

逐步回归的性能分析

中央气象台 裴国庆

为了适应天气预报客观化、量化的需要，自六十年代以来，统计预报获得了迅速发展，目前已成为一个独立的预报体系。从理论、计算方法和预报效果来看，逐步回归是统计预报中比较成熟的一个预报方案。但是，在预报实践中，却常常发现效果时好时坏。因而有人对这个方法的性能产生怀疑。为什么会造就这种情况呢？关键在于气象问题的处理，如果气象问题处理得当，能够适应逐步回归方法的性能，就会取得比较好的预报效果；反之，效果会比较差。本文就逐步回归的性能和如何恰当地处理气象问题进行一些讨论。

一、逐步回归的三个特点

逐步回归也叫“逐步线性回归”。“逐步”、“线性”、“回归”分别描述出逐步回归方法的三个主要特点，它既说明了这个方法的优越性也反映出其局限性。

1. 回归方法可以描述多数样本的预报量与预报因子间的定量关系。所以，在多数情况下，用这个关系作预报都能获得较好的预报效果。少数情况下，也可能出现失败，但这是一般预报方法所共有的局限性。有人认为用回归方法不能作异常天气的预报，这种看法也不完全正确。所谓异常天气，系指在历史样本中没出现过或出现机率极小的天气现象。从理论上讲，回归方法是能够预报异常天气的，只要预报因子出现异常，就可以预报出预报量异常。图1清楚地表明了这一点，一旦出现 $X_{\text{异常}}$ ，根据回归直线就能报出 $Y_{\text{异常}}$ 。例如，在长江流域春播季节，我们曾经用连阴雨日数的回归预报方程计算出历史上从未出现过的预报值，而且与实况基本一致。但是，一般情况下，不符合回归规律的少数样本经常出现在回归直线的两端，因此，出于下面将要讨论的非线性原因造成异常天气预报的效果较差。我们相信，只要非线性问题处理得好，一定能够改善这个结果。

就平均情况来看，

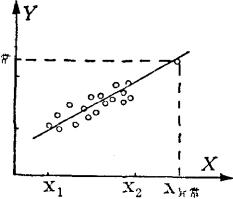


图 1

预报值（拟合值）比实况值稍小。为了说明这一点，先讨论一下统计量方差的物理意义。在统计学上，方差表示平均偏离平均值的程度，也就是平均离散程度。预报量 y 的方差表示 y 的变化大小，即所要预报信息的多少，而预报值（拟合值） \hat{y} 的方差表示回归方程能够预报出的信息的多少。回归方程的复相关系数：

$$R = \frac{\sum_{i=1}^m (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^m (y_i - \bar{y})^2} = \frac{\sigma_{\hat{y}}}{\sigma_y}$$

其中， m 为样本数。由于 $R < 1$ ，所以，回归方程所计算出的预报值（拟合值）的变化幅度比预报量实况值的变化幅度小。通常， $R > 0.90$ ，甚至 0.99，说明两者的差值很小。个别样本预报值（拟合值）也可能大于（或小于）历史样本的最大（或最小）值。

2. 回归方法能够揭示出预报量与预报因子间的线性关系。逐步回归的线性假设使得计算简单，并且在数学上能获得完满解决（非线性回归至今在数学上仍不能解决）。在气象问题中，虽然严格的线性关系很少，但在预报量和预报因子取值的一定范围内，多数可以近似看作线性关系。图 2 给出 $y = ax^2$ 在 (x_1, x_n) ， (y_1, y_n) 定义域内的近似线性关系。当 x 落在历史样本取值范围内时，预报（拟合）误差很小； x 落在历史样本取值范围之外时，预报（拟合）误差 $y - \hat{y}$ 将会加大。而对那些不能近似看成线性关系的问题，目前用逐步回归方法寻找其内在联系尚存在一定困难。这正是该方法的局限性之一。

由于线性假定，逐步回归方法不能充分考虑到两个或几个因子相互配合而发挥的作用。譬如，某地区受北槽影响而产生降水（不管南涡是否存在），降水量平均只有 3 毫米，如果受南涡影响而产生降

