# 相关比和线性相关系数的对比分析

朱光宇　胡成群　留小强
（江西省气象科研所）

**提　要**

本文对相关比和线性相关系数进行了理论比较，并用实际资料对两者的相关评价特点进行了计算对比，分析结果表明相关比确实是一种优于线性相关系数的相关度量：1）它不对相关形式作任何要求，用它来反映相关比线性相关系数更为全面；2）在相关比计算的过程中，我们可以同时得到相关程度和相关形式两项信息。另外，我们还就相关比计算中的问题作了讨论，指出了应该注意的问题。

## 一、引言

统计预报中一个重要问题．是如何评价因子和预报量之间的相关程度。目前在我国，因子评价多以线性相关系数为度量。然而，当因子与预报量之间为非线性相关时，仅用线性相关系数去度量，则势必会歪曲因子真实的相关程度。为此，近年来有许多工作致力于讨论非线性相关度量，例如采用将因子或预报对象先进行函数变换再来计算线性相关系数的方法[1—2]。变换通常采用某类特定的解析函数形式。这样的做法对那些相关形式不符合这些解析函数特征的因子，仍然不能正确反映其相关程度。我们认为文献[3]中的相关比，能适合于各类因子的相关度量评价，是一个较好的非线性相关度量。J．N．佩格尔（1975）应用相关比来挑选降水预报因子，取得了较好的效果[4]。我们在MOS预报研制工作中，应用相关比进行因子分析，并在此基础上提出了一种新的非线性模式——相关比回归模型*。本文首先对相关比和线性相关系数进行理论分析，随后讨论相关比的一些计算问题，并将相关比和线性相关系数的相关评价特点进行实际计算分析。

## 二、相关比和线性相关系数理论

随机变量X和Y之间的线性相关系数$\rho_{xy}$为

$$\rho_{xy}=\frac{E[(X-E[X])(Y-E[Y])]}{\sqrt{E[(X-E[X])^2]E[(Y-E[Y])^2]}} \tag{1}$$

在已知N个样本$(X_i,Y_i)$时，其估计值$r_{xy}$为：

$$r_{xy}=\frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2\sum_{i=1}^{N}(Y_i-\bar{Y})^2}} \tag{2}$$

Y依X的相关比$\eta_{yx}$定义为[3]

$$\eta_{yx}=\sqrt{1-\frac{E[(Y-E[Y|X])^2]}{E[(Y-E[Y])^2]}} \tag{3}$$

由等式

$$E[(Y-E[Y])^2]=E[(E(Y|X]-E[Y])^2]+E[(Y-E[Y|X])^2] \tag{4}$$

可以得到相关比的另一等价表达式

$$\eta_{yx}=\sqrt{\frac{E[(E[Y|X]-E[Y])^2]}{E[(Y-E[Y])^2]}} \tag{5}$$

$\rho_{xy}$和$\eta_{xy}$均满足$|\rho_{xy}|\leqslant 1$，$|\eta_{yx}|\leqslant 1$。

$\rho_{xy}$和$\eta_{yx}$具有很自然的内在联系，现略加分析。对于随机变量X和Y，如要通过X的观测值来估计Y，可选择一个函数g(x)，以$\hat{Y}=g(X)$来估计Y，而以均方偏差

*朱光宇等，相关比回归模型（待发表）。

$E[(Y-\hat{Y})^2]=E[(Y-g(X))^2]$ 来反映 $g(X)$ 对Y的估计精度。显然，这一估计的好坏与选取的函数$g(x)$有关。称$g(x)$为X对Y的一个估计函数。如果将$g(x)$限制在线性函数类中，求出使均方误差达到最小的最优线性估计函数$\beta_0+\beta_1 x$，该函数的均方误差为$E[(Y-\beta_0-\beta_1 x)^2]$，而线性相关系数$\rho_{xy}$可写成

