第3卷 第3期 应用气象学报 Vol. 3, №3
1992年8月 QUARTERLY JOURNAL OF APPLIED METEOROLOGY August 1992

稳健回归的反复加权最小二乘迭代解法及其应用

施 能 王建新
(南京气象学院,南京,210035)

提 要

用反复加权最小二乘迭代法求得稳健回归方程,将稳健回归方程与最小二乘的回归方程作了比较。结果表明,稳健使残差绝对值之和逐步减小并收敛。这种方法用于长江中下游降水预报,独立与非独立样本资料的检验表明,稳健回归方程比最小二乘回归效果好。

一、前 言

目前,根据残差平方和达到最小求解回归方程回归系数的最小二乘方法已得到广泛应用。如果${}_nX_{m+1}$表示观测到的因子矩阵,n是观测次数,m为自变量数,则$X^T=(x_1^T\ x_2^T\cdots\cdots x_n^T)$,$x_i^T=(1\ x_{i1}\ x_{i2}\cdots\cdots x_{im})$。预报对象向量$Y^T=(y_1\ y_2\cdots\cdots y_n)$,右上角"$T$"表示转置,$t=1,2,\cdots\cdots,n$。则回归系数的最小二乘解为

$$\hat{\beta}=(X^TX)^{-1}X^TY \tag{1}$$

$\hat{\beta}^T=(\hat{\beta}_0\ \hat{\beta}_1\ \hat{\beta}_2\cdots\cdots\hat{\beta}_m)$。式(1)所表示的解具有无偏性,并且在一切无偏估计类中,具有最小方差。由于计算简单,又能在正态假定下应用统计检验理论,所以得到广泛应用。但是,近来所积累的经验表明,特别是处理大型回归问题时,X矩阵的列向量往往具有复共线性($X^T\cdot X$矩阵中至少一个特征值很小的例子在气象预报问题中屡见不鲜),从而使($X^T\cdot X$)接近奇异,式(1)的估计性能明显变坏。近代回归所研究的狭义岭回归、广义岭回归、压缩估计、主成分回归、特征根回归颇有成效地改进了最小二乘估计,并用于解决实际问题[1—4]。最小二乘估计的另一缺点是受突出值的影响大,这是因为最小二乘估计是以残差平方和达到最小求解回归系数的,这会使突出值的作用明显增加,从而使回归方程缺少稳健性。所谓稳健回归方法就是设计一个比残差(e_i)平方增长速度慢的函数$\rho(e_i)$去代替e_i^2,例如残差绝对值之和。求解方法涉及到线性规划问题。本文使用一种反复改变权重的加权最小二乘迭代解法,这种解法在国内尚未见到应用。可以利用最小二乘解法的子程序,计算结果表明,这种方法得出的回归方程的残差绝对值之和比最小二乘法减小,收敛速度快,是一种求解稳健回归方程较实用的方法。

1990年10月27日收到,1991年1月11日收到修改稿。

二、原理和计算方法

1. 原理

最小二乘法是使残差平方和 $Q=\sum_{i=1}^{n}(y_i-x_i^T\hat{\beta})^2$ 达到最小，这导致解下列方程组

$$\sum_{i=1}^{n}e_ix_i=0 \quad 或 \quad \sum_{i=1}^{n}\frac{e_i}{\sigma}x_i=0 \tag{2}$$

其中 $e_i=y_i-x_i^T\hat{\beta}$，σ 是 e_i 的总体均方差。式(2)化为更一般的形式为

$$\sum_{i=1}^{n}\psi(\frac{e_i}{\sigma})x_i=0 \tag{3}$$

(3)式的解就是稳健回归的最大似然估计，简称 M-估计[5]。$\psi(\cdot)$在稳健回归中称为影响函数，可以通过选择 $\psi(\cdot)$的形式减少大的残差点的影响，已经研究出多种 $\psi(\cdot)$的函数形式。

