第2卷第2期　　气象科学研究院院刊　　Vol.2, No.2
JOURNAL OF ACADEMY OF METEOROLOGICAL
1987年10月　　SCIENCE, S. M. A., CHINA　　Oct., 1987

回归预报方程优良性的判断

杨洪昌　　杨喜寿　　徐淑云
（山东省气象台）　（山东大学）　（山东省气象台）

提　　要

本文列举了基于回归剩余平方和Q的几个自变量选择准则，重点讨论了一种着重预报精度的变量选择准则，并从预报角度对其进行了讨论和比较。

一、引　　言

逐步回归是目前国内气象业务部门使用较多的一种多元回归计算方法。这种方法的不足之一是控制选入和剔除变量的F_α值的选择没有客观标准，因此在使用该方法组建回归预报方程时，经常选取不同的F_α值，得到几个可供选用的方程，但并没有办法确定使用哪一个方程好。本文则给出一种合理的准则予以判断，以便为解决这方法的不足作一尝试。

二、判断回归预报方程优良性的常用准则

记预报量为y，预报因子为x_1，x_2，……x_n，将其观测数据写成矩阵形式

$$X=\begin{pmatrix} x_{11}\,x_{12}\cdots\cdots x_{1m} \\ x_{21}\,x_{22}\cdots\cdots x_{2m} \\ \vdots \qquad\qquad \vdots \\ x_{n1}\,x_{n2}\cdots\cdots x_{nm} \end{pmatrix}$$

$$y'=(y_1 y_2\cdots\cdots y_m)$$

其中，x_{it}是第i个因子的第t次观测值，y_t是预报量的第t次观测值。

本文1986年4月28日收到，1986年6月24日收到修改稿。

令

$$\bar{x}=\begin{matrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_n\end{matrix},\quad \bar{x}_i=\frac{1}{m}\sum_{t=1}^{m}x_{it},\qquad \bar{y}=\frac{1}{m}\sum_{t=1}^{m}y_t$$

$$X_0=\begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_1\cdots\cdots x_{1m}-\bar{x}_1\\ x_{21}-\bar{x}_2 & x_{22}-\bar{x}_2\cdots\cdots x_{2m}-\bar{x}_2\\ \vdots & \vdots \quad\quad \vdots\\ x_{n1}-\bar{x}_n & x_{n2}-\bar{x}_n\cdots\cdots x_{nm}-\bar{x}_n\end{pmatrix}$$

$$\boldsymbol{y}_0'=(y_1-\bar{y}\ \ y_2-\bar{y}\cdots\cdots y_n-\bar{y})$$

最小二乘方意义下的回归方程为

$$\hat{y}=\bar{y}=\boldsymbol{b}'(\boldsymbol{x}-\bar{\boldsymbol{x}}) \tag{1}$$

其中$\boldsymbol{x}'=(x_1x_2\cdots\cdots x_n)$是预报因子，$\boldsymbol{b}'=(b_1b_2\cdots\cdots b_n)$是回归系数，由下式确定

$$\boldsymbol{b}=(X_0X_0')^{-1}X_0\boldsymbol{y}_0 \tag{2}$$

设(1)就是由逐步回归方法算出来的一个回归方程，判断其优良性，传统使用的准则和做法是：

1．计算回归剩余平方和$Q=\sum_{t=1}^{m}(\hat{y}_t-y_t)^2$，其中，$\hat{y}_t=\boldsymbol{b}'(\boldsymbol{x}_t-\bar{\boldsymbol{x}})+\bar{y}$。$Q$越小，表示回归拟合越好；或考察与$Q$等价的统计量复相关系数$R=\sqrt{1-\frac{Q}{S_{yy}}}$，其中$S_{yy}=\sum_{t=1}^{m}(y_t-\bar{y})^2$，R越接近于1，认为回归方程越好。

但是，这些观点，在一定的假设下，通过F-检验，已确定地体现在逐步回归计算过程之中[1]。如果由取不同的F_α值得到两个不同的回归方程，肯定地含因子个数多者Q较小(等价地R较接近于1)。因此，如果按Q越小越好的准则来选择因子子集，则毫无疑问，应选择全部因子。因为方程中含的因子个数越多，Q越小。

另外，Q越小越好的想法，是从试验结果与统计模型的拟合程度出发的。当问题在于用一个简单的公式来概括观测结果时，这是适用的。但考虑到统计性质，拟合最好的模型并不一定统计性质也最好。因此Q（与其等价地R）并不能作为选择因子子集的判据量。