$$\rho^2_{xy}=1-\frac{E[(Y-\beta_0-\beta_1 x)^2]}{E[(Y-E[Y])^2]} \quad (6)$$

由此可见，线性相关系数反映了用X的线性函数估计Y时所能达到的最优估计精度。显然$\rho_{xy}$的局限性在于将估计函数$g(x)$作了线性约束，所以用它作为相关度量是不全面的，因为若Y不能用X的线性函数估计得很好，这时表现为线性相关系数很小，但却不能说X和Y之间一定相关不密切。因为有可能用一个X的非线性函数得到Y的更好估计。文献[2]拓宽了线性函数类的限制而采用$a+bx^A$的函数形式，发展了一种非线性相关度量，尽管它比线性相关系数能更全面一些，但对估计函数$g(x)$仍作了相当的约束。

若不对估计函数$g(x)$作任何限制，即在一切估计函数中选取使均方偏差取最小的最优估计函数$\psi(x)$，理论上可以证明[5]：这个函数存在而且就是Y依X的条件期望函数$\psi(x)=E[Y\mid X=x]$。它的均方偏差为$E[(Y-\psi(X))^2]=E[(Y-E[Y\mid X])^2]$，对照相关比的定义（3）式可知：$\eta_{yx}$正是反映了用最优估计函数$\psi(x)=E[Y\mid X=x]$去估计Y时的估计精度。由此可见：相关比能全面评价因子和预报量的相关程度，而线性相关系数是相关信息中线性部分的评价；线性相关系数所能反映的相关信息必定能为相关比度量，反过来则不成立，这在数学上反映出来就是$\eta_{xy}^2\geqslant\rho_{xy}^2$，两者差别的大小取决于最优估计函数$\psi(x)$偏离线性的程度。而$\psi(x)$的函数特性则反映了X与Y之间的具体相关特征。

## 三、相关比估值计算方法

可以直接从（2）式去计算线性相关系数的估计值$r_{xy}$，若要从（5）式相关比定义来估计相关比值却有如何从有限样本来估计条件期望函数$E[Y\mid X=x]$的问题。在样本数较大时，可以采用下面离散化的简单估值法。

首先，我们将因子X在其取值的数轴上依它的概率分布密度，划分为H个相互独立的连续区间，记为$h_1,\cdots,h_H$，对任何X的取值x，必有而且仅有一个区间$h_r$，使$x\in h_r$。现将N个样本$(X_i, Y_i)$依$X_i$值分属的区间分成对应的H组。每组内所属样本数分别记为$N_1, N_2\cdots, N_H$，满足$N=\sum_{i=1}^{H}N_i$，N为总样本数。现在对每个区间的因子分布概率$W_i=P(X\in h_i)$和区间条件期望$E[Y\mid X\in h_i]$通过样本进行估值，其值设为$\hat{W}_i$和$\hat{E}_i$ $(i=1,\cdots,H)$。离散化原则是用下面的阶梯函数

$$f(x)=\hat{E}_i \quad (当x\in h_i) \quad (7)$$

来近似代替真正的条件期望函数$E[Y\mid X=x]$。按照（5）式和（7）式，可得到相关比估值公式

$$\hat{\eta}^2_{yx}=\frac{\sum_{i=1}^{H}\hat{W}_i(\hat{E}_i-\bar{Y})^2}{\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2} \quad (8)$$

上式中$Y_{ij}$表示第i个区间$h_i$中第j个样本的预报量值$\bar{Y}$，$=\frac{1}{N}\sum_{i=1}^{H}\sum_{i=1}^{N_i}Y_{ij}$为Y的总体平均值。由（8）式可见，用不同的方法估计$W_i$和$E[Y\mid X\in h_i]$时，所得到的相关比估值也有不同，在样本数较多时，可采用下面的估计方法

$$\hat{W}_i=N_i/N,\ \hat{E}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}Y_{ij}\ (i=1,\cdots,H) \quad (9)$$