最小二乘法的影响函数是残差 e_i 的线性函数，也就是 $\psi(e_i/\sigma)=e_i/\sigma$。但是，更合理的影响函数应是有界的。Huber[5]研究的影响函数是

$$\psi(e_i/\sigma)=e_i/\sigma=\begin{cases}e_i^* & |e_i^*|\leqslant R^0\\ R^0 & e_i^*>R^0\\ -R^0 & e_i^*<-R^0\end{cases} \tag{4}$$

它的图形如图 1。这种影响函数不允许残差 e_i^* 大于 R^0 的突出点对回归系数产生大的影响，从而使回归方程具有稳健性。

2. 计算方法

式(3)不能象式(2)那样得到 β 的显式解。本文用反复加权的最小二乘迭代解[6]。

首先，式(3)的 σ 必须用估计量代替。但是方差、均方差都是稳健性较差的统计量，分布的运尾部分的细微改变就可以大大改变方差、均方差的计算值。然而，中位数是稳健性好的量，考虑到这一点，合理的选择是[6]

$$\hat{\sigma}=1.5\mathrm{med}(e_i)$$

$\mathrm{med}(e_i)$是 e_i 的中位数。

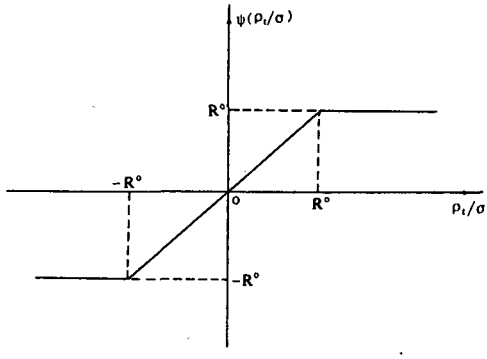

图 1 稳健回归的影响函数

式(3)可化为

$$\sum_{i=1}^{n}\frac{\psi(e_i^*)}{e_i^*}\cdot e_i^*\cdot x_i=0 \tag{5}$$

令 $w_i=\psi(e_i^*)/e_i^*$，则式(5)化为

$$\sum_{i=1}^{n}w_ie_i^*x_i=0 \tag{6}$$

注意到式(2)的由来可知，式(6)相当于 $\sum_{i=1}^{n}w_i(y_i-x_i^T\hat{\beta})^2$ 达最小时的解，也就是加权最小二

乘法。所以式(3)的求解步骤是:

①求初始估计量 $\hat{\beta}^{(0)}$(通常用最小二乘估计代替),从而得到初始的残差及 $e_{i,0}^*$。

②根据 $e_{i,0}^*$算出初始权重 $w_{i,0}=\psi(e_{i,0}^*)/e_{i,0}^*$。

③利用加权最小二乘法解得第一步的 β 的稳健估计 $\hat{\beta}_R^{(1)}$

$$\hat{\beta}_R^{(1)} = (X^T W_0 X)^{-1} X^T W_0 Y \tag{7}$$

其中 W_0 矩阵是一个对角阵,对角线元素为 $w_{i,0}$。

式(7)可沿用已有的最小二乘解法的计算程序,只需作如下变换:

$$M = CX, \quad Z = CY$$

其中 C 是一个以 $\sqrt{w_{i,0}}$为对角元的 n 阶对角矩阵,$C^T \cdot C=W$。则式(7)化为

$$\hat{\beta}_R^{(1)} = (M^T \cdot M)^{-1} M^T Z$$

与式(1)完全一致。这样就用最小二乘法的解法计算了加权最小二乘法。

④用第 3 步所解得的 $\hat{\beta}_R^{(1)}$ 返回并代替第一步的 $\hat{\beta}^{(0)}$,又可得到新的残差 $e_{i,1}$、$e_{i,1}^*$,从而得新的权重 $w_{i,1}$