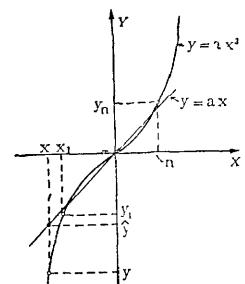


图 2 函数 $y = ax^2$ 在 (y_1, y_n) 定义域内的近似线性关系

水（不管北槽是否存在），降水量平均有 15 毫米，一旦北槽与南涡相互配合，该地区降水量立即可以达到 50 毫米。在这种情况下，用逐步回归方程作预报，只能报出 18 毫米。报不出两个因子相互配合后所造成的降水量剧增。

因为，在逐步回归计算过程中，只要引入一个新的因子，先前引入因子对预报量的贡献将会改变。假设，在第 k 个因子引入前，已引入方程和未引入方程的因子对 y 的贡献为：

$$v_i^0 = \frac{(r_{in}^0)^2}{r_{ii}^0}$$

第 K 个因子引入后，各因子对 y 的贡献为：

$$v_i = \frac{(r_{in})^2}{r_{ii}} \quad (1)$$

而

$$r_{in} = r_{in}^0 - \frac{r_{ik}^0 \times r_{kn}^0}{r_{kk}^0},$$

$$r_{ii} = r_{ii}^0 - \frac{(r_{ik}^0)^2}{r_{kk}^0}$$

所以

$$v_i = (r_{in}^0 - \frac{r_{ik}^0 \times r_{kn}^0}{r_{kk}^0})^2 / [r_{ii}^0 - \frac{(r_{ik}^0)^2}{r_{kk}^0}] \quad (2)$$

r_{ij} 代表相关矩阵的元素，当 $j=n$ 时， r_{in} 表示 x_i 和 y 之间的相关系数； $i=j$ 时， r_{ii} 表示第 i 个因子的方差。从 (2) 式看出，第 K 个因子引入后，第 i 个因子对 y 贡献发生变化的原因可以分成两部分。一部分，由 (2) 式分子表示。由于第 i 个因子与第 k 个因子相关，同时第 k 个因子又与 y 相关，致使第 i 个因子对 y 贡献的一部分包括在第 k 个因子对 y 的贡献中。

造成对 y 贡献改变原因的另一部分，由 (2) 式分母表示。由于第 K 个因子与第 i 个因子相关，使第 i 个因子的方差减少，因此，第 i 个因子对 y 的贡献也随之变化。换言之，如果 $r_{ik}^0=0$ ，那么 $V_i=V_i^0$ ，即第 i 个因子与第 k 个因子相互独立，第 k 个因子的引入将不会改变第 i 个因子对 y 的贡献。综上所述，问题的关键就在于两个因子存在相关性。然而，两个因子的相关与两个因子相互配合所发挥的作用，就如同前面所举例子中，北槽出现和南涡出现的相关性与两个天气系统相互配合所起作用一样是完全不同的两回事。说明逐步回归方法不能充分反映出两个或几个因子相互配合而起的作用，这是该方法的又一局限性。

3. 逐步回归方法采用了逐步引入和剔除的方式挑选因子。能够定量地选择预报因子是逐步回归方法的

最大优越性。然而，这种挑选因子的方式，实质上是一种普查，因此常常会引入假指标或错误地考虑了预报因子和预报量之间的虚假数量关系。这是由于在逐步回归中虽然使用了 F 检验，但是，这里使用的这种普查并不符合 F 检验中应用的“小概率事件在一次试验中不出现”这条统计学原则的前提，而是进行了几次试验。结果使引入方程的预报因子的可信度降低。