（8）式可相应地写成

$$\hat{\eta}^2_{yx}=\frac{\sum_{i=1}^{H}N_i(\bar{y}_i-\bar{y})_2}{\sum_{i=1}^{H}\sum_{j=1}^{N_i}(Y_{ij}-\bar{y})^2},\quad \bar{y}_i=\frac{1}{N_i}\sum_{j=1}^{N_i}Y_{ij} \tag{10}$$

特别，当Y为只取0、1时，若记$\hat{P}$为Y=1出现的气候概率估值，$\hat{P}_i$为$P(Y=1\mid X\in h_i)$的估值（$i=1,\cdots,H$），则与（8）式对应的公式为

$$\hat{\eta}^2_{yx}=\frac{\sum_{i=1}^{H}\hat{W}_i(\hat{P}_i-\hat{P})^2}{\hat{P}(1-\hat{P})} \tag{11}$$

简单情况下，$\hat{W}_i$，$\hat{P}_i$和$\hat{P}$可作与（9）式相对应的估值：$\hat{W}_i=N_i/N$，$\hat{P}_i=M_i/N_i$，$\hat{P}=M/N$。其中M为N个样本中Y=1出现的总次数，$M_i$为在第i区间$h_i$的$N_i$个样本中Y=1出现的次数，$M=\sum_{i=1}^{H}M_i$。

我们在进行因子与降水的实际相关分析时，一般都将降水量作了0、1化处理，采用（11）式来进行计算，并且用阶梯函数

$$f(x)=\hat{P}_i \quad (当x\in h_i) \tag{12}$$

来具体分析因子与预报量之间的相关形式特征。

在相关比估值中，区间划分得是否合适，直接影响到（9）、（7）式作的阶梯函数逼近真值$E[Y\mid X=x]$的精度和可靠性，因而也影响到用有限样本去估计变量之间相关比的优良性。理论上，由于

$$E[Y\mid X=x]=\lim_{\varepsilon\to 0}E[Y\mid X\in(x-\varepsilon,\ x+\varepsilon)] \tag{13}$$

故需要将每个区间划分到充分地小，这样才能不致因区间划分过大而对$E[Y\mid X=x]$作过多的平滑，以保证用阶梯函数（9）代替$E[Y\mid X=x]$时有足够的精度。

另一方面，我们的估值是建立在有限样本上的，它反映的是在给定条件X=x时，Y的统计特性。因此，为了保证每个区间的估计值$\hat{E}_i$具有统计稳定性，就必须使每个区间划分得大些，使其含有足够的样本数。否则所估计的$\hat{E}_i$值就会受个别随机样本的影响而产生太大的波动，失去了统计上的意义，相应的相关比估值也会产生虚假的高相关。鉴于上述，在区间划分时，需要根据具体情况，统筹权衡这两个方面。一般地说，为保证估值的统计意义，我们取每个区间所包含的样本数一般须在10—15个以上，在X的概率分布密度较大的地方区间划分得小些，而在分布密度较小的地方则划分得大些。

## 四、相关比和相关系数的实际对比分析

本节以两个具体实例来说明整个计算，并将计算结果与上述理论分析结果加以比较。取两个预报因子和一个预报量

$X_1$——08时桂林700hPa实测风向；

$X_2$——08时长沙850hPa和700hPa温度露点差的两层平均值。

Y——抚州站未来24小时晴雨（20—20时）。

预报量Y取0、1两值，Y=1表示有雨（降水量R≥0.1mm），Y=0表示无雨。分析的样本取自1972—1985年5—6月的逐日资料。分别计算出因子$X_1$和$X_2$与预报量Y的相关比及线性相关系数值。相关比按(11)式计算。

表1和表2分别列出了因子$X_1$和$X_2$在整个相关比计算中的各项参数：$M_i$，$N_i$，$\hat{P}_i$和$\hat{W}_i$。$X_1$划分了11个区间组（H=11），$X_2$划分了10个区间组（H=10），在表中都给出了它们的具体划分范围。$\hat{P}_i=M_i/N_i$，$\hat{W}_i=N_i/N$，分别表示区间的降水条件概率和因子分布概率估计值。表3给出了相关比和线性相关系数的最终计算结果。