2．为了克服Q的不足，防止选进一些不必要的因子，通常用剩余标准差$\hat{\sigma}_{y,1,2,\cdots n}=\sqrt{\frac{1}{m-n-1}Q}$作为选择因子子集的判据量。

由于用逐步回归方程做预报时，通常需要假设y服从正态分布，其预报值的95%的置信区间为$\hat{y}\pm1.96\,\hat{\sigma}_{y,1,2,\cdots\cdots n}$，因此，可用$\hat{\sigma}_{y,1,2,\cdots\cdots n}$表示回归方程的优良性。$\hat{\sigma}_{y,1,2,\cdots\cdots n}$越小，表示预报量95%的置信区间越小，方程的预报能力越强。

判断回归预报方程的优良性，使用$\hat{\sigma}_{y,1,2,\cdots\cdots n}$比直接使用Q和R更为合理。因为它不仅表示了回归方程的拟合能力，而且在一定程度上体现了其预报性能。

三、一种回归方程优良性判断准则

回归方程优良性的判据量，可通过从m样本中除去一个样本，用余下的m－1个样本建立回归方程，试报除去的样本而得到。

例如将X划去第k列，相应地y划去第k个元素。

令

$$\bar{x}_i^{(k)}=\frac{1}{m-1}\sum_{\substack{t=1\\t\neq k}}^{m}x_{it},\quad(i=1,2,\cdots\cdots n),\quad \bar{y}^{(k)}=\frac{1}{m-1}\sum_{\substack{t=1\\t\neq k}}^{m}y_t,$$

并记

$$\bar{X}^{(k)}=\begin{pmatrix}\bar{x}_1^{(k)}\\ \bar{x}_2^{(k)}\\ \vdots\\ \bar{x}_n^{(k)}\end{pmatrix}$$

$$X_k=\begin{pmatrix}x_{11}-\bar{x}_1^{(k)} & x_{12}-\bar{x}_1^{(k)} & \cdots\cdots & x_{1\,k-1}-\bar{x}_1^{(k)} & x_{1\,k+1}-\bar{x}_1^{(k)} & \cdots\cdots & x_{1m}-\bar{x}_1^{(k)}\\ x_{21}-\bar{x}_2^{(k)} & x_{22}-\bar{x}_2^{(k)} & \cdots\cdots & x_{2\,k-1}-\bar{x}_2^{(k)} & x_{2\,k+1}-\bar{x}_2^{(k)} & \cdots\cdots & x_{2m}-\bar{x}_2^{(k)}\\ \vdots & \vdots & & \vdots & \vdots & & \\ x_{n1}-\bar{x}_n^{(k)} & x_{n2}-\bar{x}_n^{(k)} & \cdots\cdots & x_{n\,k-1}-\bar{x}_n^{(k)} & x_{n\,k+1}-\bar{x}_n^{(k)} & \cdots\cdots & x_{nm}-\bar{x}_n^{(k)}\end{pmatrix}$$

$$y_k'=(y_1-\bar{y}^{(k)}\ y_2-\bar{y}^{(k)}\cdots\cdots y_{k-1}-\bar{y}^{(k)}\ \ y_{k+1}-\bar{y}^{(k)}\ \cdots\cdots y_m-\bar{y}^{(k)})$$

由X_k，y_k得到最小二乘方意义下的回归方程为

$$\hat{\hat{y}}-\bar{y}^{(k)}=b_k'(x-\bar{x}^{(k)}) \tag{3}$$

其中，回归系数

$$b_k=(X_kX_k')^{-1}X_ky_k \tag{4}$$

方程(3)对除去的那个样本y_k的预报值为

$$\hat{\hat{y}}_k=b_k'(x_k-\bar{x}^{(k)})+\bar{y}^{(k)} \tag{5}$$

对于所有$k=1,2,\cdots\cdots m$重复以上过程，定义：$Q^*=\sum_{k=1}^{m}(\hat{\hat{y}}_k-y_k)^2$为方程(1)的预报偏差平方和。$Q^*$越小，则方程（1）的预报能力越强。

