相关分析

1 相关系数及偏相关系数

考察相关性大小的时候既要考虑样本相关系数，也要考虑样本量大小。

相关系数度量了随机变量之间的线性关联程度，是研究变量关系最重要的工具之一（另一个是回归模型）。辛普森悖论表明两个不直接关联的随机变量如果都受第三个变量的调控，那么这两个变量就会呈现关联性，而如果一旦第三个变量给定，那么这种关联可能就会消失，我们以偏相关系数度量这种“条件线性相关性”。

1.1 相关系数

英国博物学家 Francis Galton 于十九世纪八十年代提出了相关系数的概念，英国统计学家 Karl Pearson 对此做了修正，提出了后来通用的 Pearson 相关系数。

定义 1.1 对任何存在二阶矩的随机变量 x, y，定义（总体）Pearson 相关系数

$$\rho = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)}\sqrt{\operatorname{var}(y)}}$$

相关系数度量了随机变量之间线性关联的程度。相关系数取值于 -1 和 1 之间，$|\rho| \le 1$。若 $\rho = 0$, 我们称 x, y 不相关, 若 $\rho > 0$ (< 0) 称为正 (负) 相关。若 $|\rho| = 1$, 则 x, y 是完美相关的，即两者存在严格的线性函数关系。另外，若 x, y 独立则它们一定不相关，但反之一般不成立，除非 x, y 联合服从二元正态分布或两者都是二值 (binary) 变量。

定义 1.2 假设 $(x_i, y_i), i = 1, \ldots, n$ iid $\sim (x, y)$, 样本相关系数定义为

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

样本相关系数是总体相关系数的渐近无偏估计（参见后面渐近正态结果），下面介绍基于样本相关系数的总体相关系数的统计推断。我们主要介绍正态假设下样本相关系数 r 的精确分布和小样本统计推断，简单介绍较弱条件的渐近分布及大样本统计推断。

1.1.1 精确检验

在正态假设下，考虑独立性零假设 $H_0 : \rho = 0$。检验统计量

$$t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$$

在零假设下的分布（称为零分布）为 t 分布，是一个精确分布，对应的检验是精确的，即检验的 I 型错误率能精确地控制在给定的水平。

引理 1.1 假设 $\mathbf{x} \sim N(0, I_n)$, 即 $x_1, ..., x_n$ iid $\sim N(0,1)$。

(a) 假设常数向量 $\mathbf{a} \in R^n$, $||\mathbf{a}|| = 1$, 则 $\mathbf{a}^\top\mathbf{x} \sim N(0,1)$, $||\mathbf{x}||^2 - (\mathbf{a}^\top\mathbf{x})^2 \sim \chi^2_{n-1}$, 两者独立，且

$$\sqrt{n-1}\frac{\mathbf{a}^\top\mathbf{x}}{\sqrt{||\mathbf{x}||^2 - (\mathbf{a}^\top\mathbf{x})^2}} \sim t_{n-1}.$$

(b) 假设常数向量 $\mathbf{a}, \mathbf{b} \in R^n$, $||\mathbf{a}|| = ||\mathbf{b}|| = 1$, $\mathbf{a}^\top\mathbf{b} = 0$, 则

$$\sqrt{n-2}\frac{\mathbf{a}^\top\mathbf{x}}{\sqrt{||\mathbf{x}||^2 - (\mathbf{a}^\top\mathbf{x})^2 - (\mathbf{b}^\top\mathbf{x})^2}} \sim t_{n-2},$$

其中分子分母独立。

证明：(a) 构造 $n \times n$ 正交矩阵 $A = \begin{pmatrix} \mathbf{a}^\top \\ * \end{pmatrix}$, 第一行为 $\mathbf{a}^\top$，令 $\mathbf{y} = A\mathbf{x} = \begin{pmatrix} \mathbf{a}^\top\mathbf{x} \\ * \end{pmatrix}$，其中 $y_1 = \mathbf{a}^\top\mathbf{x}$。因为 A 正交，$||\mathbf{y}|| = ||\mathbf{x}||$, $\mathbf{y} \sim N(0, I_n)$, 其中 $y_1, ..., y_n$ iid $\sim N(0,1)$, 则 $||\mathbf{x}||^2 - (\mathbf{a}^\top\mathbf{x})^2 = ||\mathbf{y}||^2 - y_1^2 = y_2^2 + ... + y_n^2 \sim \chi^2_{n-1}$, 且与 $y_1 \sim N(0,1)$ 独立，则由 t 分布定义

$$\sqrt{n-1}\frac{\mathbf{a}^\top\mathbf{x}}{\sqrt{||\mathbf{x}||^2 - (\mathbf{a}^\top\mathbf{x})^2}} = \frac{y_1}{\sqrt{(y_2^2 + ... + y_n^2)/(n-1)}} \sim t_{n-1}.$$