对于因子$X_1$，相关比值和线性相关系数具有很大的差异，相关比$\hat{\eta}_{yx}=0.442$，线性相关系数$r_{xy}=0.111$。这说明桂林的700hPa风向与抚州站的晴雨具有较显著的相关

表 1　因子$x_1$的相关比计算参数分布　　$N=708, \hat{P}=0.63, H=11$

| 项 \ 区间 | $h_1$ | $h_2$ | $h_3$ | $h_4$ | $h_5$ | $h_6$ | $h_7$ | $h_8$ | $h_9$ | $h_{10}$ | $h_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $M_i$ | 3 | 7 | 6 | 4 | 10 | 28 | 42 | 185 | 122 | 28 | 9 |
| $N_i$ | 20 | 22 | 19 | 12 | 23 | 47 | 62 | 226 | 163 | 74 | 40 |
| $\hat{P}_i$ | 0.15 | 0.32 | 0.32 | 0.33 | 0.43 | 0.60 | 0.68 | 0.82 | 0.75 | 0.38 | 0.22 |
| $\hat{W}_i$ | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.07 | 0.09 | 0.32 | 0.23 | 0.10 | 0.06 |
| 区间范围 | <45° | ≥45° <90° | ≥90° <135° | ≥135° <150° | ≥150° <180° | ≥180° <200° | ≥200° <225° | ≥225° <250° | ≥250° <270° | ≥270° <315° | ≥315° |

表 2　因子 $X_2$ 相关比计算参数分布　　$N=725,\ \hat{P}=0.63,\ H=10$

| 项 \ 区间 | $h_1$ | $h_2$ | $h_3$ | $h_4$ | $h_5$ | $h_6$ | $h_7$ | $h_8$ | $h_9$ | $h_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $M_i$ | 66 | 93 | 101 | 70 | 61 | 22 | 23 | 18 | 5 | 0 |
| $N_i$ | 69 | 109 | 133 | 107 | 107 | 57 | 58 | 60 | 24 | 1 |
| $\hat{P}_i$ | 0.96 | 0.85 | 0.76 | 0.65 | 0.57 | 0.39 | 0.40 | 0.30 | 0.21 | 0.00 |
| $\hat{W}_i$ | 0.10 | 0.15 | 0.18 | 0.15 | 0.15 | 0.08 | 0.08 | 0.08 | 0.03 | 0.00 |
| 区间范围 | ≥0.0 <1.0 | ≥1.0 <2.0 | ≥2.0 <3.0 | ≥3.0 <4.0 | ≥4.0 <6.0 | ≥6.0 <8.0 | ≥ 8.0 <12.0 | ≥12.0 <18.0 | ≥18.0 <25.0 | ≥25.0 |

表 3　相关计算结果

| 项目 \ 因子 | $x_1$ | $x_2$ |
|---|---|---|
| 相关比 | 0.442 | 0.444 |
| 线性相关系数 | 0.111 | −0.411 |

