相关系数和回归系数的空间几何解释

常州工学院理学院　姚　俊

[摘　要] 将变量的样本观察值看作空间向量，从空间几何角度阐明了样本线性相关系数和回归系数的意义和作用。
[关键词] 相关系数　回归系数　几何解释

变量间除了确定性的函数关系之外，还可能是不确定性的相关关系，相关分析和回归分析就是处理变量与变量之间关系的重要统计方法。相关分析主要测定变量之间关系的密切程度和变化方向，回归分析在相关分析和因果关系分析的基础上建立回归模型描述变量之间具体的变动关系，其中最基本的理论是处理两个变量的简单相关分析和一元回归分析。简单相关分析通过相关系数反映两个变量线性相关程度的强弱，一元回归分析通过回归系数说明解释变量对因变量具体影响的大小，但在教学中，学生较难理解相关系数和回归系数的统计意义和作用，本文旨在从几何角度更直观的揭示这两个统计概念的本质。

一、线性相关系数的几何解释

在相关分析中，变量 X 和 Y 的相关程度用英 Pearson 提出的积矩相关系数 ρ_{XY} 来衡量，计算公式为

$$\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\cdot\sigma_Y}$$

其中，σ_{XY} 为变量 X 和 Y 的协方差，σ_X、σ_Y 分别是变量 X 和 Y 的标准差。然而在实际中，由于变量的确切分布往往是未知的，无法直接计算 ρ_{XY}，所以只能通过变量 X 和 Y 的随机样本的观察值 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_n$ 计算样本相关系数 r，来反映变量 X 和 Y 的相关程度。样本相关系数的计算公式为：

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \quad (1)$$

其中，$\bar{x}=\frac{1}{n}\sum x_i$，$\bar{y}=\frac{1}{n}\sum y_i$ 为样本均值。

将样本观察值 $x_i,y_i(i=1,2,\cdots,n)$ 看作是 n 维欧氏空间 R^n 中的两个向量 $\vec{x}=(x_1,x_2,\cdots,xn)^T$ 和 $\vec{y}=(y_1,y_2,\cdots,yn)^T$，并记 $\vec{x'}=\vec{x}-\bar{x}\vec{e}$，$\vec{y'}=\vec{y}-\bar{y}\vec{e}$，其中 $\vec{e}$ 为 n 维单位向量 $(1,1,\cdots,1)^T$，那么在 R^n 中向量 $\vec{x'}$，$\vec{y'}$ 的夹角 $\langle\vec{x'},\vec{y'}\rangle$ 的余弦为

$$\cos\langle\vec{x'},\vec{y'}\rangle=\frac{\vec{x'}^T\vec{y'}}{\|\vec{x'}\|\cdot\|\vec{y'}\|}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

此结果与 (1) 式右端相等，即

$$r=\cos\langle\vec{x'},\vec{y'}\rangle \quad (2)$$

(2) 式表明变量 X 和 Y 的样本相关系数在几何上就是样本观察值向量的分量之间的夹角的余弦，如图 1 所示。

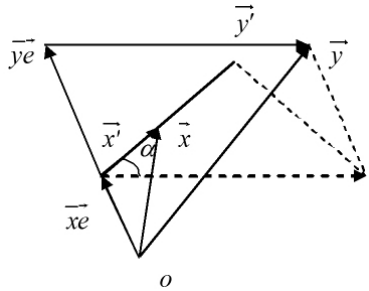

图 1　相关系数的几何解释

(图中夹角 α 的余弦即为变量间的相关系数)

由于 $|\cos\langle\vec{x'},\vec{y'}\rangle|\leq1$，这和 $|r|\leq1$ 是一致的。当夹角 $\langle\vec{x'},\vec{y'}\rangle$ 是锐角时，变量 X 和 Y 正相关，夹角越小正相关程度越强；当夹角 $\langle\vec{x'},\vec{y'}\rangle$ 是钝角时，变量 X 和 Y 负相关，夹角越大，负相关程度越强；当夹角 $\langle\vec{x'},\vec{y'}\rangle$ 是直角时，变量 X 和 Y 不相关。

