

岭回归在区域洪水频率分析中的应用

邓正波

(广东省潮州供水枢纽管理处,广东 潮州 521011)

摘要 探讨自变量多重相关条件下岭回归方法在区域洪水频率分析中的应用.首先确定水文子区域,建立洪峰流量与流域特征变量的区域回归方程,然后用岭回归分析确定回归方程的系数,建立该区域估算洪峰流量的区域回归方程,并对计算结果进行评价,分析表明多重相关条件下岭回归在区域洪水频率分析中的应用优于最小二乘回归.

关键词 设计洪水;洪水频率分析;最小二乘回归;岭回归

中图分类号:P333.9 文献标识码:A 文章编号:1006-764X(2007)S2-0049-03

在水利工程设计和水资源管理工作中,多数洪水频率分析技术需要足够长的水文时间系列,而资料系列偏短或没有水文资料的情况普遍存在.大量研究显示,区域分析方法能有效提高资料缺乏地区设计洪水的可靠性,假定在一致区内的站点具有同一概率分布,综合一个区域的若干个测站的实测样本资料去估计某个单站的参数,有利于充分利用区域范围的各种信息,这种频率分析比单站分析估计更加可靠.

国内近年来对区域洪水频率分析也做了一些研究^[1-3],用最小二乘法和加权最小二乘法进行区域回归法洪水频率分析,并比较两者的推算结果,认为这两种方法都不能有效地消除站点间观测资料的多重相关性,计算结果不理想,计算方法需要改进.本文应用岭回归方法进行区域洪水频率分析,以解决多重相关导致计算结果不理想的问题.

1 多重相关性

多重相关性是指自变量之间存在着线性相关的现象.在实际工作中,多重相关性十分普遍,且对回归计算造成了很大危害^[4].在自变量完全相关的情况下,无法估计最小二乘回归系数;不完全相关时,回归系数的估计方差将随着自变量之间相关程度的增强而迅速扩大,在高度相关条件下,回归系数估计值不稳定,对用最小二乘法得到的回归模型将很难解释回归系数的物理意义,对回归系数的一般解释也将不再适用.

常用的多重相关性的诊断方法是方差膨胀因子.自变量 x_j 的方差膨胀因子为 $(VIF)_j$:

$$(VIF)_j = (1 - R_j^2)^{-1} \quad j = 1, 2, \dots, n \quad (1)$$

式中 R_j^2 是因变量 x_j 与其回归因变量的复测定系数. $\max(VIF)_j > 10$ 表示多重相关性将严重影响最小二乘法的估计值,可以认为该自变量之间高度相关.

2 岭回归基本原理

通用的流量与相关影响因子的指数关系如下:

$$Q = \alpha_0 A_1^{\alpha_1} A_2^{\alpha_2} \dots A_n^{\alpha_n} \epsilon_0 \quad (2)$$

式中: Q 为流量; $A_1, A_2, \dots, A_n$ 为流域特征变量(影响因子); $\alpha_0, \alpha_1, \dots, \alpha_n$ 为模型参数; n 为特征变量的个数; ϵ_0 为乘积误差项.式(2)两边取对数变换后得到通用多元线性回归方程:

$$Y = b + a_1 X_1 + a_2 X_2 + \dots + a_n X_n + \eta_0 \quad (3)$$

式中: $Y = \lg Q$; $X_i = \lg A_i$ ($i = 1, 2, \dots, n$); $\eta_0 = \lg \epsilon_0$.

岭回归是一种修正的最小二乘估计法,能较为有效地消除多重相关性的不良影响,当自变量系统中存在多重相关性时,能提供一个比最小二乘法更为稳定的估计.

根据高斯-马尔柯夫定理,在线性回归模型的基本假定被满足时,用最小二乘法得到的回归系数估计量是无偏的,且具有最小方差.可以证明,即使在高度多重相关的情况下,最小二乘法的回归系数估计量在无偏估计中依然方差最小,但可能比有偏估计量的方差大.有偏估计量有微小的偏差,但精度却能够大大提高,且更接近真实参数(见图1).