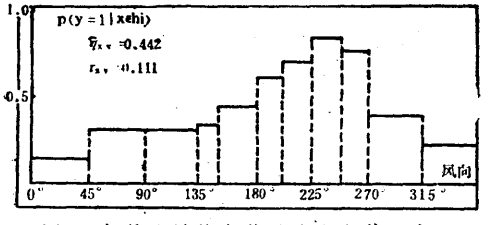


图1　相关比计算中估计降水条件概率 $P(Y=1|X_1=x)$ 的阶梯函数 $\hat{P}_i$（当$x\in h_i$），$i=1,\cdots,H$

性，但这部分相关信息不能由线性相关系数所度量。图 1 的阶梯函数是在相关比计算中得到的降水条件概率 $P(Y=1\mid X_1=x)$ 的估计曲线。我们可以从该函数的分布特征来分析因子 $X_1$ 与 Y 的相关形式特点。由图 1 可见，$X_1$ 与 Y 的相关具有非常明显的非线性特点。将风向按0—360°展开，整个阶梯函数呈抛物线型分布，这样的降水条件概率分布与天气学经验是一致的。当桂林风向为0—180°和270—360°之间时，降水概率均在0.5以下，不利抚州有降水发生；而当风向位于180—270°区间时，降水条件概率在 0.6 到 0.82之间，为抚州站未来24小时出现降水的有利条件；特别当风向为225—250°之间时，降水条件概率高达0.82，而落于这一区间的样本却占总样本的32%（226/708）。相关比中正确反映了这一相关信息。由于用任何一条直线来取代图 1 中的阶梯函数都会产生很大的偏差，因此，线性相关系数对该因子与 Y 的相关失去了评价能力。

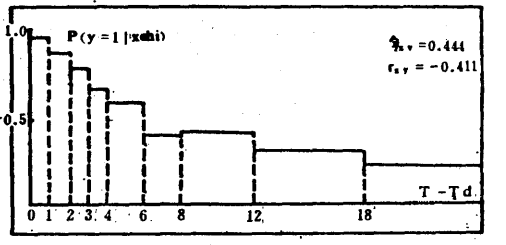


图2　降水条件概率 $P(Y=1|X_2=x)$ 的估值阶梯函数 $\hat{P}_i$（当$x\in h_i$），$\cdots,H$

由表3可见，因子$X_2$与Y的相关比和线性相关系数计算的量值相当，$\hat{\eta}_{yx}=0.444$，略大于线性相关系数的绝对值（0.411）。从描述$P(Y=1 \mid X_2=x)$分布特征的阶梯函数（图2）可以看出：因子$X_2$与Y之间的相关基本上是线性的。这说明当主要呈线性相关时，相关比也能对这样的因子相关信息作出正确的评价。

## 五、结束语

由上面的理论和计算分析可见：在评价因子与预报量的相关程度时，使用相关比要比用线性相关系数或一些以解析函数进行因子变换的非线性相关度量为佳。

在相关比计算过程中，我们可以得到象表1和图1那样的参数分布表和相关特征图，批量计算时可由计算机统一输出。这些信息可以帮助我们透彻地了解因子的具体相关特点，这在选择统计模型和因子处理时显得非常重要。

本文介绍的相关比算法对于处理离散型因子也非常适用，尤其当因子是只取状态符号而无数量和次序之分的变量时（如云图因子取：晴空、少云、多云和阴），只需将计算中的区间与因子状态符号相对应，就可计算该因子的相关比，而在线性相关系数中对这样的因子却难以正确处理。

当样本数较少时，用（10）式估计的相关比值易受分区间中样本的随机性影响而出现虚假的增值。一种有效的改进可能是相关比的近邻权估值方法。

## 参考文献


〔1〕王双一，预报对象和因子间的相关性探讨，《气象》，1987，第7期。

〔2〕冯耀煌等，最优化方法在天气预报中的应用，《气象》，1987，第8期。

〔3〕么枕生，《气候统计学基础》，p235—240，科学出版社，1984。

〔4〕Paegle等，根据一种图象识别法作降水概率预报，《统计天气预报译文集》，农业出版社，1979。

〔5〕周概容，《概率论与数理统计》，p304，高等教育出版社，1985。


# A comparison between the correlation ratio and the linear correlation coefficient


Zhu Guangyu　Hu Chengqun　Liu xiaoqiang

(Institute of Meteorological Science, Jiang Xi Province)


## Abstract


In this paper, a comparison between the correlation ratio and the linear correlation coefficient are given, and their characteristics in evaluating correlation are considered by computing them with practical data. The analysis shows that correlation ratio has advantage over linear correlation coefficient and other correlation measure.