实际上Q^*可以在计算回归方程(1)的同时得到，不必建立另外m个方程（即留一个样本出来的回归方程）。证明如下：

记x_k为X的第k列，y_k为y的第k次观测值。

由于，

$$\bar{x}_i^{(k)}=\frac{1}{m-1}\sum_{\substack{l=1\\l\neq k}}^{m}x_{il}=\frac{m}{m-1}\bar{x}_i-\frac{1}{m-1}\bar{x}_i+\frac{1}{m-1}\bar{x}_i\frac{1}{m-1}x_{ik}$$

$$=\bar{x}_i-\frac{1}{m-1}(x_{ik}-\bar{x}_i)\quad(i=1,2,\cdots\cdots n)$$

$$\bar{y}^{(k)}=\bar{y}-\frac{1}{m-1}(y_k-\bar{y})$$

因此有：

$$X_k X_k'=\sum_{j=1}^{m}(\mathbf{x}_j-\bar{\mathbf{x}}^{(k)})(\mathbf{x}_j-\bar{\mathbf{x}}^{(k)})'-(\mathbf{x}_k-\bar{\mathbf{x}}^{(k)})(\mathbf{x}_k-\bar{\mathbf{x}}^{(k)})'$$

$$=\sum_{j=1}^{m}\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'+\frac{1}{m-1}(\mathbf{x}_k-\bar{\mathbf{x}})\sum_{j=1}^{m}(\mathbf{x}_j-\bar{\mathbf{x}})'$$

$$+\left[\sum_{j=1}^{m}(\mathbf{x}_j-\bar{\mathbf{x}})\right]\frac{1}{m-1}(\mathbf{x}_k-\bar{\mathbf{x}})'+\frac{m}{(m-1)^2}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})'$$

$$-\frac{m^2}{(m-1)^2}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})'$$

$$=\sum_{j=1}^{m}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'-\frac{m}{m-1}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})'$$

$$=X_0X_0'-\frac{m}{m-1}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})'\qquad(6)$$

同理，

$$X_k\mathbf{y}_k=X_0\mathbf{y}_0-\frac{m}{m-1}(\mathbf{x}_k-\bar{\mathbf{x}})(y_k-\bar{y})\qquad(7)$$

可以证明[2]：

$$(X_kX'_k)^{-1}=(X_0X'_0)^{-1}+\frac{m}{m-1}\cdot\frac{(X_0X_0')^{-1}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})'(X_0X_0')^{-1}}{1-\frac{m}{m-1}(\mathbf{x}_k-\bar{\mathbf{x}})'(X_0X_0')^{-1}(\mathbf{x}_k-\bar{\mathbf{x}})}\qquad(8)$$

令，$\mu=\frac{m}{m-1}$，$\lambda_k=(\mathbf{x}_k-\bar{\mathbf{x}})'(X_0X'_0)^{-1}(\mathbf{x}_k-\bar{\mathbf{x}})$　　(9)

将(7)、(8)代入(4)，并注意到(2)得：

$$\mathbf{b}_k=\mathbf{b}-\mu(X_0X_0')^{-1}(\mathbf{x}_k-\bar{\mathbf{x}})(y_k-\bar{y})+\frac{\mu}{1-\mu\lambda_k}(X_0X_0')^{-1}(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})'\mathbf{b}$$

$$-\frac{\mu^2\lambda_k}{1-\mu\lambda_k}(X_0X_0')^{-1}(\mathbf{x}_k-\bar{\mathbf{x}})(y_k-\bar{y})\qquad(10)$$

(10)代入(5)两边同减去y_k得：

$$\hat{\hat{y}}_k-y_k=\mathbf{b}'(\mathbf{x}_k-\bar{\mathbf{x}}^{(k)})-\mu(y_k-\bar{y})'(\mathbf{x}_k-\bar{\mathbf{x}})'(X_0X_0)^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}^{(k)})$$

$$+\frac{\mu}{-\mu\lambda_k}\mathbf{b}'(\mathbf{x}_k-\bar{\mathbf{x}})(\mathbf{x}_k-\bar{\mathbf{x}})'(X_0X'_0)^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}^{(k)})$$

$$-\frac{\mu^2\lambda_k}{1-\mu\lambda_k}(y_k-\bar{y})'(\mathbf{x}_k-\bar{\mathbf{x}})'(X_0X_0')^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}^{(k)})+\bar{y}^{(k)}-y_k$$