从图1可以直观地看出小方差的有偏估计量可能比大方差的无偏估计量更接近真值.估计量 b

作者简介:邓正波(1979—),男,广东韶关人,硕士研究生,从事水资源管理研究. E-mail: xingmubo@163.com

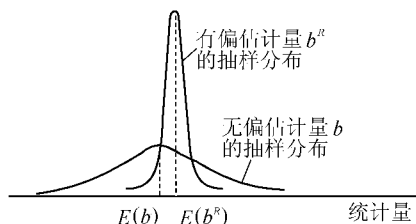


图1 小方差有偏估计量示意图

是真值 β 的无偏估计,但它的抽样分布方差很大,且不精确,而估计量 b^R 的精度却有小的偏差。在一次抽样结果中, b^R 落在真值 β 附近的概率远远大于无偏估计量 b 落在 β 附近的概率。

有偏估计量 b^R 与真实参数 β 离差平方和的期望值称为均方误差:

$$E = \|b^R - \beta\|^2 = \text{Var}(b^R) + [E(b^R) - \beta]^2 \quad (4)$$

这时,均方误差就等于估计量的方差加上偏差的平方。如果估计量是无偏的,均方误差就等于估计量的方差。

标准的回归模型矩阵表达式为

$$Y = X\beta + \varepsilon \quad (5)$$

式中: Y 为因变量向量; X 是 $N \times M$ 阶的自变量矩阵; β 为回归系数向量; ε 为剩余向量。

岭估计 $\hat{\beta}_k$ 具有如下形式:

$$\hat{\beta}_k = (X'X + kI)^{-1}X'Y \quad (6)$$

式中 k 为岭参数 $0 < k < 1$ 。显然 $k = 0$ 时即为普通最小二乘估计 $\hat{\beta} = (X'X)^{-1}X'Y$ 。采用较为简单有效的一种方法来选择 k 值使岭估计优于普通最小二乘估计,即

$$0 < k < k_{\max} \quad (7)$$

式中 $k_{\max} = \sigma^2 / (\max_{1 \leq i \leq M} \gamma_i^2)$, $\gamma = P'\hat{\beta}_k$, P 为一正交矩阵,由 $X'X = P'\Lambda P$ 得出, Λ 为一对角矩阵,其主对角元素为 $X'X$ 的特征值,则式(6)所定义的岭估计量 $\hat{\beta}_k$ 便优于最小二乘估计 $\hat{\beta}$ 。

实际上,由于 σ^2 和 β 都不是可实测的,这里以最小二乘估计得到的 $\hat{\beta}$ 代表 β ,而 σ^2 由下式估计:

$$\sigma^2 = (Y - X\hat{\beta})(Y - X\hat{\beta})'(N - M) \quad (8)$$

令 $k = ik_{\max}$ ($i = 0, 0.1, \dots, 1$),计算回归系数的估计值,进行回代,此时,称由普通最小二乘估计的 $\hat{\beta}$ 得到的回代结果 $\hat{Y} = X\hat{\beta}$ 为回归值。显然此处 $\hat{Y}$ 亦为 N 维向量。当回归系数由岭估计求得时,回代的公式在形式上完全相同,只是将 $\hat{\beta}$ 换成 $\hat{\beta}_k$ 即可。

3 岭回归模型的建立与检验

3.1 建立模型

本次研究选择淮河息县以上流域 5 个水文站的流量资料和 35 个雨量站的雨量资料作为研究对象,

首先进行水文资料的审查与分析,确保频率分析的有效性。经分析,这些流量资料随机独立,不需要作修正,用模糊聚类法进行水文相似性分区,得到具有水文相似性的区域;然后用逐步回归法筛选影响较大的流域特征变量,建立洪峰流量与流域特征变量的区域回归方程,用岭回归分析确定回归方程的系数,建立该区域估算洪峰流量的区域回归方程,最后推求流域设计洪水。

逐步回归法筛选得影响较大的流域特征变量因子为: A_1 为年平均流量; A_2 为年径流模数; A_3 为逐年 7 月降雨量; A_4 为逐年 8 月降雨量。计算 $(VIF)_j$ 发现其最大值均大于 10,说明因子间存在多重相关性,故选用岭回归进行估计。

对流量及其流域特征变量因子做对数变换后构造出该分区的形如式(3)的多元线性回归模型:

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} + \theta \quad (9)$$

式中: y_i 为因变量; x_{ij} 为自变量($i = 1, 2, \dots, n$; $j = 1, 2, 3, 4$; n 为资料长度),用岭回归估计求得回归方程系数的估计值,这里偏量 $k = 0, 0.001, 0.002, \dots, 1$,岭回归分析部分结果见表 1。表中 SSE 为残差平方和, R^2 为复相关系数,用来判断线性回归模型拟合程度的综合测定指标, R^2 值越大,线性方程拟合程度越好。

表1 息县站岭回归分析部分结果

k	b_0	b_1	b_2	b_3	b_4	SSE	R^2
0.000	9.372	-7.085	8.133	0.158	-0.138	1.788	0.746
0.001	2.652	-0.410	1.489	0.154	-0.148	1.837	0.740
0.002	2.205	0.034	1.046	0.154	-0.149	1.843	0.739
0.003	2.042	0.195	0.885	0.154	-0.148	1.846	0.738
0.004	1.958	0.278	0.801	0.154	-0.148	1.847	0.738
0.005	1.907	0.329	0.750	0.154	-0.148	1.848	0.738

最佳 k 值是要牺牲最小的偏差来获取最优的拟合值,使估计量的偏差和方差的组合效应达到一个最佳水准,即

$$SR = \min \left(\frac{SSE_{k=k_i} - SSE_{k=k_0}}{R_{k=k_i}^2 - R_{k=k_0}^2} \right) \quad i = 1, 2, \dots, 10 \quad (10)$$

从表 1 可以看出,随着 k 的增加,模型 SSE 从 $k = 0$ 时的 1.788 增加到 $k = 0.001$ 时的 1.837,而 R^2 则从 0.746 下降到 0.740,即 SR 取得最小,可以判断 $k = 0.001$ 时的岭分析结果是满意的。可以得到该区的岭回归方程:

$$y = 2.652 - 0.410x_1 + 1.489x_2 + 0.154x_3 - 0.148x_4 \quad (11)$$

3.2 模型检验

回归方程显著性检验的目的是检验因变量 y

是否与自变量 $x_1, x_2, \dots, x_p$ 存在线性关系。如果在总体数据中,确实存在这种线性关系,或者说,确实可以用 $x_1, x_2, \dots, x_p$ 的线性形式来解释 y , 则至少应该存在一个 x_k 使得 y 对 x_k 的总体参数 β_k 不等于零,否则,所有的总体参数 $\beta_j (j = 1, 2, \dots, p)$ 均等于零。因此,对回归方程进行显著性检验。

原假设:

$H_0: \beta = \alpha$ 即 $\beta_1 = \beta_2 = \dots = \beta_p = 0$

$H_1: \beta \neq \alpha$ 至少存在 β 的一个分量 $\beta_k \neq 0$);

统计量:

$$F = \frac{SSR/p}{SSE/n - p - 1} \sim F(p, n - p - 1) \quad (12)$$

式中:SSR 是回归平方和;SSE 是残差平方和;SST 是方差总和。

由给定的显著水平 α ,可查 F 分布表,得到拒绝域的临界值 $F_\alpha(p, n - p - 1)$ 。若 $F \leq F_\alpha(p, n - p - 1)$ 则接受 H_0 假设,认为 Y 与 X 无显著的线性关系;若 $F > F_\alpha(p, n - p - 1)$ 则否定 H_0 ,认为 y 可以用 $x_1, x_2, \dots, x_p$ 的线性模型拟合。回归方程(11)的显著性检验见表2,从表中看出,回归方程通过 F 检验,说明其具有显著性。

表2 岭回归分析显著性检验($\alpha = 0.05$)

方差来源	方差和	自由度	方差	方差比 F	临界值 F_α
SSR	5.216	4	1.304		
SST	7.053	49	0.144	31.952	2.174
SSE	1.837	45	0.041		

