DOI:10.13203/j.whugis1993.01.003

第 18 卷第 1 期　　武汉测绘科技大学学报　　Vol. 18 No. 1
1993 年 3 月　　Journal of Wuhan Technical University of Surveying and Mapping　　Mar. 1993

大坝变形监测数据回归分析中的因子选择

吴　子　安

摘　要　本文对我国大坝变形资料分析中常用的逐步回归分析进行了探讨，指出这种方法通常所选的因子数偏少，其原因来自自变量之间的复共线性的影响。为了克服复共线性对因子筛选的影响，文中对因子筛选提出了若干有益的建议。

关键词　大坝；变形分析；逐步回归；因子筛选；复共线性

我国各类工程建筑物变形观测中，尤以大坝变形观测最为重视，其观测手段完善、设备先进、资料系统完整、处理及时等方面是其它工程建筑物变形观测难以比拟的。目前大坝变形监测正开始进入数据自动采集与建立监测资料自动安全管理系统阶段，在大坝监测资料分析系统中，回归分析方法已被人们接纳为统计学建模程序中主要方法之一[1,2]。因此探讨回归分析方法在大坝变形建模中的适用性，对於完善大坝变形监测资料分析系统显然是极有实际意义的课题。

1　大坝变形资料回归分析中因子的选择

1.1　自变量因子的初选

在大坝变形分析中，当用统计学方法建模时，自变量因子的初步选择主要考虑三类主要因子：上下游水位变化，温度变化和不可逆变化。水位变化可用多项式拟合[1]，温度变化可用积累平均气温作因子[2]。例如在建立陆水大坝变形预报的回归方案时，初步选择如下自变量因子。

水位因子：考虑到库水位 H 与变形呈抛物线关系，故选择 $(H-H_0)$，$(H-H_0)^2$ 与 $(H-H_0)^3$ 三个因子作为水位因子；此外还选取上下游水位差 DH 因子。

温度因子：选择水库水温因子与气温因子(对未埋设内部观测仪器的大坝，这是常用的方法)。考虑到实测资料反映了坝体变形迟后於温度这一特点，初选 T_1、T_{10}、T_{30}、T_{60}、T_{90}作为气温因子，它们分别表示观测当天，观测前 10、30、60 与 90 天的气温平均值。

时效因子：对于已运行很长时间的大坝(如陆水坝)，可简单地认为时效与大坝运行时间成比例，也即直接选时间因子作为时效因子。

用上述因子选法，在西津大坝资料分析时，选择了四个水位因子，二个水温因子，七个气温因子，一个水头因子与一个时效因子。对丹江口大坝，则选择了三个水位因子，四个气温因子，一个时效因子。

收稿日期：1992-06-29

这种因子初选方式，很自然地会导致各因子之间的相关性。

1.2 回归分析中因子的筛选

大坝变形分析中，为了获得最佳回归方程，对初选因子的筛选，通常采用逐步回归分析方法。考虑到初选因子之间的相关性，一般，对回归系数显著性比较时，每进行一次检验后，只能剔除(或选入)其中一个因子，然后重新建立新的回归方程，再进行新的一次检验[3]。考虑到增加一个因子后，还可能引起原有因子之间的相关性，在逐步回归中，从选进第三个因子开始，每选入一个因子后，需对已进入回归方程的因子的显著性进行重新检验，以便剔除不显著的因子。这种处理方法是否能达到克服自变量之间相关性的影响有待进一步分析。

文选[4]提出："在回归问题中采用逐步方法选择变量要谨慎……逐步方法是在非共线情形下的选择变量的有用工具"。显然，该书作者对逐步回归分析方法用於大坝变形分析(自变量密切相关)是持怀疑态度的。事实上，因子间的相关性正是变形分析的障碍，为此，对逐步回归所选因子的合理性作深入分析，对於判断我国目前大坝资料分析中所得回归方程的预测精度是有实际意义的。