⑤再继续返回到第 3 步,计算第 2 步的稳健估计 $\hat{\beta}_R^{(2)}$。如果用 $\hat{\beta}_R^{(j)}$、$\hat{\beta}_R^{(j+1)}$ 分别表示第 j 步和第 $j+1$ 步的稳健回归系数向量,则可规定一个迭代收敛的误差标准 ε,当相邻两步之间回归系数差的最大绝对值小于 ε 时,迭代收敛,即

$$\max(|\hat{\beta}_R^{(j+1)} - \hat{\beta}_R^{(j)}|) < \varepsilon \tag{8}$$

实际上,迭代时的每步 $\sum_{i=1}^{n}|e_i|$逐步减少,然后稳定在某个定值附近摆动,所以,也可用相邻两步残差绝对值之和的差小于 ε 时收敛,结果是一样的。

三、在梅雨降水预报中的应用

长江中下游梅雨预报一直是一个很重要的研究课题。因为降水极易受到异常因素的影响,而使降水量出现突出值。也就是降水量分布不对称,运尾部分的概率并不很小。为了减少突出值的影响而采用稳健回归方法,效果可望得到改善。

所预报的地区是长江中下游雨量场大致均匀分布的 38 个测站。预报对象是这 38 个站梅雨期的雨量。梅雨期雨量由逐日雨量累加算得。为了与最小二乘的结果相比较,预报因子沿用文献[7]中的 3 个因子。这 3 个因子是从 13 个因子中在 $F=3.0$ 标准下挑选的,即 x_1(上年 10 月 500hPa(40°N,80°E)与(30°N,80°E)高度距平的平均)、x_2(上年 12 月 500hPa(20°N,80°W)的高度距平)、x_3(当年 2 月 500hPa(80°N,40°E)与(70°N,40°E)高度距平的平均)。

用逐步回归组成的回归方程(最小二乘法)是

$$y = 272.55 - 46.64x_1 + 32.92x_2 + 5.78x_3 \tag{9}$$

这个方程的复相关系数 $R=0.84$。由于 y 的均方差高达 132mm,所以预测 y 有相当的难度。

方程(9)经 F 统计量检验

$$F = \frac{R^2/m}{(1-R^2)/(n-m-1)} = 7.07$$

大于信度 0.01 的 F 临界值($F_{0.01}(3,26)=4.76$)。

对同一份资料,改用稳健回归,影响函数取 Huber 的式(4),R^0 临界值取 1.0。为比较起见,对 $R^0=1.5$ 也进行了类似计算。

表 1 给出最小二乘时的残差 $e_{t,0}$,第一次权重 $w_{t,0}$(分别对 $R^0=1.0$、1.5)以及标准化残差 r_t。

$$r_t=\frac{e_t}{\sigma\sqrt{1-h_{tt}}}$$

$$h_{tt}=\boldsymbol{x}_t^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_t$$

$$\sigma^2=\frac{1}{n-m-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$$

h_{tt}就是帽子矩阵 $\boldsymbol{X}(\boldsymbol{X}^T\cdot\boldsymbol{X})^{-1}\boldsymbol{X}^T$ 的对角元。由表 1 看出,序号 1、18、21、27 具有最大的残差及标准化残差,这 4 年分别是 1954、1971、1974、1980,除 1971 年以外都是涝年。从初始权重看出,这些年份 $w_{t,0}$相应最小。