是否可以通过提高 α 水平而增加预报因子的可信度呢？假如一次试验的信度为 α ，则 n 次试验的信度只有 $K \times n \times \alpha$ ，其中 K 表示由于 n 个因子间不独立而引入的订正系数。当样本数取 100， K 取 0.80 时，对 30 个因子进行普查，即使 F 提高到 6.90，其信度也不过只有 $0.80 \times 30 \times 1\% = 24\%$ 。而一次试验时，信度却可达到 1%。如果再提高 F 水平，必然引入的因子数目极少甚至不能引入任何因子。所以，事实上，这个途径是行不通的。

关于普查引起假指标这一问题曾有许多人研究过。对一些简单的普查尚可以从理论上予以解决，逐步回归普查中所引入的假指标问题比较复杂，目前还没有一个解决的办法。Cund (1970) 提出对特定的一组样本和因子用数值模拟的方法判断逐步回归预报方程的可靠程度，即“Monte Carlo”有效性检验。美国国家飓风中心用 NHC-73 系统建立台风移动预报方程时，曾利用这个方法进行有效性检验。其中，样本数取 127，因子数取 120，选用因子数为 12。对 120 个因子任意选择 12 个因子的方法共有 C_{120}^{12} 种，假如随机选取 100 种，每种选法可以建立一个回归方程，其复相关系数为 R ，100 个 R 值可看作随机变量，通常呈正态分布。令 S_R 表示 R 的方差， \bar{R} 为 R 的平均值，那么。

$$R_{0.95} = \bar{R} + 1.645S_R$$

如逐步回归方程的复相关系数 $R^* > R_{0.95}$ ，则认为这个回归方程不是随机的，其可信度为 95%。

二、使用逐步回归方法时，气象问题的处理

气象问题的处理是为了适应逐步回归方法的性能，以便更好地反映预报量与预报因子间的内在联系。这对预报效果的好坏将起决定性作用。气象问题的处理包括样本的选择、预报对象的描述、预报因子的选取和处理以及预报方程的建立等。这些问题的处理是相互联系的，选择样本时，必须考虑到预报量和预报因子将如何处理，在描述预报量时，也要想到样本和预报因子应如何选择。下面分别予以介绍。

1. 样本选择：

选择样本时，首先应该注意样本数目不能过少，

一般取 100 个左右，以利提高 F 水平从而增加选用因子的可靠程度，减少假指标、假关系出现的可能性。 F 水平和样本数目的关系可以表示成：

$$F_k = \frac{\Delta Q^k}{Q/m - N - 1} \quad (3)$$

这里， F_k 表示第 k 个因子的 F 值， $\Delta Q^k = Q^{k-1} - Q^k$ 表示引入第 k 个因子后 y 的剩余方差减少， Q^k 是引入第 k 个因子后的剩余方差。从 (3) 式可以看出， F_k 和样本数 m 成正相关。样本愈多， F 值愈大。

另外，逐步回归方程是表示多数样本的预报因子和预报量间的内在联系，所以选择样本时，应尽量选取规律相同的样本。如果样本中掺杂了一些不同规律的样本，不但不能在回归方程中得到反映，而且会破坏多数样本的定量关系（图 3）。然而，由于我们对预报问题所具有的规律性仍不完全掌握，故确定样本规律是否相同更有困难。目前，只能采用下面一些方法，尽可能使选用样本的规律相同。

(1) 同一季节：一般，不同季节的天气形势、形势演变、系统活动规律以及产生的天气现象都有所不同。因此，样本应在一个季节内选取。至于季节如何恰当的划分则取决于具体预报对象。

(2) 地理分区：天气系统活动规律和天气现象都具有明显的区域性，位于 20°N 以北的台风和 20°N 以南的台风其移动规律差别很大；南海台风

生成规律与太平洋地区也有所不同，这些都说明许多天气学问题都存在清楚的地理区分。当然，某些情况下，这种界线不十分突出而已。为了考虑地理位置所造成的天气规律的差异，也可以采用连续分区法。美国台风移动预报的 NHC-73 模式就利用了这种方案。