4 岭回归方程在区域洪水频率中的应用

应用区域回归方程估算洪峰流量有多种情况^[3]。本文用各个子区域的流域面积占总流域面积之比对洪峰流量进行加权来推求区域无因次洪水频率设计值。对区域内的任何一站点,概率 p 所对应的洪水流量 $Q_p = Q_i q(p)$, Q_i 为各站流量均值。岭回归方程推求的息县地区无因次洪水频率设计值结果见表3,表中 $Q_{\text{设计}}$, $Q_{\text{最小二乘}}$, $Q_{\text{岭回归}}$ 分别为由实测值、最小二乘回归方程、岭回归方程推求频率 p 所对应的流量值; $S_{\text{最小二乘}}$, $S_{\text{岭回归}}$ 分别为由最小二乘回归方程、岭回归方程推求流量值的误差。

用区域洪水频率曲线推求的单站流量设计值跟单站实测资料推求的流量设计值相比,大部分误差较小,可以认为,区域洪水频率分析中的回归法可用于估算无资料地区的设计洪水,且岭回归估计值的误差方差比最小二乘估计值的误差方差小,表明在区域洪水频率分析中岭回归方法比最小二乘回归方法好,岭回归可以提高区域洪水频率分析的精度。

表3 回归方程推求的息县地区无因次洪水频率设计值及设计流量值

频率 $p/\%$	$Q_{\text{设计}}/(\text{m}^3 \cdot \text{s}^{-1})$	$Q_{\text{最小二乘}}/(\text{m}^3 \cdot \text{s}^{-1})$	$S_{\text{最小二乘}}$	$Q_{\text{岭回归}}/(\text{m}^3 \cdot \text{s}^{-1})$	$S_{\text{岭回归}}$
0.01	21601	23385	0.08	22358	0.04
0.02	20131	22299	0.11	21315	0.06
0.05	18176	20868	0.15	19939	0.10
0.10	16555	19788	0.20	18902	0.14
0.20	15198	18712	0.23	17868	0.18
0.50	13209	17296	0.31	16508	0.25
1.00	11691	13084	0.12	12484	0.07
2.00	10158	8879	-0.13	8468	-0.17
3.00	9250	8266	-0.11	7880	-0.15
5.00	8095	7499	-0.07	7145	-0.12
10.00	6499	6475	0.00	6165	-0.05
20.00	4850	3902	-0.20	3709	-0.24
25.00	4303	3652	-0.15	3471	-0.19
30.00	3846	3337	-0.13	3171	-0.18
40.00	3108	2949	-0.05	2802	-0.10
50.00	2508	2658	0.06	2526	0.01
60.00	1991	2428	0.22	2309	0.16
70.00	1527	1296	-0.15	1231	-0.19
75.00	1305	1230	-0.06	1168	-0.10
80.00	1088	1138	0.05	1082	-0.01
90.00	642	375	-0.42	356	-0.45
误差方差			0.030		0.028

6 结 语

实际工作中,多重相关性普遍存在,其形成原因主要有两个:一是某些变量的物理含义决定了它们之间的相关性,二是样本长度不足。变量之间的多重相关性严重影响了回归结果的可靠性。因此,研究出能消除变量间多重相关的回归方法甚为重要,岭回归是修正变量间多重相关性的有效方法。

本文建立流量与其相关因子的经验指数关系,对数线性化后进行岭回归分析,并进行相关统计检验,确保回归方程符合最初构建方程时的理论假设,且可用于预测。分别用岭回归和最小二乘回归进行区域洪水频率分析,分析结果与由实测资料推求的频率设计值的比较表明,区域洪水频率分析用于估算设计洪水是合适的,且岭回归结果优于最小二乘结果。对于资料短缺或无资料的区域,容易获得回归方程中的特征变量值较延长,且回归方程应用简单方便。

参考文献:

[1] 张静怡,徐小明.极值分布和 P-III 型分布线性矩法在区域洪水频率分析中的检验[J].水文,2002,22(6):12-16.
[2] 周芬,郭生练,方彬,等.区域回归法对无资料地区设计洪水的估算[J].水力发电,2004,30(7):10-13.
[3] 周芬,郭生练,肖义,等.无资料地区设计洪水的区域频率分析[J].人民长江,2004,35(5):29-31.
[4] 王惠文.偏最小二乘回归方法及其应用[M].北京:国防工业出版社,1999:1-368.

(收稿日期:2007-01-23 编辑:骆超)