相关系数只能说明变量间的线性相关关系而不能说明非线性关系，这也可以从几何上来解释。当 $|r|=1$，即 $\cos\langle\vec{x'},\vec{y'}\rangle=\pm1$ 时，$\langle\vec{x'},\vec{y'}\rangle$ 为 0 或 π，此时向量 $\vec{x'}\parallel\vec{y'}$，那么根据向量平行的充要条件：一定存在常数 $k\neq0$ 使得 $\vec{y'}=k\vec{x'}$，即 $\frac{y_i-\bar{y}}{x_i-\bar{x}}=k$，这说明样本观察值 $x_i,y_i(i=1,2,\cdots,n)$ 都落在直线 $y_i-\bar{y}=k(x_i-\bar{x})$ 上，变量 X 和 Y 具有完全的直线关系。若 $|r|$ 越接近于 1，那么向量 $\vec{x'}$ 几乎平行于 $\vec{y'}$，从而样本观察值 $x_i,y_j(i=1,2,\cdots,n)$ 落在直线 $y_i-\bar{y}=k(x_i-\bar{x})$ 附近，说明变量 X 和 Y 具有较强的线性关系。

二、回归系数的几何解释

要了解具有线性相关关系的变量之间具体的变动关系，在因果关系分析的基础上，可建立一元回归模型：$Y=\beta_0+\beta_1X+\varepsilon$

其中 β_1 为回归常数，在统计上表示解释变量 X 变化一单位时引起的因变量 Y 的平均变动单位。用最小二乘法可得回归系数 β_1 的估计值为

$$\hat{\beta}_1=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}$$

对上式变形可得

$$\hat{\beta}_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

$$=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}\cdot\frac{\sqrt{\sum(y_i-\bar{y})^2}}{\sqrt{\sum(x_i-\bar{x})^2}}$$

$$=\frac{\|\vec{y'}\|\cos\langle\vec{x'},\vec{y'}\rangle}{\|\vec{x'}\|}$$

上式分子为 $\vec{y'}$ 在 $\vec{x'}$ 上的投影 (图 1 中粗线所示)，即

$$\hat{\beta}_1=\frac{\mathrm{pro}_{\vec{x'}}\vec{y'}}{\|\vec{x'}\|} \quad (3)$$

(3) 式表明：回归系数 β_1 的估计值在几何上表示因变量 Y 的样本观察值向量的分量 $\vec{y'}$ 在解释变量 X 的样本观察值向量的分量 $\vec{x'}$ 上的投影与分量 $\vec{x'}$ 的模的比值。该几何意义可解释因变量随解释变量的变化大小和方向。一方面，该比值的绝对值越大，解释变量 X 引起的因变量 Y 的变动也就越大；另一方面，当投影 $\mathrm{pro}_{\vec{x'}}\vec{y'}>0$ 时，$\hat{\beta}_1>0$，解释变量 X 与因变量 Y 同向变动；当 $\mathrm{pro}_{\vec{x'}}\vec{y'}<0$ 时，$\hat{\beta}_1<0$，X 和 Y 反向变动；当 $\mathrm{pro}_{\vec{x'}}\vec{y'}=0$ 时，$\hat{\beta}_1=0$，X 能引起 Y 变动，X 对 Y 没有解释作用。

三、小结

从空间几何角度看，变量间的相关系数是空间两个向量的夹角的余弦，回归系数是空间向量的投影与另一个向量模的比值，直观的理解这两个概念的意义和作用，对以后学习和掌握复相关分析、偏相关分析及多元回归分析是很有帮助的。

参考文献

[1] 姚菊香，王盘兴等.相关系数显著性检验的几何意义 [J].南京气象学院学报，2007，30 (4)：566-570.

[2] 庞浩.计量经济学 [M].成都：西南财经大学出版社，2002.14-25.

基金项目：本文系常州工学院大学生创新实践训练项目 (编号：D090352)。
作者简介：姚俊 (1979-)，硕士，讲师，研究方向：统计理论及应用。