表 1 最小二乘时的残差 $e_{t,0}$,标准残差 r_t 以及 $R^0=1.0$、1.5 时的初始权重 $w_{t,0}$

序号	最小二乘法		初始权重 $w_{t,0}$		序号	最小二乘法		初始权重 $w_{t,0}$	
t	$e_{t,0}$	r_t	$R^0=1.0$	$R^0=1.5$	t	$e_{t,0}$	r_t	$R^0=1.0$	$R^0=1.5$
1	113.67	2.065	0.681	1.0	15	−22.03	−0.318	1.00	1.00
2	7.65	0.108	1.0	1.0	16	−51.60	−0.775	1.00	1.00
3	−68.35	−1.071	1.0	1.0	17	84.47	1.131	0.92	1.00
4	−54.00	−0.821	1.0	1.0	18	−116.30	−1.600	0.67	0.99
5	−31.20	−0.475	1.0	1.0	19	−92.32	−1.234	0.84	1.00
6	−90.39	−1.228	0.856	1.0	20	−58.12	−0.871	1.00	1.00
7	76.66	1.087	1.0	1.0	21	146.40	2.050	0.53	0.79
8	−89.50	−1.194	0.865	1.0	22	11.99	0.161	1.00	1.00
9	12.27	0.169	1.0	1.0	23	23.16	0.322	1.00	1.00
10	−25.87	−0.379	1.0	1.0	24	−24.02	−0.325	1.00	1.00
11	8.44	0.119	1.0	1.0	25	−41.45	−0.563	1.00	1.00
12	−9.18	−0.124	1.0	1.0	26	50.50	0.694	1.00	1.00
13	73.26	1.002	1.0	1.0	27	148.10	2.027	0.52	0.78
14	17.78	0.254	1.0	1.0					

据表 1 中的 $e_{t,0}$值,算得中位数$(e_{t,0})=51.60$。残差绝对值之和 $\sum_{t=1}^{n}|e_{t,0}|=1548.678$。利用上述算法,若给出收敛标准为 $\varepsilon=0.01$,则 $R^0=1.0$ 时迭代到第 11 步,$R^0=1.5$ 时迭代 6 步。表 2 给出 $R^0=1.0$ 时迭代 11 步的回归系数 $\boldsymbol{\beta}$ 的稳健估计以及各步的残差绝对值之和。

从表 2 看出,经过反复加权迭代,回归方程的回归系数已经收敛。残差绝对值之和也在迭代过程中减少并稳定在一个最小值附近。当继续迭代时,回归系数仅在小数第 3 位改变 0.001,残差绝对值和仅在小数第 2 位改变 0.01。所以,取最后的稳健回归方程为

$$y = 257.71 - 46.34x_1 + 28.89x_2 + 4.89x_3 \tag{10}$$

这个回归方程与最小二乘法的式(9)不同。表 2 已表明，方程(10)的拟合效果从 $\Sigma|e_i|$ 看比方程(9)好。表 3 是对独立样本资料的 7 次预报结果。可以看出，用稳健回归方程后，7 年中有 5 年预测效果得到改善。7 年总的残差绝对值也以(10)式为好。但从预报趋势看，稳健回归与最小二乘方法经常是一致的。

表 2　反复加权迭代的各步回归系数及残差绝对值和($R^0=1.0$)

步数	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\sum_{i=1}^{n}\|e_i\|$
最小二乘	272.55	−46.64	32.92	5.78	1548.678
1	267.39	−46.46	32.16	5.34	1538.141
2	264.32	−46.27	31.20	5.16	1531.016
3	262.23	−46.19	30.51	5.04	1526.035
4	260.47	−46.19	29.87	4.98	1521.904
5	259.11	−46.27	29.39	4.92	1518.590
6	258.32	−46.32	29.12	4.89	1516.728
7	257.90	−46.35	28.98	4.88	1515.764
8	257.76	−46.35	28.92	4.88	1515.485
9	257.72	−46.34	28.90	4.89	1515.419
10	257.71	−46.34	28.89	4.89	1515.416
11	257.71	−46.34	28.89	4.89	1514.421

表 3　稳健回归与逐步回归对独立样本资料的预报误差(单位：mm)

序号	逐步回归(9)	稳健回归(10)	效果
1	−134.8	−125.6	改善
2	−164.4	−155.6	改善
3	98.2	105.7	未改善
4	−170.2	−147.6	改善
5	−241.4	−220.2	改善
6	−136.6	−122.3	改善
7	−5.5	9.2	未改善
$\Sigma\|e_i\|$	950.9	886.2	改善

