第12卷 第4期　　大 气 科 学　　Vol. 12, No. 4
1988 年 12 月　　SCIENTIA ATMOSPHERICA SINICA　　Dec., 1988

# 试用最优子集与岭迹分析相结合的方法确定回归方程

俞善贤　　汪 铎
（浙江省气象科学研究所）（浙江师范大学地理系）

**提　　要**

本文针对逐步回归在气象业务应用中出现的一些问题，提出采用岭回归等方法加以改进．在太湖平原春粮产量的气象预报统计模式试验中，运用最优子集与岭迹分析相结合的方法，取得了较好的效果．

多元回归分析是气象业务中一个重要的工具，并取得了一定的成效，但也存在不少问题．比较突出的是对样本拟合好，而作外推时（实际预报应用）效果往往不理想，其原因取决于：（1）对所研究问题内在机制的认识程度和选取相应模型的正确程度；（2）数学处理技巧是否得当，如模式中变量的选入和回归系数估计方法的恰当程度等．近年来回归分析方法的研究进展较快，提出了一些新方法，企图改进古典的以最小二乘法为主体的回归分析．我们在气象业务应用中发现目前常用的回归方程有很多问题，有必要引用这些新方法，并在微型机上予以实现，从初步工作看是有意义的，本文对其中一些主要问题作初步评述和探讨．

## 一、逐步回归在应用中的一些问题

逐步回归方法是气象业务中最常用的方法之一，这个方法表面上看来比较"完美"，但在大量的实际应用中和理论上都发现有一定缺陷．

（1）逐步回归最大特点是在选入或剔除变量时都基于统计检验，常用的是$F$检验．但实际上正如一些统计学者所指出：除了某些特殊的、在应用上不太现实的条件外，不能认为所涉及的$F$检验是正确的[1]．因而从理论上并不能以任何概率保证所挑选自变量的"显著性"，亦即由逐步回归建立的方程很可能不满足实际应用（做预报）的要求．

（2）我们在实例中发现（见表1），a）选入和剔除的$F$值不是单调下降，有时会回升这个现象十分普遍；b）有时信度虽然较高，但选进因子的回归系数与经验不符，如因子单相关系数的符号与回归系数的符号相反；c）在实际使用中，使用者往往希望通过调节信度，使方程中含有的因子数与预定的一致（同样本数相匹配）．由于$F$值有跳动现象，信度

1985年11月4日收到，1986年8月23日收到再改稿．

**表 1　14 个样本作逐步回归试验分析**

| 因子个数 | 自变量序 | 剔除 $F$ 检验 | 选入 $F$ 检验 | 单相关系数 | 回归系数 |
|---|---|---|---|---|---|
| 1 | 8 | | 7.53 | | |
| 2 | 2 8 | −7.53 | 3.58 | | |
| 3 | 2 6 8 | −3.58 | 3.09 | $r_6 = 0.46$ | $\beta_6 = -103.5$ |
| 4 | 2 6 7 8 | −3.09 | 9.25 | | |
| 4 | 2 6 7 8 | −0.44 | | | |
| 3 | 2 6 7 | −21.19 | | | |
| 4 | 2 6 7 11 | | 10.7 | $r_{11} = -0.34$ | $\beta_{11} = 0.229$ |
| 5 | 2 3 6 7 11 | −10.7 | 3.59 | | |
| 6 | 2 3 6 7 10 11 | −3.59 | 2.65 | | |
| 7 | 2 3 5 6 7 10 11 | −2.65 | 8.74 | $r_5 = 0.367$ | $\beta_5 = 0.399$ |
| 8 | 2 3 5 6 7 9 10 11 | −8.74 | 4.31 | $r_9 = 0.122$ | $\beta_9 = -0.313$ |
| 9 | 1 2 3 5 6 7 9 10 11 | −4.31 | 21.01 | $r_1 = -0.26$ | $\beta_1 = 0.247$ |
| 10 | 1 2 3 4 5 6 7 9 10 11 | −21.01 | 7.66 | | |

注：表中单相关系数 $r$ 和回归系数 $\beta$ 的下标均是自变量序。

稍低时选进因子数太多，信度稍高时选进的因子又太少，需要反复调节试验(有时仍难以取得理想结果)大大增加了工作量.