(b) 构造正交矩阵 A 使得其第一、二行分别为 $\mathbf{a}^\top, \mathbf{b}^\top$，其它证明与 (a) 类似。 ■

定理 1.2 假设 $y_i, i = 1, ..., n$ iid $\sim N(\mu, \sigma^2)$, 假设 x_i, y_i 独立 $(\rho_{xy} = 0)$，基于 (x_i, y_i) 的样本相关系数记为 r, 则

$$t \triangleq \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}.$$

证明：因为样本相关系数关于平移、刻度变换不变，不妨假设 $y_1, ..., y_n$ iid $\sim N(0,1)$, 即 $\mathbf{y} = (y_1, ..., y_n)^\top \sim N_n(0, I_n)$。给定 $\mathbf{x} =$

$(x_1,...,x_n)^\top$, 记 $\mathbf{a}=(x_1-\bar{x},...,x_n-\bar{x})^\top/\sqrt{s_{xx}}$, $\mathbf{b}=(1,...,1)^\top/\sqrt{n}$, 则 $||\mathbf{a}||=1, ||\mathbf{b}||=1$, 且 $\mathbf{a}^\top\mathbf{b}=0$, 则

$$\frac{r}{\sqrt{1-r^2}}=\frac{s_{xy}/\sqrt{s_{xx}}}{\sqrt{\sum y_i^2-n\bar{y}^2-(s_{xy}/\sqrt{s_{xx}})^2}}=\frac{\mathbf{a}^\top\mathbf{x}}{\sqrt{||\mathbf{x}||^2-(\mathbf{b}^\top\mathbf{x})^2-(\mathbf{a}^\top\mathbf{x})^2}}$$

由引理1.1 (b), 在给定 $\mathbf{x}$ 条件下 $t|\mathbf{x}\sim t_{n-2}$, 该分布与条件 $\mathbf{x}$ 无关, 所以 t 与 $\mathbf{x}$ 独立并且

$$\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}\sim t_{n-2}.$$

■

定理1.2的结果可用来检验正态分布的独立性假设。

独立性假设的 t 检验（精确检验）

假设 $(x_i,y_i), i=1,...,n$ iid 服从二元联合正态分布。若

$$\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}\geq t_{n-2}(\alpha/2),$$

则在 α 水平下拒绝 $H_0: x,y$ 独立。

例 1.1 (两样本 t 检验) 假设 $y_1,...,y_{n_1}$ iid $\sim N(\mu_1,\sigma^2)$, $y_{n_1+1},...,y_{n_1+n_2}$ iid $\sim N(\mu_2,\sigma^2)$, $n=n_1+n_2$, $H_0:\mu_1=\mu_2$ 的两样本 t 检验定义为

$$t_1=\sqrt{\frac{n_1n_2}{n_1+n_2}}\frac{(\bar{y}_1-\bar{y}_2)}{s}\overset{H_0}{\sim}t_{n-2},$$

其中 $\bar{y}_1,\bar{y}_2$ 分别是两组的样本均值，s^2 是 σ^2 的估计，

$$s^2=\frac{1}{n-2}(\sum_{i=1}^{n_1}(y_i-\bar{y}_1)^2+\sum_{i=n_1+1}^{n_1+n_2}(y_i-\bar{y}_2)^2).$$

下面我们从相关分析角度考虑该检验问题。引入两组样本的标号 $x_i=1, 1\leq n_1, x_i=2, n_1+1\leq i\leq n_1+n_2$，$H_0$ 的含义是，测量 y's 与分组 x's 无关。假设所有 $(x_i,y_i), i=1,...,n$ 的样本相关系数为 r，按照定理1.2, 我们可以应用 $t_2=\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$ 检验 $y's, x$'s 的相关性。容易验证 $t_2=t_1$。因此，两样本 t 检验是一个相关检验。