(3) 环流分型：大气环流形势与天气系统的生、消以及活动都有密切关系。所以，常常把环流形势归纳成各种类型，对不同环流型建立不同的逐步回归方程。

(4) 确定起报条件：

这种方法用途极广。有时，可以把某系统的强度界线定作起报条件。譬如台风路径预报，由于处在热低压阶段的活动规律很难掌握，故常常决定只有达到台风强度时，才开始制作台风路径预报。这个规定即

称起报条件。可以选择某种天气系统通常活动规律可能趋于一致的一个地区警界线，作为起报条件。

以上这些做法都是广大气象工作者实践中总结出来的宝贵经验，一般通称之为“分类”预报，其目的是把不同规律的样本分成各种样本组，尽可能使同类样本组的样本规律基本相同。从而提高回归方程的预报效果。但是，经分类以后，每种样本组的样本数将会减少，必然导致方程可信度下降。这个矛盾还需在实践中合理解决。

2. 预报对象的描述

利用逐步回归方法制作预报时，首先要选择 1 个或几个预报量，用以描述预报对象。选择预报量时，必须考虑与可利用的资料有关的预报量的可预报性。为了避免由于预报量中包括了较多不可预报的信息而使计算过程中引入假指标，还要对预报量做预处理。预处理的办法很多，如参数化、分类法、时间和空间平均等等。

一般，台风中期路径预报不象短期预报那样，把每 12 小时（或 6 小时）经、纬向位移作为预报量，而首先将台风路径用二次曲线展开。那么只要报出二次曲线的系数或者台风平均移向、平均曲率，则台风路径的形状基本确定。这种处理办法，即所谓参数化。预报量经过参数化处理以后，预报量可预报性提高。

天气预报中，天气现象的定量化预报常常是很困难的。为了提高这种预报量的可预报性，我们可以对预报量作合理的分类处理。所选分类方案既要符合统计学原理又要兼顾服务需要。目前，分类方案居多，聚类分析中的极差分割法就是其中之一。这里再介绍一种与前者分类原则不同的方法——频数分布图分类法。分类时，首先确定“界值”，“界值”附近的样本数要求比较少，以避免由于分类而造成界值附近样本

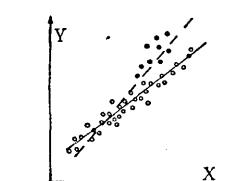


图 3 由于掺杂了少数不同规律的样本，回归直线发生变化。实线：掺杂前的回归线，虚线：掺杂后的回归线。实心点为不同规律的样本。

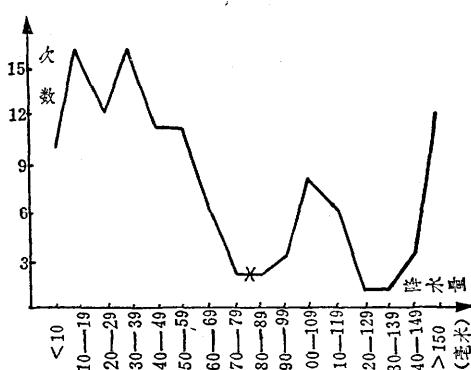


图 4 1951—1970 年 7—8 月北京旬降水量频数分布图

间差异人为扩大。根据这个原则，可以利用频数分布图确定出恰当的“界值”。图4给出1951—1970年7—8月北京旬降水量的频数分布图。图中“X”表示选取的界值。有些人把平均值选作界值，分成正、负距平两类。但是，对正态分布的要素来说，平均值附近样本频数最高，显然，这种分类方法是不够恰当的。

降水量预报也经常采用区域平均或者时间滑动平均的处理办法。这样可以去掉小尺度系统以及地形的影响，使其可预报性增大。

再有，选择预报量时，还应考虑线性化问题。譬如，在台风移动预报中，选择预报量的方案可能有两个，一个是用每6小时台风经、纬向位移作为预报因子，另一个是选预报时刻台风经、纬度作预报量。明显地，前一种方案选出的预报量容易与预报因子建立线性关系；而后者，预报时刻台风所在经、纬度是台风位移与目前台风所在经、纬度的函数，即：

预报时刻台风经纬度 = f (位移、目前台风经度、纬度)