$$=\mu\mathbf{b}'(\mathbf{x}_k-\bar{\mathbf{x}})-\mu^2\lambda_k(y_k-\bar{y})-\frac{\mu^2\lambda_k}{1-\mu\lambda_k}\mathbf{b}'(\mathbf{x}_k-\bar{\mathbf{x}})$$

$$-\frac{\mu^3\lambda^2_k}{1-\mu\lambda_k}(y_k-\bar{y})+\bar{y}^{(k)}-y_k$$

$$=\frac{\mu}{1-\mu\lambda_k}\mathbf{b}'(\mathbf{x}_k-\bar{\mathbf{x}})-\frac{\mu^2\lambda_k}{1-\mu\lambda_k}(y_k-\bar{y})-\mu(y_k-\bar{y})$$

$$=\frac{\mu}{1-\mu\lambda_k}\mathbf{b}'(\mathbf{x}_k-\bar{\mathbf{x}})-\frac{\mu}{1-\mu\lambda_k}(y_k-\bar{y})$$

$$=\frac{\mu}{1-\mu\lambda_k}(\hat{y}_k-y_k)$$

因此有：

$$Q^*=\sum_{k=1}^{m}\left(\frac{\mu}{1-\mu\lambda_k}\right)^2(\hat{y}_k-y_k)^2 \tag{11}$$

我们还证明了$\left|\frac{\mu}{1-\mu\lambda_k}\right|\geqslant 1$，且随方程中因子数的增加而增大（证明略）。

由(11)可见，当回归方程含有一定量的因子后再增加因子，Q^*将逐渐增大。Q^*可以作为判断回归方程优良性的判据量，其出发点是基于对未来的一次预报。选取不同的F_a值，用逐步回归建立的所有方程中，Q^*的最小值所对应的方程即为最佳预报方程。

四、实 例 分 析

我们用上年九月至当年三月北半球500百帕月平均高度预报山东省夏季(六至八月)降水量距平百分率，从14个初选因子中，通过改变F_a值，共建含不同因子数的9个方程，资料年代为1952—1984年。下边以含三个因子（x_5，x_9，x_{14})的方程为例说明Q^*的计算方法

表1中第一行为资料年序号，第二至第四行分别为第五、九、十四个因子的距平值。其因子距平资料矩阵为：

$$X_0=\begin{pmatrix} x_{51}-\bar{x}_5 & x_{52}-\bar{x}_5 & \cdots\cdots & x_{5m}-\bar{x}_5 \\ x_{91}-\bar{x}_9 & x_{92}-\bar{x}_9 & \cdots\cdots & x_{9m}-\bar{x}_9 \\ x_{141}-\bar{x}_{14} & x_{142}-\bar{x}_{14} & \cdots\cdots & x_{14m}-\bar{x}_{14} \end{pmatrix}=\begin{pmatrix} 4.73 & 3.73 & \cdots\cdots & 0.73 \\ -5.93 & 3.07 & \cdots\cdots & 7.07 \\ -6.63 & -3.63 & \cdots\cdots & 22.37 \end{pmatrix}$$

计算得：

表 1　Q^* 计算表

k	1	2	3	4	5	……	33
$x_{5k}-\bar{x}_5$	4.73	3.73	5.73	0.73	8.73	……	0.73
$x_{9k}-\bar{x}_9$	−5.93	3.07	−2.93	−5.93	5.07	……	7.07
$x_{14}-\bar{x}_{14}$	−6.63	−3.63	−2.63	−31.63	−8.63	……	22.37
λ_k	0.04	0.02	0.03	0.14	0.10	……	0.09
$\hat{y}_k$	−16	−2	−15	0	−3	……	−4
y_k	−29	16	−1	−12	9	……	3
$(y_k-y_k)^2$	169	324	196	144	144	……	49
$\left(\frac{\mu}{1-\mu\lambda_k}\right)^2(\hat{y}_k-y_k)^2$	195.5	359.2	222.0	209.2	190.4	……	63.4

$$X_0X_0' = \begin{Bmatrix} 1052.24 & -206.45 & -2.72 \\ -206.45 & 1075.87 & -589.72 \\ -2.72 & -589.72 & 9419.63 \end{Bmatrix}$$

$$(X_0X_0')^{-1} = \begin{Bmatrix} 0.000973 & 0.000120 & 0.000007 \\ 0.000120 & 0.000614 & 0.000038 \\ 0.000007 & 0.000038 & 0.000108 \end{Bmatrix}$$