1.1.2 独立性的大样本检验

当总体不是正态分布的时候，独立性检验同样是统计学中的重要问题。考虑零假设

$$H_0 : x, y \text{ 独立},$$

我们将基于样本相关系数构建检验。当 H_0 成立时，可以验证定理 A1.1的渐近分布为标准正态。

命题 1.3 假设 $(x_i, y_i), i = 1, ..., n$ iid, x_i, y_i 独立，则

$$\sqrt{n}r \xrightarrow{d} N(0,1)).$$

证明：附录定理 A1.1的推论。 ■

因此在独立性假设 "$H_0 : x, y$ 独立" 成立时，近似地有 $\sqrt{n}r \sim N(0,1)$, 由此得到如下独立性假设的大样本检验。

独立性假设的大样本检验

若 $z = \sqrt{n}|r| > z_{\alpha/2}$ 或等价地 $z^2 = nr^2 > \chi_1^2(\alpha)$, 则在 α 水平下拒绝 H_0.

例 1.2 假设 $(x_i, y_i), i = 1, ..., n$ iid 样本相关系数 $r = 0.1$, 试对于 $n = 10, 20, 100, 500$, 分别检验 $H_0 : x, y$ 独立 (水平 $\alpha = 0.05$)。

解：二元正态假设下, 精确检验 $t = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}$，对于给定的 t, p 值 $p = P(|t_{n-2}| > |t|)$; 对于一般总体，大样本检验 $z = \sqrt{n}r$, $p = P(|N(0,1)| > |z|)$。结果如下

	精确检验		大样本检验	
n	t	p	z	p
10	0.284	0.783	0316	0.752
20	0.426	0.675	0.447	0.655
100	0.995	0.322	1.000	0.317
500	2.24375	0.025	2.236	0.025

当样本量较小时，两个检验的 p 值有差别，但都不能拒绝 H_0; 当 n 较大时，两个检验的 p 值几乎没差异, 且都拒绝 H_0。因此，对于独立性或相关性检验来说，样本量较大时（比如 $n \geq 20$），精确检验和大样本检验相差不大。

我们知道相关系数只能度量线性相关性而不能度量非线性相依关系，一个自然的问题是，以样本相关系数检验独立性是

不是有效？容易验证，当总体是二元正态或二元伯努利分布时，$\rho_{xy}=0$ 等价于 x,y 独立，[1] 因此在这两种最为常见的情况下，大样本检验是有效的方法。例如，当 x_i,y_i 都是二值变量的时候，可以验证 $z^2=nr^2$ 是 2×2 列联表的 Pearson 卡方统计量，因而是一个有效的检验方法（练习）。而对于其它总体，可能会不够有效。例如，若 (x,y) 服从平面单位圆周上的均匀分布，则 $\rho_{xy}=0$, 但 (x,y) 满足非线性函数关系 $x^2+y^2=1$，在这种情况下基于相关系数的大样本检验没有任何功效。

1: 需要注意的是, 两个不相关的正态随机变量如果不是联合服从二元正态，则它们未必独立。

当总体不是正态分布而且对大样本近似检验没有信心的时候，我们可以应用置换检验（参见附录1.4.3）

1.1.3 置信区间

为了构建总体相关系数 ρ 的置信区间，我们需要 $\rho\neq 0$ 时 r 的分布。附录定理 A1.1给出了一般总体的样本相关系数的渐近正态分布，其渐近方差依赖于总体的四阶矩，形式复杂。但在二元正态假设下，作为定理 A1.1的特殊情况，有如下渐近分布。

命题 1.4 假设 $(x_i,y_i),i=1,...,n$ iid 服从相关系数为 ρ 的二元正态分布, 则当 $n\to\infty$ 时

$$\sqrt{n}(r-\rho)\xrightarrow{d}N(0,(1-\rho^2)^2)).$$

基于该渐近分布，对于二元正态分布我们可以构建 ρ 的置信水平为 $1-\alpha$ 的大样本置信区间

$$\left\{\rho:\frac{\sqrt{n}|r-\rho|}{1-\rho^2}\leq z_{\alpha/2}\right\},$$

注意到命题1.4中 r 的近似正态分布 $N(\rho,(1-\rho^2)^2/n)$ 中，方差与均值 ρ 有关，这被称为是 “方差不稳定的”, 对样本相关系数做所谓的 Fisher’s z-变换, 可得到方差稳定的渐近正态分布（渐近正态分布的方差与均值无关）。

