相关性

在许多情况下，观测值 x_i 并不是单个数值，而是向量 $x_i=(x_{i,1},\ldots,x_{i,d})$。我们通常对不同坐标之间的相关性(一个数据点的几个分量之间的关系)感兴趣。在本节中，我们将限制讨论只有两个坐标的向量，并将它们记作 (x_i,y_i)（而不是 $(x_{i,1},x_{i,2})$）。

样本相关系数

一个由对 (X_1,Y_1)，(X_n,Y_n) 组成的样本的样本相关系数为：

$$r_{X,Y}=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{(n-1)\sqrt{S_X^2}\sqrt{S_Y^2}}$$

注意这里是样本相关系数, 所以, 分母为 $\frac{1}{n-1}$.

数据对 (其实是一个数据点, 但是是二维的有两个分量) $(x_1,y_1),\ldots,(x_n,y_n)$ 的样本相关系数 $r_{x,y}$ 是线性相关性的强度的数值度量(无单位量)，取值在 -1 到 1 之间。其值的解释如下：
(i) 如果 $r_{x,y}=1$，那么散点图中的 n 个点完全落在直线 $y=\bar{y}+(s_y/s_x)(x-\bar{x})$ 上（完全正相关）。
(ii) 如果 $r_{x,y}=-1$，那么散点图中的 n 个点完全落在直线 $y=\bar{y}-(s_y/s_x)(x-\bar{x})$ 上（完全负相关）。(注意这里的系数是 $\frac{s_y}{s_x}$, 是一个值得注意的点)
(iii) 如果 $X_1,\ldots,X_n$ 和 $Y_1,\ldots,Y_n$ 是独立的样本，得到的 $r_{x,y}$ 将接近 0。

前两条陈述以及不等式 $|r_{x,y}|\leq 1$ 的结论来自柯西-施瓦兹不等式。

独立随机变量是不相关的，且样本相关系数会趋向于总体相关系数，但是不相关向量不一定独立 (考虑 $y=x^2$ 的情况).

当样本量 n 很大时，相关系数 ρ 趋向于：

$$\rho=\frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}X}\sqrt{\operatorname{var}Y}}=\frac{\mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)}{\sqrt{\mathrm{E}(X-\mathrm{E}X)^2}\sqrt{\mathrm{E}(Y-\mathrm{E}Y)^2}}$$

因为 $\operatorname{cov}(X,Y)=\mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)=\mathrm{E}(XY)-\mathrm{E}X\mathrm{E}Y$，因此对于独立的随机变量 X 和 Y，此系数 ρ 等于 0：独立随机变量是不相关的。我们将在讨论线性回归时进一步解释样本相关系数。

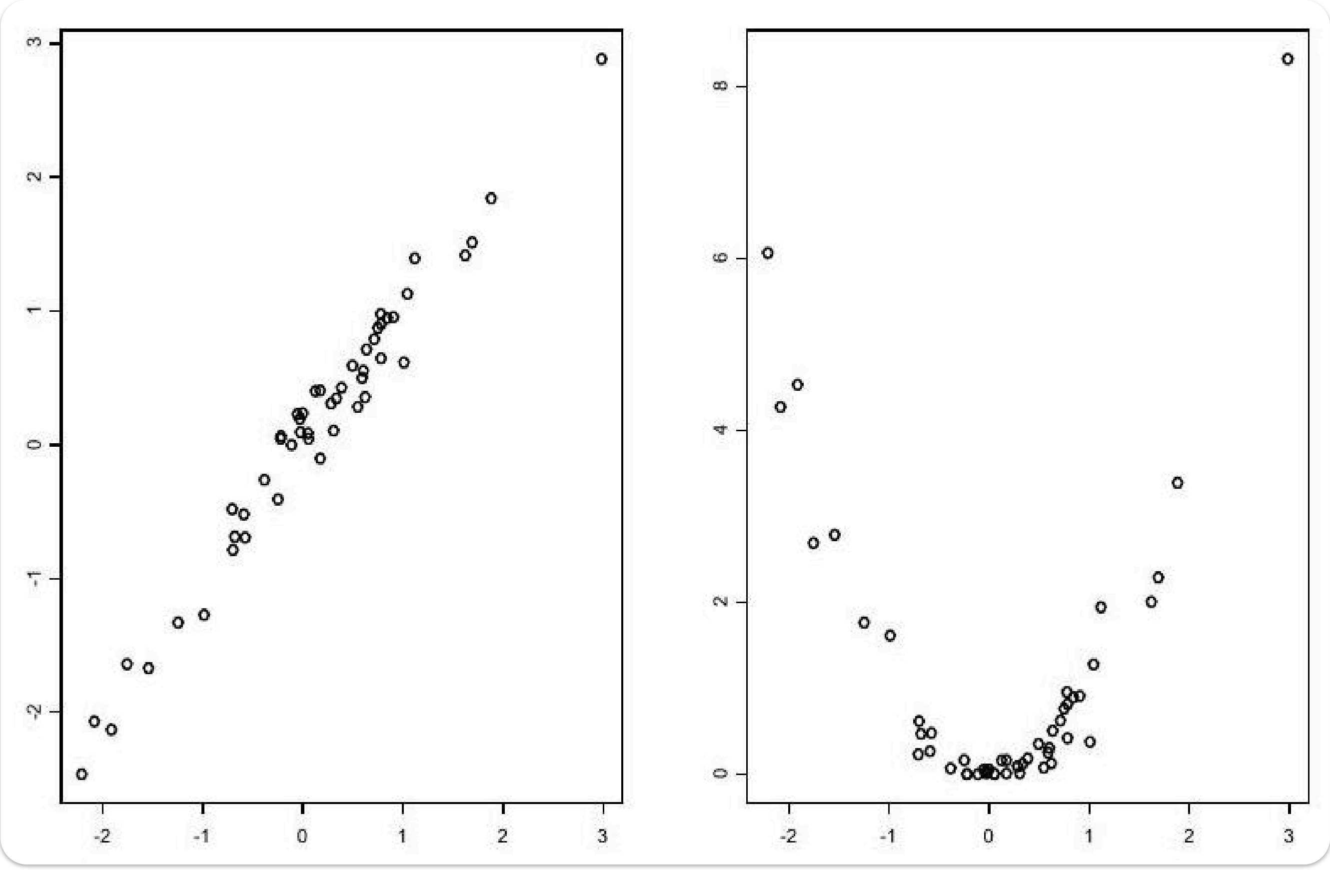

图: 两个样本的 50 个数据点的散点图，样本相关系数分别为 0.98 和 -0.05。右图给出了点 (x_i, y_i^2) 对应的散点图。

不相关不一定独立

不能将陈述 (iii) 反过来说，认为接近 0 的相关性意味着两个坐标是独立的。这在上图中得到了说明。左图显示了明显的线性相关性，其相关系数为 0.98。右图是点 (x_i, y_i) 对应的 (x_i, y_i^2) 的散点图，表现出明显的二次相关性。右图中的两个坐标之间的“相关强度”不亚于左图中的强度。然而，右图中的样本相关系数为 -0.05。显然，相关系数考虑的是线性意义上的相关,对存在的二次关系不敏感。