2 大坝变形分析中初选因子间的相关性和复共线性

由前所述，大坝变形分析中，经常由同一变量(如水位，气温)演变出几个初选因子，这种演变必然会导致自变量之间的相关性。相关性不仅仅出现在同一变量演变的因子之间，而且还存在於不同变量(如水位与气温)之间。上述现象已为大坝分析资料所证实。

当初选自变量 x_1、x_2、……x_r 之间存在线性相关时，则由线性代数可知，必存在不全为零的数 k_1、k_2、……k_r，使

$$k_1x_1 + k_2x_2 + \cdots\cdots + k_rx_r = 0 \tag{1}$$

在上述情况成立时，则由初选自变量组成的正规方程 $A=X^TX$，它的特征值 λ_1、λ_2、……λ_r 中，将存在数值上为零的特征根。

在大坝变形分析中，初选自变量之间严格满足(1)式的情况几乎是不存在的，但存在

$$k_1x_1 + k_2x_2 + \cdots\cdots + k_rx_r \approx 0 \tag{2}$$

例如对陆水大坝，由计算求得的 $A=X^TX$ 之特征根与特征向量可写出式(2)的两个近似复共线关系式：

$$-0.0257(H-48)^2 - 0.0023T_5 - 0.0049N + 0.0183DH + 0.0016T_{90} - 0.0083T_1$$
$$-0.4010(H-48) + 0.8022T_{10} - 0.4394(H-48)^3 + 0.0397T_{30} - 0.0007T_{60} = 0.0006$$
$$\simeq 0$$

$$0.0026(H-48)^2 + 0.3542T_5 - 0.4233N + 0.0542DH - 0.1767T_{90} - 0.3867T_1$$
$$+0.6256(H-48) + 0.1368T_{10} - 0.3113(H-48)^3 + 0.0197T_{30} + 0.0161T_{60} = 0.0051 \simeq 0$$

为了判断大坝变形资料中复共线性的严重程度，根据文选[5]所得出之经验系数：

$0 \leqslant k \leqslant 100$	不存在复共线性
$100 \leqslant k < 1000$	存在较严重的复共线性
$k > 1000$	存在极严重的复共线性

式中，k 为最大特征根与最小特征根的比值，即 $k=\frac{\lambda_1}{\lambda_r}$。

对於陆水大坝

$$k = \frac{\lambda_1}{\lambda_{10}} = 11525$$

对於丹江口大坝

$$k = \frac{\lambda_1}{\lambda_8} = 216$$

对于西津大坝

$$k = \frac{\lambda_1}{\lambda_{15}} = 3154$$

显见复共线性在上述三个大坝资料中严重存在。

3 自变量之间的复共线性对逐步回归分析因子选择的影响

逐步回归分析中因子显著性检验[3]所用零假设

$$H_0: \quad \beta_j = 0 \tag{3}$$

检验所用统计量

$$\frac{b_j - \beta_j}{\sqrt{C_{jj}\hat{\sigma}^2}} \sim \mathrm{N}(0,1) \tag{4}$$

或

$$\frac{b_j - \beta_j}{\sqrt{C_{jj}\hat{\sigma}^2}} \sim \mathrm{t}(N-p-1) \tag{5}$$

此处：C_{jj}为矩阵 $C=A^{-1}$中对角线上第 j 个元素；b_j 为回归系数 β_j 的估值。

$$\hat{\sigma} = \frac{S_{剩}}{N-p-1}$$

由线性代数可知，矩阵 X^TX 可谱分解为：

$$X^TX = \Phi\Lambda\Phi^T \tag{6}$$

此处 $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots\cdots,\lambda_p)$，$\lambda_1\geqslant\lambda_2\geqslant\cdots\cdots\geqslant\lambda_p>0$ 为 X^TX 的特征根

$\Phi=(\varphi_1,\varphi_2\cdots\cdots,\varphi_p)$，$\varphi_1,\varphi_2\cdots\cdots,\varphi_p$ 为对应於特征根 $\lambda_1,\lambda_2,\cdots\cdots\lambda_p$ 的标准正交化特征向量