推论 1.5 假设总体是二元正态 (相关系数为 ρ)，当 $n\to\infty$ 时，

$$\sqrt{n}\left(\zeta(r)-\zeta(\rho)\right)\xrightarrow{d}N(0,1),$$

其中 Fisher’s z-变换 $\zeta(r)=\operatorname{atanh}(r)=\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$。

证明：由推论1.4 和 delta 方法立得。 ■

基于 Fisher z-变换构造置信区间如下

$$\left\{\rho:|\zeta(r)-\zeta(\rho)|\leq z_{\alpha/2}\right\}.$$

一般认为，该区间比基于 r 的渐近分布构造的区间具有更精确的覆盖率。

1.2 偏相关系数

1.2.1 干扰因素与辛普森悖论

变量 x, y 相关可能是因为它们都与共同的一个变量或向量 $\mathbf{z}$ 相关，这种情况下，$\mathbf{z}$ 对于研究 x, y 之间的关系是一个干扰因素或混杂因素 (confounder), 如图1.1所示。

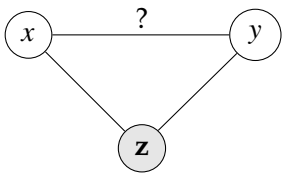

图 1.1: 变量 $\mathbf{z}$ 相关与 x, y 都相关，$\mathbf{z}$ 是研究 x, y 之间关系的干扰因素。

如果只研究感兴趣的变量 x, y 的关系而不考虑干扰因素 z 的影响，那么由此得到的结果与控制 z 时的结果可能不同，甚至相反，这种现象称为辛普森悖论 (Simpson's paradox)。下面是一个著名的辛普森悖论案例 [1]。

例 1.3 加州大学伯克利分校 1973 年研究生招生的女生录取率 35% 显著低于男生的录取率 44%, 因而校方担心被控告存在女性歧视。但分别考察各个系的录取率，反而大部分系的女生录取率高于男生。这种看似矛盾的现象就是典型的辛普森悖论。这里我们构造一个简单例子演示这种现象，假设两个招生系的数据如下（分母为申请人数，分子为录取人数）:

第一个系录取率: $4/10$ (男生) $< 1/2$ (女生);

第二个系录取率: $2/10$ (男生) $< 5/20$ (女生);

总录取率：$6/20$ (男生) $> 6/22$ (女生).

这里，录取情况 ($y = 1, 0$) 和性别 ($x = 1, 0$) 分布在两个系都不同：第一个系录取率较高，女生更倾向于报考第二个系。这里 x, y 都是二值 (binary) 变量，两个系的 (x, y) 相关系数分别为 $-0.076, -0.057$（负数）, 但合并两个系的数据之后相关系数为 0.030（正数）。因此，招生系在研究录取与性别的关系的时候是干扰因素。，我们应该设法消除掉招生系的影响。

1.2.2 去相关化

如何消除干扰因素的影响？理想的方法是采用随机化控制试验设计，但大部分研究不能实施试验而只能被动观察，在观察研究中我们可以应用去相关化或者回归分析控制变量的方法消除干扰因素的 (线性) 影响。偏相关系数的定义将依赖于这一技术，回归分析也基本如此。

假设 $\mathbf{x}, \mathbf{y}$ 分别是 $p \times 1, q \times 1$ 随机向量, 它们的方差-协方差矩阵为

$$\operatorname{var}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix}$$

若 $\Sigma_{\mathbf{xy}} \neq 0$, 我们希望消除 $\mathbf{y}$ 中含有的与 $\mathbf{x}$ 有关的线性成分，即求解线性变换矩阵 A 使得 $\mathbf{y} - A\mathbf{x}$ 与 $\mathbf{x}$ 不相关。由

$$0 = \operatorname{cov}(\mathbf{y} - A\mathbf{x}, \mathbf{x}) = \operatorname{cov}(\mathbf{x}_2, \mathbf{x}_1) - A\operatorname{cov}(\mathbf{x}_1, \mathbf{x}_1) = \Sigma_{21} - A\Sigma_{11},$$