取 $R^0=1.5$ 时，迭代 6 步得到稳健回归方程：

$$y = 270.55 - 47.73x_1 + 34.01x_2 + 5.52x_3 \tag{11}$$

这个方程更接近逐步回归式(9)，$\Sigma|e_i|=1546.96$，仍比最小二乘法的式(9)好。式(11)对 7 年独立样本预报，效果也有改善。误差绝对值之和为 942.6，比逐步回归的 950.9 有改进，但改进很有限。如果再增加 R^0 的值，取 $R^0=2.0$ 时，则由于 e_i^* 全部小于 2.0，所以结果与最小二乘完全相同，不能建立稳健回归方程。

我们还用上述方法制作了宜昌逐日流量中期预报，样本容量 $n=140$。用最小二乘法建立回归方程的残差绝对值之和为 $6362.29\times100\text{m}^3/\text{s}$。取 $R^0=1.0$，用稳健回归方法，计算到第 14 步收敛，这时的残差绝对值之和为 $6134.59\times100\text{m}^3/\text{s}$，拟合效果改善。进行独立样本预报时，28 次预报用逐步回归的残差绝对值平均为 $69.89\times100\text{m}^3/\text{s}$，用稳健回归

方程的预报残差绝对值平均为 67.29×100m³/s,预报效果也有改善。

四、小 结

1. 稳健回归的反复加权迭代法简单易行,这种方法减少突出值对回归系数的影响,使残差绝对值和减小,从而使预测效果比最小二乘法有所改进。但是,这种改进仅仅是数值方面的改进,一般不能改变预测趋势。

2. 根据稳健回归理论,对于容易产生突出值,运尾概率分布不很小的变量(例如降水量)使用稳健回归效果比较好。此外,根据式(4),e_i^* 大于 R^0 的点认为是突出值,它们对回归系数的影响应受到抑制。根据我们的比较以及正态分布理论,R^0 取 1.0、1.5 是较为合适的(取决于所给资料的概率分布),R^0 的取值小于 1.0 或大于 2.0 均是不合适的。

3. 稳健回归方法并不涉及到选择最优变量子集。它是在变量子集已给定时,建立模型的方法。如果能与选择最优变量子集方法结合起来(逐步回归仅是其中一个方法),可以有更好的预报效果。

4. 通过大量计算表明,取最小二乘法的回归系数为迭代初值,进行逐次迭代运算,$\sum_{i=1}^{n}|e_i|$ 逐步减少并趋稳定,因此计算结果是收敛的。然而,理论上证明其收敛性以及收敛条件仍是值得进一步研究的。

参 考 文 献

[1] 陈希孺、王松桂,近代回归分析,217—268,安徽教育出版社,1987 年。
[2] 俞善贤、汪 铎,试用最优子集与岭迹分析相结合的方法确定回归方程,大气科学,12,4,1988。
[3] Weisberg,s. ,Applied Linear Regression,Second edition,260—276。John Wiley & Sons,New York,1985.
[4] 冯翺煌、吴达三,岭回归在预报集成中的应用,气象,11,11,1985。
[5] Huber,P. J. ,Robust statistics:A review,*Annals of Mathematical Statistics*,43,1041—1067,1972.
[6] Myers,H. R. ,Classical and Modern Regression with Applications,195—217,Duxbury Press,Boston,1986.
[7] 王建新,长江中下游地区梅雨期雨量场与 500 百帕月平均高度场的相关分析,气象科学,9,3,1989。

A METHOD OF ITERATIVELY REWEIGHTED LEAST SQUARE AND ITS APPLICATION

Shi Neng Wang Jianxin
(*Nanjing Institute of Meteorology*, *Nanjing*, 210035)

Abstract

The robust regression is obtained by using the method called the iteratively reweighted least squares (IRWLS). The comparison between robust regression and least squares (LS) regression is also made. The results point out that the procedure may lead to gradual reduction and convergence of the sum of absolute residuals. This procedure is also used to forecast the rainfall over the middle and lower reaches of the Changjiang River. The tests on dependent and independent sample data prove that the IRWLS is better than the LS.