令人满意的结论。然而,传统的方法往往重视短期预报,相比较而言,对于长期预报(预报百年一遇的特大自然灾害等)的研究较少。极值理论作为长期预报的有效研究工具,解决好其参数估计问题,具有非常重要的理论意义和实际意义。

1 最小绝对偏差方法的基本原理

要引入最小绝对偏差方法,就要先了解一下回归模型

$$y=f(x,b_0,b_1,\cdots,b_p)+\varepsilon \tag{1}$$

其中x为自变量,y为因变量,$b_0,b_1,\cdots,b_p$为待定参数,ε为误差项。如果$f(x,b_0,b_1,\cdots,b_p)$是关于参数$b_0,b_1,\cdots,b_p$的线性函数,(1)式表示线性回归模型;如果是非线性函数,(1)式表示非线性回归模型。问题通常是:当x和y有一组观察组$(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\cdots,(x^{(n)},y^{(n)})$,且真实值与理论值的误差分别为$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$时,如何估计出未知参数向量$\beta$的值,记$\beta=(b_0,b_1,\cdots,b_p)^T$。当$\{\varepsilon_i\}$的分布满足Gauss-Markov假设

Gauss-Markov假设

$$(a)E(\varepsilon_i)=0;(b)Var(\varepsilon_i)=\sigma^2;(c)Cov(\varepsilon_i,\varepsilon_j)=0,(i\neq j) \tag{2}$$

收稿日期:2005-04-11;**修改稿日期**:2005-06-07
基金项目:本文由国家自然科学基金项目(编号:70301010)资助
第一作者简介:吴香华(1980-),女(汉),江苏南通,南京信息工程大学硕士生在读,主要从事气象数据统计研究
E-mail:wuxianghua@nuist.edu.cm

采用经典的最小二乘法(LSE)可以获得较满意的效果,但如果出现异常点或重尾分布(Gauss－Markov 假设不成立),用 LSE 获得的结果就令人难以接受了。

为解决这个问题,Roger 提出了最小绝对偏差估计。此时,估计模型(1)的参数向量 β 时,关键是考察误差绝对值 $r^{(i)}=|y^{(i)}-f(x^i,\beta)|$ $(i=1,\cdots,n)$;然后选择合适的 β^*,当 $\beta=\beta^*$ 时,使误差绝对值总和最小,即

$$min\sum_{i=1}^{n}r_i=min\sum_{i=1}^{n}|y^{(i)}-f(x^i,\beta)| \tag{3}$$

理论已经证明:不满足 Gauss-Markov 假设时,LAD 的统计性能优于 LSE,且稳健性也更优,特别适合厚尾分布。以上方法的缺点是(3)中出现了绝对值,具有不可微性,导致所有需要导数的优化方法都不可行,所以一般无法明确写出解 β^* 的表达式。但是,可以通过运筹学问题中解无约束最优化问题的方法间接估计出 β^*(参见[9]),如松弛算法、目标规划法、搜索算法等。也可以运用泰勒展式先化成线性回归模型,然后再用 LAD 法来估计参数。

2　极值理论及其参数估计

2.1　极值分布

极值理论是在研究最值随机变量的性质时提出的。若 $X_1,\cdots,X_n$ 是独立同分布的随机变量,共同的分布函数为 F,令 $M_n=\max\{X_1,\cdots,X_n\}$,以 G(x)记 M_n 的分布函数,表示 M_n 取小于 x 的值的可能性大小。极值分布函数 G(x)具有 Gumbel 分布、Fuechet 分布、Weibull 分布三种形式,都能够统一于广义极值模型

$$G(x)=exp\{-[1+\zeta(\frac{x-\mu}{\sigma})]^{-1/\zeta}\} \tag{4}$$

其中 μ 为位置参数,σ(>0)为尺度参数,ξ(一般在－0.5～0.5 范围内)为形状参数。研究广义极值分布模型后,根据形状参数 ξ 估计值的符号来确定具体是 Gumbel 分布(ξ→0),Frechet 分布(ξ>0),还是 Weibull 分布(ξ<0)。

2.2　参数估计

极值分布模型常用的参数估计方法有极大似然法、矩法、可靠性矩法、概率权值法、回归法和百分点法等。极大似然估计法(MLE)是最基本的方法,它使参数估计结果在总体上反映样本的统计信息,具有很好的统计性质,因而应用最为普遍。MLE 的主要思想是根据极值分布函数引入对数似然函数

$$l(\mu,\sigma,\zeta)=-n\log\sigma-(1+1/\zeta)\sum_{i=1}^{n}\log[1+\zeta(\frac{x_i-\mu}{\sigma})]-\sum_{i=1}^{n}[1+\zeta(\frac{x_i-\mu}{\sigma})]^{-1/\zeta}\} \tag{5}$$