得 $A = \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}$。

定义 1.3 $\mathbf{y}^{\perp} = \mathbf{y} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x}$ 称为 $\mathbf{y}$（关于 $\mathbf{x}$）的去相关化。

基于上述定义，下面我们试图定义两个随机向量之间的相关性大小的数字度量（而不是矩阵）。因为 $\mathbf{y}^{\perp}$ 与 $\mathbf{x}$ 不相关，所以也与 $\hat{\mathbf{y}} = \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x}$ 不相关，形式上我们记作 $\hat{\mathbf{y}} \perp \mathbf{y}^{\perp}$。因此有如下“正交” 分解

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{y}^{\perp},\ \hat{\mathbf{y}} \perp \mathbf{y}^{\perp},$$

两边同时求方差，得方差分解

$$\Sigma_{\mathbf{yy}} = \operatorname{var}(\mathbf{y}) = \operatorname{var}(\hat{\mathbf{y}}) + \operatorname{var}(\mathbf{y}^{\perp}) = \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}} + \Sigma_{\mathbf{yy}\bullet\mathbf{x}},$$

即 $\mathbf{y}$ 的方差分解为 $\mathbf{x}$ 所能解释的部分 $\operatorname{var}(\hat{\mathbf{y}})$ 和不能解释的部分 $\operatorname{var}(\mathbf{y}^{\perp}) = \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}} \triangleq \Sigma_{\mathbf{yy}\bullet\mathbf{x}}$。我们以前者在总方差 $\operatorname{var}(\mathbf{y}) = \Sigma_{\mathbf{yy}}$ 中的 “比例” 表示 $\mathbf{x}, \mathbf{y}$ 的相关程度:

$$\Phi = \Sigma_{\mathbf{yy}}^{-1/2}\left(\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\right)\Sigma_{\mathbf{yy}}^{-1/2},$$

当 $q > 1$ 时，Φ 是一个 $q \times q$ 矩阵（$0 \leq \Phi \leq I_q$），我们可以以该矩阵的某个数字特征包括最大特征根、行列式、迹 (trace) 等作为两个随机向量相关性的度量。古典多元分析中称 Φ 的最大特征根的平方根 $\lambda_{\max}^{1/2}(\Phi)$ 为第一典则相关系数。本课程主要关心的是 $q = 1$ 的情形，此时 Σ_{yy} 是实数，$\Sigma_{y\mathbf{x}}$ 是 $1 \times p$ 行向量，$\Sigma_{\mathbf{x}y}$ 是 $p \times 1$ 列向量,

$$\Phi = \frac{\Sigma_{y\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}}{\Sigma_{yy}}$$

是介于 0 和 1 之间的实数，称为决定系数，是回归分析中最重要的几个概念之一。

特别地，当 $p = q = 1$ 即 x, y 都是随机变量时，Φ 中诸 Σ 都是实数，则 Φ 等于相关系数的平方:

$$\Phi = \frac{\Sigma_{xy}^2}{\Sigma_{xx}\Sigma_{yy}} = \rho_{xy}^2.$$

1.2.3 偏相关系数

存在干扰因素 z 的情况下，为了得到 x, y 之间 “真正” 的关系，我们需要控制 (control) 变量 $\mathbf{z}$。“控制” 的意思是给定一个变量使之保持不变，我们可理解为在 x, y 中消除该变量的影响，在线性回归或相关分析中就是去相关化。

假设 $(x, y, \mathbf{z}^\top)^\top$ 的协方差矩阵为

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{x\mathbf{z}} \\ \Sigma_{yx} & \Sigma_{yy} & \Sigma_{y\mathbf{z}} \\ \Sigma_{\mathbf{z}x} & \Sigma_{\mathbf{z}y} & \Sigma_{\mathbf{zz}} \end{pmatrix}$$

为了排除 $\mathbf{z}$ 的影响，我们分别对 x, y 做关于 $\mathbf{z}$ 的去相关化，即在 x, y 中消除与 $\mathbf{z}$ 有关的线性成分，令

$$x^\perp = x - \Sigma_{x\mathbf{z}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z}, y^\perp = y - \Sigma_{y\mathbf{z}}\Sigma_{\mathbf{zz}}^{-1}\mathbf{z}$$