(3) 如果单纯从回归方程残差平方和的观点看，逐步回归方法决定的子集残差平方和 RSSp 可能比包含同样自变量数的子集的 RSSp 大得多. 逐步回归方法只能提供一个子集回归，而对不同的因子数，这些子集变量又是包含关系，这样就很难对各个自变量的重要性及相互关系作出客观的分析.

下面给出一个虚构数组来说明.

$y$: 29 −48 18 −12 44 57 47 10 86 46

$x_1$: 7 −19 38 45 −5 15 −38 38 59 −27

$x_2$: 7 −12 39 49 −7 12 −40 39 53 −29

$y$ 与 $x_1$ 之间的单相关系数为 0.104，$y$ 与 $x_2$ 之间的单相关系数为 −0.00635，由于逐步回归每步只考虑一个因子对 Y 的作用，在本例情况下，$x_1$ 和 $x_2$ 很难入选方程. 选入 $x_1$ 时的 $F$ 值只有 0.08，由此将得出 $y$ 对 $x_1$ 和 $x_2$ 的回归没有必要的假象，然而 $y$ 对 $x_1$，$x_2$ 的回归方程复相关系数为 0.997，是高度相关的.

## 二、最优子集和岭回归方法

### 1. 最优子集方法

最优子集方法是随着计算机的发展，针对上节 (3) 中指出的问题应运而生的. 该方法是对各个因子进行组合，从中选出最优的方程[2]. 例如对 4 个因子有 1、2、3、4、12、

13、14、23、24、34、123、124、134、234、1234 十五种组合，从中选出最优的.

这种方法计算工作量大，内存量也大，十几个因子是可行的，超过二十个因子时计算时间和内存量的矛盾就极为突出. 针对气象业务需要，可预先指定方程含有因子数的范围，这样就不必对所有组合都运算，用"分支定界法"可大大节省计算机的时间和空间.

**2. 岭回归方法**

岭回归方法是针对回归系数估计的正确性提出的. $\hat{\beta}$ 作为 $\beta$ 的估计是否良好，应考虑 $\hat{\beta}$ 与 $\beta$ 的接近程度，可以用均方差 MSE 作为一个较好的数量指标. 即:

$$\mathrm{MSE}(\hat{\beta}) = E[\|\hat{\beta} - \beta\|^2] = E[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] \tag{2.1}$$

可以证明在线性模型 $Y = X\beta + e$ 下有

$$\mathrm{MSE}(\hat{\beta}) = \sigma^2 \,\mathrm{tr}(S^{-1}) \tag{2.2}$$

若假定 $e \sim N(0, \sigma^2 I)$，则

$$\mathrm{Var}(\|\hat{\beta} - \beta\|^2) = 2\sigma^4 \,\mathrm{tr}(S^{-2}) \tag{2.3}$$

由线性代数理论可知: $S = X^TX$ 是对称正定矩阵，其特征根皆为正数，设

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots\cdots \geqslant \lambda_t > 0$$

则 $S^{-1}$ 的特征根为 $\frac{1}{\lambda_i}$，$i = 1, 2, \cdots\cdots t$，于是得:

$$\mathrm{tr}(S^{-1}) = \sum_{i=1}^{t} \frac{1}{\lambda_i} \tag{2.4}$$

将 (2.4) 式代入 (2.2) 和 (2.3) 式得:

$$\mathrm{MSE}(\hat{\beta}) = \sigma^2 \sum_{i=1}^{t} \frac{1}{\lambda_i} \tag{2.5}$$

$$\mathrm{Var}(\|\hat{\beta} - \beta\|^2) = 2\sigma^4 \sum_{i=1}^{t} \frac{1}{\lambda_i^2} \tag{2.6}$$