(6)式两边求逆得

$$C = (X^TX)^{-1} = \Phi^T\Lambda^{-1}\Phi \tag{7}$$

采用符号 $\varphi_i=(\varphi_{i1},\varphi_{i2},\cdots\cdots,\varphi_{ip})^T$ $\quad i=1,2,\cdots\cdots,p$ 则可求得矩阵 C 之对角线元素为：

$$C_{ii} = \sum_{j=1}^{p}\varphi_{ji}^2\frac{1}{\lambda_j} \qquad i=1,2,\cdots\cdots,p \tag{8}$$

顾及到回归系数估计精度

$$m_{b_i} = \hat{\sigma}\sqrt{C_{ii}}$$

即可求得 β 的均方误差：

$$\mathrm{MSE}(\hat{\beta}) = \sum_{i=1}^{p}m_{b_i}^2 = \hat{\sigma}^2\sum_{j=1}^{p}\frac{1}{\lambda_j} \tag{9}$$

当自变量之间存在复共线关系时，则存在接近於零之特征根 λ_i，使式(9)中 $1/\lambda_i$ 变得很大，导致 $\mathrm{MSE}(\hat{\beta})$变得很大，这一现象表明，由回归方程计算之回归系数虽然具有无偏性，但本身的估值 $\hat{\beta}$是不稳定的。表 1 为陆水坝按因子显著性逐个引入回归方程后，计算的回归系数值。由表 1 可看出，回归系数 b_2、b_7 在因子数增加过程中甚至产生符号上的变化，这一情况从自变

量对因变量影响的专业分析角度来说是不能接受的。

表 1

因子数	b_0	$H-48$ b_1	$(H-48)^2$ b_2	$(H-48)^3$ b_3	T b_4	S b_5	T_{20} b_6	T_{60} b_7	T_{90} b_8	T_{10} b_9	DH b_{10}	x b_{11}
0	-1.0000											
1	-4.143							0.1863				
2	-4.2078				0.0474			0.1448				
3	-5.3504				0.0498			0.1424				0.0002
4	-5.8458	-0.0381			0.0465			0.1501				0.0003
5	-10.0531	-0.2437			0.0429			0.1352			0.2281	0.0002
6	-11.1143	-0.3509	0.0145		0.0457			0.0152			0.2562	0.0003
7	-11.7679	-0.4350	0.0275		0.0181		0.1013	0.0680			0.3263	0.0002
8	-11.5375	-4558	0.0309		0.0187		0.1569	-0.0769	0.0991		0.3005	0.0002
9	-11.5743	-0.4915	-0.0110	0.0060	0.0224		0.1811	-0.1344	0.1324		0.3159	0.0002
10	-12.0586	-0.4934	-0.0058	0.0053	0.0289	-0.0435	0.1990	-0.1142	0.1248		0.3376	0.0002
11	-11.8829	-0.4888	-0.0064	0.0054	0.0277	-0.0427	0.2012	-0.0943	0.0760	0.0298	0.3273	0.0002

自变量之间的复共线性导致回归系数不稳定性，还反映在随观测资料的增减，使逐步回归分析中所选因子的变化与回归系数数值的变化上。表 2 为根据陆水大坝对同一观测点按不同观测期数用逐步回归分析所选因子与计算求得之回归系数数值。

表 2

观测时段	观测期数	b_0	H b_1	H^2 b_2	H^3 b_3	DH b_4	T_1 b_5	T_5 b_6	T_{10} b_7	T_{30} b_8	T_{60} b_9	T_{90} b_{10}	x b_{11}
69—69	13	−2.9530								0.1971			−0.0087
69—71	37	−2.6769									0.5015	-0.3740	
69—72	52	−2.6296								0.1181	0.2766	-0.2681	
79—79	12	−4.8821						0.1887			0.0479		
79—80	24	−4.2700						0.0590	0.0807				