线性关系比较差。因此，预报时选用第一个方案为宜。

同样，在预报量处理时，也存在线性问题。一般地，通过对预报量作变量变换，减少非线性关系。比如，作降水量预报时，可以把降水量开平方，使 y 与 x 成线性关系。

3. 预报因子的选取和处理：

在逐步回归预报中，预报因子的选择是很关键的一环。这项工作企图用一些简单的统计量把具有预报价值的信息反映出来。（1）式也表明预报因子所包含的信息中，与预报量有关的部分应尽可能多，而与 y 无关的信息应尽可能减少。虽然，完成这个任务不容易，但是只要按照逐步回归方法的性能，处理得当，就可以取得较好的效果。

为了达到这个目的，选取预报因子时，应注意如下几点：

1) 由于逐步回归方程仅仅描述了多数样本的预报因子和预报量的关系。所以，必须选用与多数样本有贡献的预报因子。一些因子只对少数样本有预报意义，而对多数样本却没有价值，那么，在逐步回归中必然落选。有时，这样做也许会失去一些有预报价值的信息，但只要对它们加以适当地处理就可以弥补这一损失。

假设 x_1 对某些样本 n_1 有预报意义； x_2 对另一些样本 n_2 有预报意义……，为了不漏掉任何有预报价值的信息，我们对预报因子 x_1, x_2, \dots, x_n 实行逻辑加法，得到一个复合预报因子：

$$x = x_1 \vee x_2 \vee x_3 \vee \dots \vee x_n$$

当 $x_1, x_2, x_3, \dots, x_n$ 中任一指标出现时，都可以认为 x 指标出现。这样，经过处理后的复合预报因子 x 则对多数样本具有预报意义。

2) 逐步回归方程仅仅反映了预报因子和预报量间的线性关系。所以，对预报因子和预报量间的非线性关系必须作变量变换。假如 x 和 y 的关系不十分确定，可根据气象学知识推断出几种可能的函数关系 $f_i(x)$ ，对这些函数关系作变量变换 $z_i = f_i(x)$ 并且按照一个比较高的 F 水平用逐步回归方法进行挑选。最后，对选出的最佳函数关系再作变量变换 $z = F(x)$ ，以实现 y 与 z 近于线性关系。

3) 在逐步回归挑选（普查）过程中，常常会引入一些假指标或假关系。所以，必须提高预报因子的质量，同时减少参加挑选的因子数目。

首先，预报因子的描述要力求简练，一些气象意义比较明确的因子，如锋区强度、地转风速、西风指数……能够利用高度差描述就不要用两点或几点的高度值代替；再者，在因子普查以前，应对可能的预报因子作主客观分析，选取相关好、独立性强的因子参加挑选，以避免盲目性；适当的区域平均、时间平均可以去掉因子随机成分，以提高因子稳定性；还要注意，预报因子对预报量要有明显针对性。譬如，降水预报，由于地理位置的差异，往往降水成因存在各种不同。因此，在作不同地区预报时，虽然都是降水预报，但对不同地区选择的因子应有所侧重，不能用完全相同的一批因子参加挑选。

4. 预报方程的建立

采用多少预报因子建立预报方程为适宜，这个问题从理论上很难回答。业务实践中，许多人认为可以取样本数的5%，当然并非绝对。总之，尽管适当提高 F 水平，少取一些预报因子可能使拟合结果稍差，也比预报因子选得过多能获得较好的预报效果。

三、结 论

逐步回归方法的三个特点：

1. 描述大多数样本的平均关系；
2. 线性假设；
3. 普查挑选因子。

这三个特点不仅反映出这个方法的优越性，也指明了方法的局限性。但是，一旦我们真正地掌握了它的性能，对气象问题做出恰当地处理，逐步回归方法就会充分发挥其优越性并弥补其不足，从而成为获得定量化、客观化预报的一个有效的统计学方法。

由于引入回归预报方程中的预报因子是经普查而得到的，所以，在某些情况下，方程的天气学意义仍然很不明确。