由此可见，当自变量之间存在复共线关系时，即当 $S$ 呈现"病态"时，$\lambda_i$ 很接近 0，$\mathrm{MSE}(\hat{\beta})$ 和 $\mathrm{Var}(\|\hat{\beta} - \beta\|^2)$ 的值都很大，即 $\|\hat{\beta} - \beta\|^2$ 平均来说取值很大且很不稳定，因此就很难认为在这种情况下 $\hat{\beta}$ 仍然是 $\beta$ 的良好估计.

岭回归是 1970 年以来系统发展起来的一种改进最小二乘估计的方法.

设 $0 \leqslant k < \infty$ 称下式为岭回归

$$\hat{\beta}(k) = [\hat{\beta}_1(k), \cdots\cdots \hat{\beta}_t(k)]^T = (S + KI)^{-1}X^TY \tag{2.7}$$

当 $k = 0$ 时就是最小二乘估计.

据上分析，由 $S$ 的特征根 $\lambda_1 \cdots\cdots, \lambda_t$，可知 $S + KI$ 的特征根为

$$\lambda_1 + k, \cdots\cdots \lambda_t + k,$$

如果 $S$ 的最小特征根 $\lambda_t$ 很接近于 0，则 $\lambda_t + k$ 接近 0 的程度就会小些. 因而有理由期望 $\hat{\beta}(k)$ 比 $\hat{\beta}$ 有所改善，可以证明存在这样的 $k$，使 $\hat{\beta}(k)$ 均方误差比 $\hat{\beta}$ 的均方误差小，即

$$E[\|\hat{\beta}(k) - \beta\|^2] < E[\|\hat{\beta} - \beta\|^2] \tag{2.8}$$

并且使左边尽可能小. 当然要求出最优的 $k$ 值，实际计算就变得复杂起来，但当 $k$ 在

$(0, +\infty)$ 内变化时 $\hat{\beta}_j(k)$ 作为 $k$ 的函数，就描出一条曲线，这些曲线称为岭迹. 在实际应用中，可择取适当的一些 $k$ 值，由此点出的岭迹来分析变量之间的关系是很有意义的.

按 $\hat{\beta}(k) = (X^TX + KI)^{-1}X^TY$ 对每一个 $k$ 值需计算一次逆矩阵，计算量太大，下面给出计算岭迹的简单公式. 设线性回归模型为:

$$Y = X\beta + e \qquad E(e) = 0 \qquad V_{AR}(e) = \sigma^2 I \tag{2.9}$$

且设资料已经中心化和标准化处理. 记矩阵 $U$ 为:

$$U = \begin{bmatrix} Y^TY & Y^TX \\ X^TY & X^TX \end{bmatrix} \tag{2.10}$$

又 $A$ 为 $t+1$ 阶正交方阵致

$$A \cup A^T = \mathrm{diag}(\lambda_1, \lambda_2, \cdots\cdots \lambda_{t+1}) = \Lambda \tag{2.11}$$

则 $\hat{\beta}_j(k)$ 表示为:

$$\hat{\beta}_{j-1}(k) = -\left[\sum_{i=1}^{t+1} \frac{a_{i1}a_{ij}}{\lambda_i + k}\right] \Big/ \left[\sum_{i=1}^{t+1} \frac{a_{i1}^2}{\lambda_i + k}\right] \tag{2.12}$$
$$(j = 2, 3, \cdots\cdots, t+1)$$

此式表明一旦算出 $A$ 和 $\Lambda$ 就可方便算出岭迹，对每个不同 $k$ 需要重复计算矩阵.

### 3. 岭迹分析

在岭回归中，岭迹分析作为了解各变量的作用及相互关系的一种工具，有其重要作用和实用意义. 表 2 是不同岭迹的特点，参考了文献 [2] 给出的图形，我们用来分析古典回归和岭回归的不同情况.