$$\lambda_1 \doteq (x_{51}-\bar{x}_5\ x_{91}-\bar{x}_9\ x_{141}-\bar{x}_{14})(X_0X_0')^{-1}\begin{pmatrix} x_{51}-\bar{x}_5 \\ x_{91}-\bar{x}_9 \\ x_{141}-\bar{x}_{14} \end{pmatrix}$$

$$= (4.73\quad -5.93\quad -6.63)\begin{pmatrix} 0.000973 & 0.000120 & 0.000007 \\ 0.000120 & 0.000614 & 0.000038 \\ 0.000007 & 0.000038 & 0.000108 \end{pmatrix}\begin{pmatrix} 4.73 \\ -5.93 \\ -6.63 \end{pmatrix}$$

$$= 0.04$$

同理可以算出 λ_2，λ_3，…… λ_m 列入表 1 第五行。第六、七两行分别为方程计算值和实况值；本例中，$m=33$，因此 $\mu=\frac{m}{m-1}=1.031$，将表 1 中最后一行相加即得 Q^*。

表 2　山东省夏季降水距平百分率各预报方程的判断统计量

方程含因子数	1	2	3	4	5	9	11	13	14
Q	9623	7840	6337	5796	5578	4981	4904	4829	4806
R	0.62	0.71	0.77	0.79	0.80	0.82	0.83	0.83	0.83
$\hat{\sigma}_{y.1,2,\cdots\cdots n}$	17.62	16.17	14.78	14.39	14.37	14.72	15.28	15.94	16.34
Q^*	10779	9165	7698	7725	8153	11881	13735	19491	21378
1985年试报	−12	−9	−20	−8	−27	−5	0	−1	2

表 2 对于含因子数不同的各个方程，列出了各种优劣判断统计量。可见，Q 随方程中因子数 n 的增加而减少(相应地 R 增大)，但方程中含四个因子后再增加因子，Q 的变化就比较小了；$\hat{\sigma}_{y.1,2,\cdots\cdots n}$ 随 n 的增大先是减小，当 $n=5$ 时取最小值，然后随 n

的增大而增加，据此判断，应选用含5个因子的方程；随着n的增大，Q^*的变化趋势与$\hat{\sigma}_{y.1,2,\cdots n}$类似，但在$n=3$时取最小值。因此，以选含三个因子的方程为最好。

表2最后一行为各方程对于山东省一九八五年夏季降水量距平百分率的预报值，其实况值为－17，可见以含三个因子的方程预报效果最好，含全部初选因子的方程预报效果最差。

另外，针对山东省各区域七至九月各月降水量距平百分率预报，共有8个预报量，初选因子数8—16个，用逐步回归得到8组方程，用上述方法分别选择最优方程对1985年进行试报，经与实况比较，其试报效果以Q^*作为判据量选择的预报方程为最好，$\hat{\sigma}_{y.1,2,\cdots n}$次之，$Q$（等价地$R$）最差。

五、几 点 说 明

1．Q和R只能表明回归方程的拟合能力，不宜作为回归预报方程优良性的判据量。

2．$\hat{\sigma}_{y.1,2,\cdots n}$虽然在一定程度上反映了回归方程的预报能力，但它也属基于Q的判断准则，且必须以预报量为正态分布的假设为前提。

3．不必对预报量的分布做任何假设，Q^*即可表示回归方程的预报性能，同时Q^*亦可在建立回归方程的同时得到。因此，可以认为Q^*是判据回归预报方程优良性的较好判据量。

参 考 文 献

[1] A.拉尔斯登，H·S·维尔夫等，徐献瑜等译，数字计算机上用的数学方法，上海科技出版社，1963.

[2] N.C.Giri, Multivariate Statistical Inference, *Academic Press*, 1977.

[3] 张尧庭、方开泰，多元统计分析引论，科学出版社，1983.

ASSESSMENT OF THE EFFECTIVENESS OF REGRESSION EQUATION

Yang Hongchang

(Meteorological Observatory of Shandong Province)

Yang Xishou

(University of Shandong)

Xu Shuyun

(Meteorological Observatory of Shandong Province)

Abstract

Based on the sum of regression residual square (Q) and the criteria for choosing some of undependent variables, a choice criterion for improving the forecast accuracy is discussed. Also, the effectiveness of these criteria is examined and compared on the basis of skill scores in the forecast experiments.