复共线性造成的回归系数不稳定与回归系数估值误差过大，将使本来对因变量是重要的因子 x_j，可能由於(5)式之分母过大，导致统计量偏小而被剔除。表 3 为几个大坝的计算统计表，由表中数据不能不使人怀疑，逐步回归分析所选用因子数目偏少。

表 3

项　目	预选因子数	近於零的特征根数	逐步回归所选因子数
陆水 46#	11	2	2
丹江 27#	8	1	1
西　津	15	4	4

表 4[16]为陆水大坝逐步回归与多元回归所求回归方程拟合的残差平方和与残差中误差。表中数据不能得出逐步回归所得回归方程为最优的结论。其原因不能不考虑是自变量因子之间复共线性的影响。

表 4

来　源	残差平方和	自由度	残差中误差
逐步回归	2.745	19	0.3801
多元回归	0.521	10	0.2282

4　对回归分析中因子选择的几点建议

4.1　初选自变量不应轻易剔除

如前所述，在建立统计模型时，初选自变量是建立在坝工理论与实测资料分析基础上的。一般来说，水库水位变化、温度变化和不可逆变化对大坝变形均应占有一定程度的影响[1]。当然对大坝运营的不同时期，或大坝不同的部位，它们的影响将产生变化。例如大坝运行初期，不可逆变化将显著存在，又如对大坝坝顶部位的变形，温度影响将变得很显著，而对设置在大坝底部廊道的观测点，温度影响将变弱。应该指出，影响显著与影响变弱都是相对的。

当因子之间存在相关性时，影响弱的因子的影响将被影响强的因子所掩盖，逐步回归分析中将影响弱的因子剔除出了回归方程。例如陆水大坝作逐步回归分析时，求得坝顶观测点的回归方程[6]只选了温度因子，而坝底廊道观测点的回归方程则只选了水位因子。

从影响变形因子的重要性来说，所得回归方程是正确的，但从专业分析来看，不能接受坝顶变形点不受水位影响，坝底廊道中观测点不受温度影响。

逐步回归分析中用影响强的因子代替影响弱的因子（或者说剔除了影响弱的因子），其结果是所得残差不服从同一正态 $N(0,\sigma)$ 分布的要求，这一现象表明，被剔除因子（水位因子或温度因子）的残余影响未能在回归方程中得到反映。

从另一方面来说，如前所述，在大坝变形分析中，水位对变形的影响常用多项式拟合，变形与库水位之间关系选为：

$$\delta l = b_0 + b_1 h + b_2 h^2 + b_3 h^3 \tag{10}$$

采用变量变换式

$$x_1 = h, \qquad x_2 = h^2, \qquad x_3 = h^3$$

则库水位对变形值的影响因子变成 x_1、x_2、x_3 三个因子，它们联合影响的结果将反映变形与库水位的非线性关系[1]（见图 1），显见这三个因子之间存在相关性，且它们对变形的影响是互相补充的。

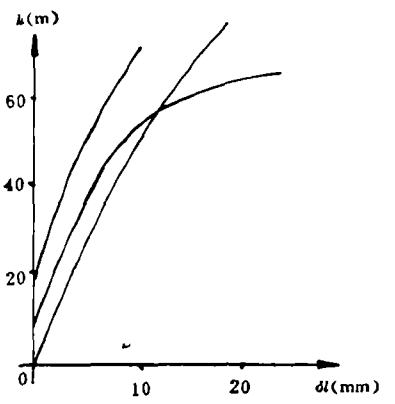

图 1

在逐步回归分析中，将上述三个水位影响因子看成是相互无关的变量，而且经逐步回归分析作因子筛选后，上述三个因子经常只选入一个因子，有时甚至连一个因子也未选入（如陆水坝顶变形分析）。