这里扼要介绍确定 $k$ 的岭迹方法，该方法为了改变古典回归中看来不合理之处（数值、符号），基于直观考虑在岭迹图上可找到 $k_0$ 值，使回归系数估计达到稳定，即图中岭迹曲线趋于平稳(见图 1g)，并且希望残差平方和不增加太多. 具体原则是:

(1) 去掉岭回归系数稳定且绝对值较小的变量，这里的岭回归系数是可以直接比较

**表 2　古典回归与岭回归对变量的分析对比**

| 岭迹图形 | 图形特点 | 古典回归 | 岭回归 |
|---|---|---|---|
| 图 1a | $\hat{\beta}_j(0) = \hat{\beta}_j > 0$ 且比较大，$\hat{\beta}_j(k)$ 很不稳定，$k$ 从零开始显著下降，且迅速趋于零 | $X_j$ 对 $Y$ 有重要影响 | 由于随 $k$ 值的增加迅速趋于零，失去“预报能力”，$X_j$ 不起重要影响，甚至可以去掉这个变量 |
| 图 1b | $\hat{\beta}_j(0) > 0$ 且比较小，$\hat{\beta}(k)$ 比较稳定，从零开始变成负值 | $X_j$ 对 $Y$ 作用不大，且影响为正 | $X_j$ 对 $Y$ 有较显著影响，且影响为负. |
| 图 1c | $\hat{\beta}_j(0) = \hat{\beta}_j > 0$ 且比较大，当 $k$ 增加时迅速下降，且稳定为负值 | $X_j$ 对 $Y$ 有正影响，且作用较大 | $X_j$ 对 $Y$ 有负影响 |
| 图 1d | $\hat{\beta}_1(k)$ 和 $\hat{\beta}_2(k)$ 不稳定，但其和大体稳定 | $X_1$ 对 $Y$ 有正影响，$X_2$ 对 $Y$ 有负影响，$\hat{\beta}_1(0)$ 和 $\hat{\beta}_2(0)$ 符号相反 | $X_1$、$X_2$ 对 $Y$ 都是正影响，$X_1$、$X_2$ 有复共线关系，只留一个变量就行，对古典回归的符号相反提供解释 |
| 图 1e | 把所有的岭迹描在一张图上，岭迹不稳定，系统较乱 | | 有理由怀疑，古典估计的回归系数是否反映真实情况. |
| 图 1f | 迹线平稳，系统稳定 | | 对古典回归的估计具有信心 |
| 图 1g | 当 $k < k_0$ 时，迹线较乱，且不稳定，当 $k > k_0$ 时迹线稳定 | | 选择 $k_0$，以 $\hat{\beta}_j(k_0)$ 作为回归系数，系数较稳定 |