容易验证

$$\operatorname{cov}(x^\perp, y^\perp) = \Sigma_{xy\bullet\mathbf{z}}, \operatorname{var}(x^\perp) = \Sigma_{xx\bullet\mathbf{z}}, \operatorname{var}(y^\perp) = \Sigma_{yy\bullet\mathbf{z}},$$

其中

$$\Sigma_{ab\bullet c} = \Sigma_{ab} - \Sigma_{ac}\Sigma_{cc}^{-1}\Sigma_{cb}. \tag{1.1}$$

定义 x, y 的偏相关系数为 $x^\perp, y^\perp$ 的 Pearson 相关系数:

定义 1.4 $\rho_{xy\bullet\mathbf{z}} = \rho_{x^\perp y^\perp} = \dfrac{\operatorname{cov}(x^\perp, y^\perp)}{\sqrt{\operatorname{var}(x^\perp)}\sqrt{\operatorname{var}(y^\perp)}} = \dfrac{\Sigma_{xy\bullet z}}{\sqrt{\Sigma_{xx\bullet z}}\sqrt{\Sigma_{yy\bullet z}}}.$

因为 $x^\perp, y^\perp$ 都与 $\mathbf{z}$ 不相关，偏相关系数度量了在排除了干扰因素 $\mathbf{z}$ 之后，x, y 之间 “真正” 的相关关系。当 z 也是（一维）随机变量时，偏相关系数可由三个变量两两相关系数表示如下:

命题 1.6 当 x, y, z 都是随机变量时

$$\rho_{xy\bullet z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1 - \rho_{xz}^2)(1 - \rho_{yz}^2)}},$$

其中 ρ_{ab} 表示随机变量 a,b 的相关系数。

证明: 只需注意到 $\Sigma_{xy\bullet z} = \Sigma_{xy} - \Sigma_{xz}\Sigma_{zy}/\Sigma_{zz} = \sqrt{\Sigma_{xx}\Sigma_{y}}(\rho_{xy} - \rho_{xz}\rho_{yz})$, 以及 $\Sigma_{xx\bullet z} = \Sigma_{xx}(1-\rho_{xz}^2)$, $\Sigma_{yy\bullet z} = \Sigma_{yy}(1-\rho_{yz}^2)$。

■

例 1.4 假设三个随机变量 x,y,z 的相关系数矩阵如下（两两之间的相关系数都是 ρ）

$$R = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

其中 $\rho > -1/2$, 则偏相关系数 $\rho_{xy\bullet z} = \rho/(1+\rho)$, 由此知 $|\rho_{xy\bullet z}| < |\rho|$。这说明在一个对称的系统里（两两相关性相同），消除其它变量影响之后，任何两个随机变量的偏相关系数相比于相关系数都会更接近于 0（绝对值变小）。

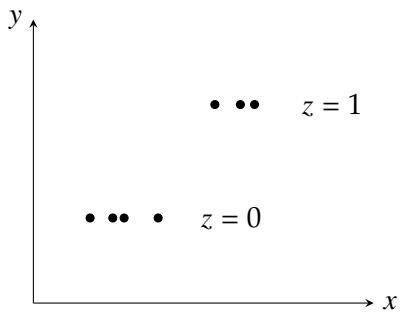

图 1.2: 辛普森悖论: 在 $z=0$ 组内和 $z=1$ 组内 x,y 都是不相关的，合并两组数据后 x,y 正相关。

例 1.5 假设 z 代表分组，辛普森悖论指的是，合并各组的 (x,y) 数据得到的相关性与分组计算的相关性不同甚至符号相反。图1.2给出了这样一个直观演示的例子，这里 $z=0,1$ 分别代表两组，在两组内部 x,y 不相关, 这意味着 $\rho_{xy\bullet z}=0$。另外，且 x,y 都随 z 增加而增加（正相关）。即 $\rho_{xz}>0, \rho_{yz}>0$, 所以不控制分组变量 z 的时候，$\rho_{xy} = \rho_{xz}\rho_{yz} > 0$，这和我们从图中观察到的现象一致。

1.3 多元正态分布

多元正态分布是统计学中最重要的一类分布，是线性统计的基础。虽然一般不明确假设多元正态模型，但在某种意义上，相关分析和线性回归模型都是以多元正态为基础建立并拓展的。