当人们试图利用所得回归方程对变形作进一步分析时，例如利用回归方程之相应自变量计算水位、温度、时效对大坝影响的变形分量时[1]，逐步回归分析中因子的筛选

方法就愈显出其不足之处。

综上所述，可见对回归分析中初选因子的筛选应持慎重态度，不应轻易剔除。

4.2 关於因子剔除数目的探讨

逐步回归分析中简单地将因子分为重要与不重要两类，而在用统计检验判断因子重要性时又忽视了自变量之间复共线性的影响，这样做极易造成自变量因子的过多剔除。为了克服这一缺点，作者建议用自变量之间的复共线关系数目作为剔除自变量个数的初控值，即首先计算 X^TX 的特征根，计算复共线判断系数 k，由此判断可能存在的复共线关系数目，例如，对陆水大坝有两个复共线关系，将它作为可能剔除的自变量数目的初控值，或者把这些复共线关系作为正规方程解算时的附加约束条件。

为了使复共线关系数目作为因子剔除的初控值能付诸实施，可以根据因子挑选时的情况对因子附加一些必要条件。例如，当对水位与变形的分析认为存在式(10)之抛物线关系时，则可以把代表水位的三个因子($x_1=h, x_2=h^2, x_3=h^3$)作为一个因子子集，逐步回归计算时，把它们的联合影响看作一个因子进行处理。类似地对温度因子则可以通过对变形过程线的分析，根据变形对温度的迟后情况，把气温因子分成不同的子集，再进行逐步回归来最后选择因子。

4.3 关於自变量对因变量影响的权

利用主成分分析，先将各变量转换成相互独立的新变量，再进行变量删减求新变量相应之回归方程，最后再转换成原来的自变量所相应的回归方程，笔者认为由原自变量通过主成分转换成新变量的作法，体现了各自变量的作用大小，类似於对各自变量取不同的“权”值，是一个值得进一步探讨的方法。有关这一计算方法，在文选[5]回归系数的主成分估计中有较详细的介绍，此处不再赘述。

4.4 大坝变形分析中，回归分析因子筛选的标准

通常评判回归方程的优劣的标准是残差平方和，从这一标准出发，回归分析中因子越多，残差平方和越小。特别需要指出的是：这一标准评判回归方程优劣时，不能反映自变量因子之间复共线性的影响。如前所述，复共线性将导致回归系数数值的不稳定性，也即将导致回归系数估计误差过大。当用回归方程

$$\hat{y} = b_0 + \sum b_i x_i$$

去作预报时，预报值 $\hat{y}$ 的正确性将受到回归系数 b_i 的误差影响。笔者认为当自变量之间存在复共线关系时，回归方程效果的好坏应该采用回归系数 $\hat{\beta}$ 的均方误差[6] $\mathrm{MSE}(\hat{\beta})=\mathrm{E}(\hat{\beta}-\beta)^2$，考虑到回归系数的真值通常无法求得，为了克服这一困难，可以采用预测回归系数误差平方和来评判回归方程的优劣。

参 考 文 献

1 李青岳等. 工程测量学. 北京：测绘出版社，1984，2～14

2 吴子安. 工程建筑物变形观测数据处理. 北京：测绘出版社，1989，2

3 上海师范大学数学系概率统计教研组. 回归分析及其试验设计. 上海教育出版社，1978，4～8

4 盛永懋，李蕙芬，钱君燕编译. 应用回归分析. 上海科学技术文献出版社，1989，4

5 陈希孺，王松桂. 近代回归分析. 合肥：安徽教育出版社，1987，7～17

6 王凤艳. 逐步回归分析预报大坝变形精度的探讨. 武汉测绘科技大学毕业设计论文，1992，12～13

The Independent Variablas-snooping in the Regression Analysis of Dam Deformation

Wu Zian

Abstract In this paper, the regression analysis, which is usually used in dam deformation analysis is discussed. It is shown that the number of selected independent rariables is less than it is real one, because of multi-collinearity between independent variables. In order to overcome effect by multi-collinearity, the author proposes some suggestions about independent variables-snooping.

Key word dam; deformation analysis; regression analysis; independent variables-snooping; multi-collinearity