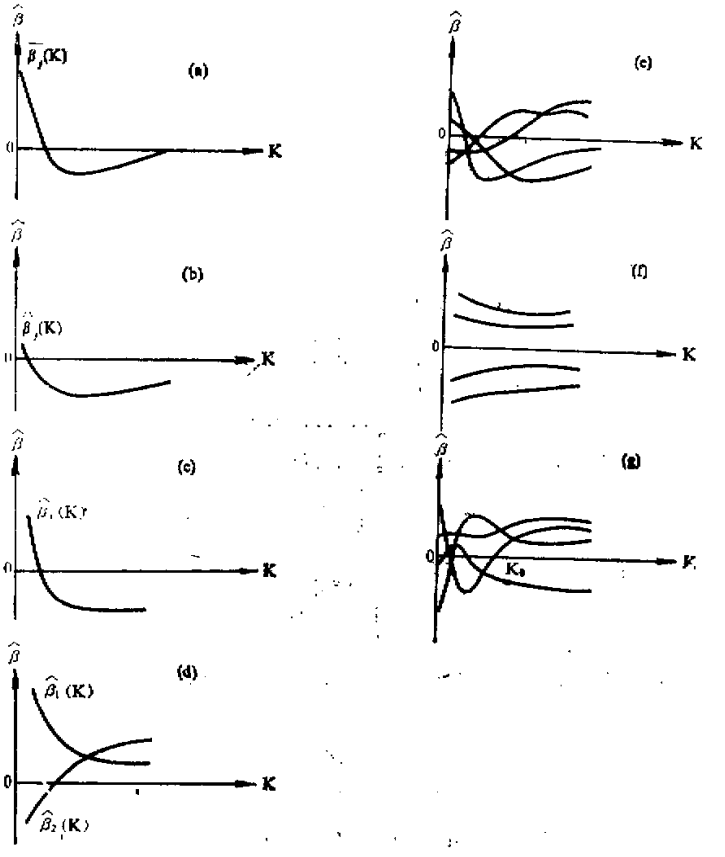


图 1 岭迹图形

大小的，因为已作中心化和标准化处理；

(2) 去掉岭回归系数不稳定又随着 $k$ 的增加迅速趋于零的变量；

(3) 去掉一个或若干个具有不稳定岭回归系数的变量，如果不稳定的岭回归系数很多，究竟去掉几个，去掉哪几个，并无一般规律可循，只有重新进行岭回归分析，由分析效果来决定.

应当指出：岭回归分析方法与由残差平方和作为指标的古典回归方法比较，在基本概念上是完全不同的两种方法. 前者作出方程的残差平方和比后者要大些，复相关系数、剩余标准差要差一些，这一点是两者考虑的目标不同所致.

## 三、应 用 实 例

针对气象业务中因子较多时，一开始用岭迹法来挑选因子，或许还带有主观性；同时

照顾到残差平方和，这一点从心理上常为人们所接受，所以我们先采用最优子集的方法确定因子，然后在众多的方程中用岭迹法来分析系数的稳定性，综合两者的优点，获得的方程较为满意.

我们在太湖平原春粮产量预报模式的究制中作了一些尝试[3]. 对 29 年样本输入 12 个因子进行最优子集回归和逐步回归两种方法的对比试验，两者确定的方程都作了岭迹分析.

12 个因子是：$X_1$、$X_4$、$X_5$、$X_1^2$、$X_3^2$、$X_1X_2$、$X_1X_4$、$X_2X_6$、$X_4X_5$、$X_3X_5$、$X_7X_9$.

$X_1$：上年 11 月至当年 2 月，四个月总雨量；

$X_2$：上年 10 月至当年 2 月下旬，旬雨量 $>10$ mm 的旬数；

$X_3$：2 月上旬至 3 月中旬雨量；

$X_4$：2 月上旬至 3 月中旬日照；

$X_5$：1 月的平均气温；

$X_6$：2 月的最低气温；

$X_7$：3 月的最低气温；

$X_8$：3 月中旬的平均气温；

$X_9$：北半球极涡中心强度，上年 12 月至当年 2 月，三个月的累积值.

用最优子集确定的因子为：

3 个因子：$X_1$、$X_5$、$X_1^2$（秩为 1）

4 个因子：$X_1$、$X_5$、$X_1^2$、$X_3X_5$（秩为 1）

5 个因子：$X_1$、$X_4$、$X_4X_5$、$X_3X_4$、$X_7X_9$（秩为 1）

逐步回归确定方程的因子为：

3 个因子：$X_3^2$、$X_1X_4$、$X_7X_9$（秩为 2）

4 个因子：$X_1$、$X_1^2$、$X_3^2$、$X_7X_9$（秩 $>5$）

5 个因子：$X_1$、$X_1^2$、$X_4X_5$、$X_3X_5$、$X_7X_9$（秩 $>5$）这里的秩指残差平方和从小到大排序的位置. 对最优子集和逐步回归确定的 4 个因子的方程作岭迹分别见图 2、图 3. 这两个岭迹都比较稳定，图 3 中 $X_7X_9$ 因子的数值较小，作用不大，可以去掉这个因子. 5 个因子的方程中上述两种方法都选进了 $X_7X_9$ 因子，它的作用同样也不大，故 5 个因子的方程也不宜选用.