然后求解函数 l(μ,σ,ξ)的驻点,得到参数 μ,σ,ξ 的估计值。但是这个过程中需要数值求解非线性方程组,运算量较大,应用起来不方便。Jenkinson(1969), Hosking(1985), Macleod(1989)等给出了具体的算法和步骤。矩法和改进后的可靠性矩法、概率权值法利用子样的数字特征"矩"来估计母体数字特征,计算简单,却对随机变量的样本信息利用不充分。百分点法通过求解独立的非线性方程来得到参数的估计。回归法是在给定参数初值后,采用线性或非线性回归得到分布参数的估计,本文着重于运用 LAD 法取代经典的最小二乘法,探讨广义极值分布模型回归方法下参数估计的改进。

3　用 LAD 法估计极值分布参数的算法实现

(1)确定样本容量 n,子样 $M_1,\cdots,M_n$,获得子样的观察值 $x_1,\cdots,x_n$。

(2)将 n 个子样的观察值按降序排列成 $x^{(1)},\cdots,x^{(n)}$,序列号依次为 $1,2,\cdots,n$。

(3)计算 $x^{(1)},\cdots,x^{(n)}$ 对应的概率 $p_i'=\frac{i}{n+1}$,$(i=1,2,\cdots,n)$。

(4)计算经验概率 $p_i=1-p'_i$,$(i=1,2,\cdots,n)$,作为广义极值模型公式中分布函数 G(x)对应的概率值,记 $y^{(i)}=p_i$。

(5)确定观察值 $x^{(i)},\cdots,x^{(n)}$ 及其对应概率的二维数组:$(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\cdots,(x^{(n)},y^{(n)})$。

(6)用 MLE 法估计出参数值 μ_0,σ_0 和 ξ_0。

(7)确定非线性回归模型 $y=\exp\{-[1+\zeta(\frac{x-\mu}{\sigma})]^{-1/\zeta}\}+\varepsilon$。

(8)选取目标函数 $\sum_{i=1}^{n}|y^{(i)}-f(x^i,\beta)|$,将 μ_0,σ_0 和 ζ_0 作为参数初始值,采用 LAD 法通过解非线性规划的方法估计出参数 $\hat{\mu}$,$\hat{\sigma}$ 和 $\hat{\xi}$。

(9)对所求出的模型作检验,分析其拟和程度的优劣。

(10)参数计算结束。

4 模型的检验

从理论上严格证明极值分布模型统计推断中的有效性是比较困难的,一般很难对模型的选用原则做理论证明。但是,根据具体的样本观测值能够从一定程度上来检验模型的合理性和拟和效果。

首先,我们可以采用 P-P 图(Probability Plot)作初步检验,基本思想是将经验分布函数和理论分布函数作对比。若理论分布函数为 $\hat{G}(x)=\exp\{-[1+\hat{\zeta}(\frac{x-\hat{\mu}}{\hat{\sigma}})]^{-1/\hat{\zeta}}\}$,点集$\{(y^{(i)},\hat{G}(x^{(i)}),i=1,0,n\}$应该分布在对角线附近。P-P 图检验的优点是比较简单和易操作;缺陷是不能作精确的检验。

其次,用 Kolmogorov 检验能够进一步分析理论分布和实际分布的一致性,设 $H:X_1,\cdots,X_n$ 的公共分布为 G。具体定理为:若 X 为一维且其分布函数 G 在 R^1 内处处连续,则

$$\lim_{n\to\infty}P(\sqrt{n}\,\|G_n-G\|<x)\equiv Q(x)=1+2\sum_{i=1}^{\infty}(-1)^k\exp(-2i^1x^2) \tag{6}$$

其中 $x\in R^1$,$\|G_n-G\|=\sup_x|G(x)-G_n(x)|$。利用这个定理,取定显著性水平 α,找 c_α,使 $Q(c_\alpha)=1-\alpha$,然后在 $\|G_n-G\|$ 大于临界值 $c_\alpha/\sqrt{n}$时否定假设 H,否则接受假设 H。

5 资料说明

本文所用的徐州市降水量(单位:mm)数据覆盖的时间段是从 1960 年 1 月到 2000 年 12 月,选取其中的年最大日降雨量值,共 41 个数据。样本的数字特征如下:

(1)均值:$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=105.985$

(2)标准差:$S=\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}=49.462$

(3)斜度:$g_1=\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{S})^3=2.325$

(4)峰度:$g_2=\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{S})^4\}-\frac{3(n-1)^2}{(n-2)(n-3)}=7.557$

从以上数据特征可以看出斜度和峰度非零,表现出了厚尾特点,由上文所述,此时不能选用最小二乘法估计回归模型的参数,故考虑用 LAD 法进行回归。

6 不同参数估计方法的比较

先采用极大似然估计法求出三个参数值作为回归估计的初值,然后分别采用 LSE 和 LAD 回归得到参数的估计值,见表 1。具体如下:

(1)与 MLE 估计的初值相比,LSE 回归方法和 LAD 回归方法使 μ 值分别改变了 −0.653、0.795,σ 值改变了 −3.809、−5.05,ξ 值改变了 0.2069、0.1594。μ 和 σ 的变化率分别不大于 1%和 17.7%,ξ 的变化率虽然比较大,但是始终满足了 −0.5~0.5 这个一般的范围。

(2)图 1 可见,LSE 回归方法和 LAD 回归方法都明显提高了拟合效果,其中 LAD 回归法拟合更好。

(3)图 2 是广义极值模型三种参数估计的 P-P 图,除了(a)之外,(b)和(c)都具有明显的线性趋势,拟和

的效果都比较理想。在概率值小于 0.5 的部分,改进后的 LAD 回归方法要明显优于 LSE 回归方法。

(4)在 MLE、LSE、LAD 方法下,模型的总误差分别为 7.899、1.050、0.858,初步可以推断新方法的稳定性较好;Kolmogorov 检验中,$\| G_n - G \|$ 的值分别为 0.3613、0.0769、0.0750,由表 2,除了 MLE,都满足不等式 $\| G_n - G \| < c_a/\sqrt{n}$,故接受假设 H,说明模型拟和的显著性明显。

表 1　参数估计的结果

Table 1　The estimates of parameters

	μ	σ	ξ
极大似然估计法	84.064	28.489	0.1603
回归法(LSE)	83.411	24.680	0.3672
回归法(LAD)	84.859	23.439	0.3197

表 2　Kolmogorov 检验

Table 2　Kolmogorov test

a	c_a	n	$c_a/\sqrt{n}$
0.05	1.358	41	0.212
0.01	1.628	41	0.254

7　总结与讨论

本文运用 LAD 回归法对广义极值分布模型中的参数进行估计,发现得到的模型具有优于 MLE 和 LSE 回归法,表明这一方法具有一定的适用性。对徐州市年最大日降雨量数据的分析有以下结论:

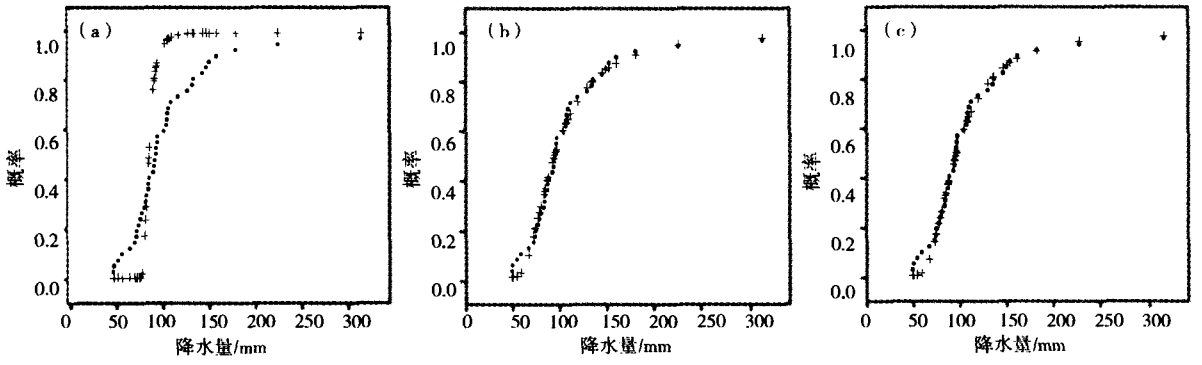

图 1　经验分布函数和理论分布函数对比图(····经验值,+++理论值)
(a) MLE　(b) LSE　(c) LAD

Fig. 1　Comparision of empirical distribution function values with model ones

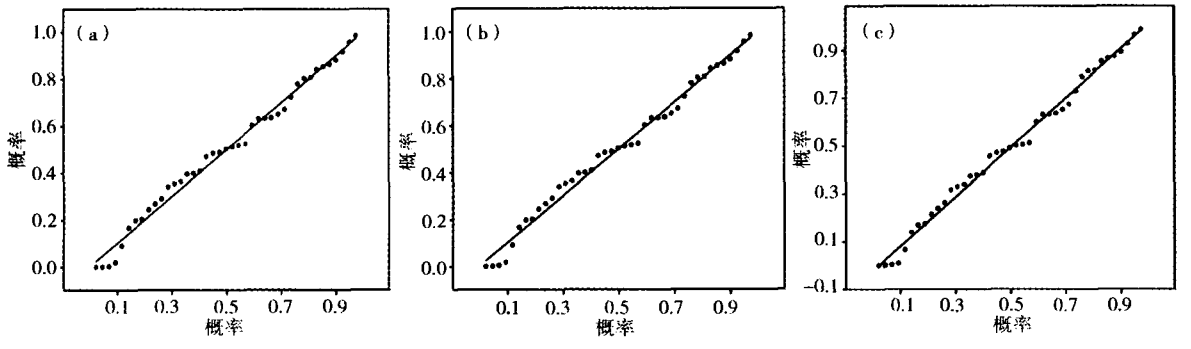

图 2　三种参数估计方法下的 P-P 图
(a) MLE　(b) LSE　(c) LAD

Fig. 2　Probability plots for different estimates

(1)斜度和峰度非零表明是重尾分布,说明降雨量具有很大的不稳定性,这一现象可以用各种突发的自然现象(地震、台风、潮汐等)来解释。那么,重视分布的尾部而不是整个分布是显然的,引进极值方法来研究这类问题具有现实意义。

(2)广义极值分布模型的形状参数在0.15～0.4之间,属于Frechet分布,这与以往相关研究相符。

(3)极值方法的关键步骤就是估计参数,采用LAD改进的回归法比传统的方法更能提高预报的准确度和精度,增强了模型的拟和效果。

(4)在估计出极值分布参数的基础上,可以进行进一步的应用研究,例如回归期问题,工程建筑强度,产品质量,股票市场等,将为我们的生产生活提供更多的科学理论指导。

参 考 文 献

1 Bassett G W. and Koenker R W. Asymptotic theory of least absolute error regression. Journal of the American Statistical Association, 1978,**73**:618～622

2 Samuel Kotz, Saralees Nadarajah. Extreme value distributions theory and applications. Imperial College Press, 2000. 61～94

3 Stuart Coles. An introduction to Statistical Modeling of Extreme Values. The Athenaum Press Ltd. ,2001. 47～53

4 杨扬,施能. 1948～2001年全球陆地9～11月降水的长期变化. 气象科学,2003, **23**(3):253～262

5 张敏,张铭. 南海夏季风爆发的EOF分析. 气象科学, 2004,**24**(3):261～268

6 Gerald L Lieberman, Ingram Olkin. Extreme Value Theory in Engineering. Academic Press Inc. ,1988. 183～208

7 金龙,袁成松,秦伟良,李永华. 基于预报误差最小的自回归模型定阶方法研究. 气象科学,1997,**17**(1):36～42

8 黄伟,傅惠民. 截尾数据线性回归分析方法. 航空动力学报, 2003,**18**(4):465～469

9 Koenker R W and Basset G W. Regression Quantile. Econometrica,1978, **46**:33～50

REGRESSIVE PARAMETER-ESTIMATING WITH LEAST ABSOLUTE DEVIATION IN EXTREME VALUE DISTRIBUTION

Wu Xianghua[1] **Qin Weiliang**[2] **Wang Xinlei**[1] **Yu Shuping**[1]

(1*Mathematic Department, Nanjing University of Information Science and Technology, Nanjing* 210044)

(2*The Institute of Economics and Management, Southeast University, Nanjing* 210096)

Abstract Least Absolute Deviation substituting Least Squares Method is introduced to establish the aim function $\sum_{i=1}^{n}|y^{(i)}-f(x^i,\beta)|$ in the estimation of position, scale and shape parameters in the extreme value distribution. The extreme value in the stochastic process is the maximum or minimum value on a given time trial, so the extreme method emphasizes particularly the tail of a distribution instead of the whole, which makes sense to forecast the disasters, reduce or avoid losses and so on. As one of the most important step, parameter-estimating attracts more and more consideration. Compared with the classical estimation, the new method using LAD and nonlinear regression model improves in many aspects. It has been found that LAD can enhance not only the accuracy of the forecast but also the calculation stability.

Key words Extreme value distribution Least absolute deviation Annual maxima of daily rainfall Kolmogorov test