逐步回归建立的方程为：

$$y_w = 213.1 - 0.4647X_1 - 0.001607X_4^2 - 0.001978X_3^2 + 0.4337X_7X_9$$

$$R = 0.87 \qquad S_y = 26$$

最优子集建立的方程为：

$$y_w = 147.8 - 0.4593X_1 + 26.86X_5 - 0.001432X_1^2 - 0.1582X_1X_5$$

$$R = 0.89 \qquad S_y = 24.7$$

取 $k = 0.1$，岭回归方程为：

$$y_w = 125.5 - 0.4134X_1 + 19.05X_5 - 0.001078X_1^2 - 0.1067X_1X_5$$

三个方程对 1983、1984、1985 年进行试报，1983 年相对误差分别为：1.62%、3.69%、2.89%；1984 年相对误差分别为：18.8%、9.95%、7.38%；1985 年相对误差分别为：

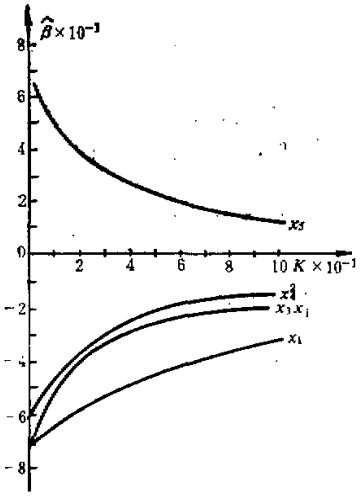


图 2 最优子集确定方程的岭迹

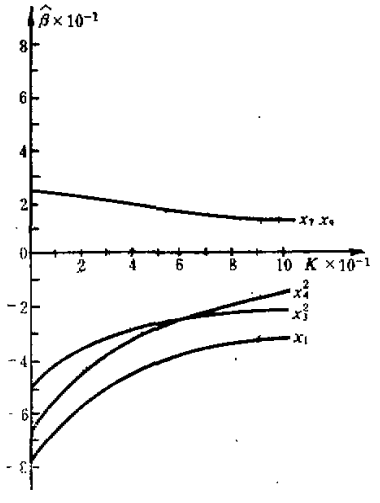


图 3 逐步回归确定方程的岭迹

7.86%、5.88%、4.00%. 从试报三年来看，岭回归方程预报能力较强，且比较稳定.

## 四、讨 论

(1) 我们认为：对回归方程预测能力问题，仅用拟合的残差平方和来衡量预测能力的好坏是不够的，应该用多个目标函数来衡量，从最优子集方法建立的众多的方程中找出各项性能较好的方程来，这样的预测能力是会提高的，对方程进行岭迹分析是其中的手段之一. 在以上的实例中 $X_1X_3$ 这个因子，在逐步回归建立的 3 个、4 个、5 个因子的方程中一直保留下来，显得很重要，但在岭迹图上反映出它并非十分重要.

(2) 在有的方程中岭迹不稳定，所建立的方程会产生单相关系数与回归系数符号相反的问题，这自然联想到偏相关系数，偏相关系数是由回归系数确定的，当因子间存在复共线关系时，回归系数的估计偏离较大，这时算出的偏相关系数是否会失去原有的意义？是否还可信？这些问题是值得进一步讨论的.

## 参 考 文 献

[1] 陈希孺、王松桂编著，1984，近代实用回归分析，广西人民出版社.
[2] Hocking, R. R., 1981，回归变量最好子集的选择，《数字计算机上用的数学方法》第二篇，上海科学技术出版社.
[3] 汪铎、俞善贤，1985，太湖平原春粮产量预报统计模式试验研究，浙江师范大学学报(自然科学版)第二期.