比如线性模型的均值线性性是正态模型下条件期望的基本性质，去相关化是多元正态条件化的一般化，正态假设下偏相关系数是条件相关系数，等等。

定义 1.5 若 p 维随机向量 $\mathbf{x}$ 的概率密度函数为

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

其中参数 $\boldsymbol{\mu} \in R^p, \Sigma$ 为 $p \times p$ 正定参数矩阵，则称 $\mathbf{x}$ 服从 p 元正态分布，记作 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$。

特别地，当 $\boldsymbol{\mu} = \mathbf{0}, \Sigma = I_p$（$p$ 阶单位阵）时，$N_p(\mathbf{0}, I_p)$ 称为 p 元标准正态分布，其概率密度函数为

$$g(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{x}\right),$$

这等价于 $\mathbf{x}$ 的各个分量服从一元标准正态分布，即

$$\mathbf{x} \sim N_p(\mathbf{0}, I_p) \Leftrightarrow x_1, ..., x_p \text{ iid } \sim N(0,1).$$

命题 1.7 若 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$，则 $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu}) \sim N_p(\mathbf{0}, I_p)$。反之，若 $\mathbf{y} \sim N_p(\mathbf{0}, I_p), B$ 是任一可逆常数矩阵使得 $BB^\top = \Sigma$, 则 $\mathbf{x} = B\mathbf{y} + \mu \sim N_p(\boldsymbol{\mu}, \Sigma)$。特别地, $\Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \Sigma)$。

证明： 假设 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$，其概率密度为 $f(\mathbf{x})$（定义1.5）。令 $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$，变换的 Jacobi 行列式

$$J = det\left(\partial \mathbf{x}/\partial \mathbf{y}\right) = det(\Sigma^{1/2}) = [det(\Sigma)]^{1/2},$$

利用随机向量函数的概率密度公式，$\mathbf{y}$ 的概率密度函数

$$g(\mathbf{y}) = f(\Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu}) \times |J| = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{y}^\top \mathbf{y}\right),$$

所以 $\mathbf{y} \sim N_p(\mathbf{0}, I_p)$。反之可类似证明。 ■

推论 1.8 假设 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$，则

(a) $E(\mathbf{x}) = \boldsymbol{\mu}$, $\text{var}(\mathbf{x}) = \Sigma$。
(b) 若 A 是 p 阶可逆常数矩阵，$\mathbf{b} \in R^p$ 是常数向量，则 $A\mathbf{x} + \mathbf{b} \sim N_p(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^\top)$.

证明： (a). 由命题1.7, $\mathbf{x}$ 可表示为 $\mathbf{x} = \Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu}$, 其中 $\mathbf{y}$ 的各个分量 $y_1, ..., y_p$ iid $\sim N(0,1)$, 因为 $E(y_i) = 0, \text{var}y_i) = 1$, 所以 $E(\mathbf{y}) = \mathbf{0}$, $\text{var}(\mathbf{y}) = I_p$, 所以 $E(\mathbf{x}) = E(\Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu}) = \Sigma^{1/2}E(\mathbf{y}) + \boldsymbol{\mu} = \boldsymbol{\mu}$, $\text{var}(\mathbf{x}) = \text{var}(\Sigma^{1/2}\mathbf{y} + \boldsymbol{\mu}) = \Sigma^{1/2}\text{var}(\mathbf{y})\Sigma^{1/2} = \Sigma$.

(b). 设 $\mathbf{x}=\Sigma^{1/2}\mathbf{y}+\boldsymbol{\mu}$, 其中 $\mathbf{y}\sim N_p(\mathbf{0},I_p)$ 则由命题1.7

$$A\mathbf{x}+\mathbf{b}=A(\Sigma^{1/2}\mathbf{y}+\boldsymbol{\mu})+\mathbf{b}=A\Sigma^{1/2}\mathbf{y}+A\boldsymbol{\mu}+\mathbf{b}\sim(A\boldsymbol{\mu}+\mathbf{b},A\Sigma A^{\top}).$$

■

命题 1.9 假设 $\mathbf{x}\sim N_p(\boldsymbol{\mu},\Sigma)$，划分

$$\mathbf{x}=\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix},\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix},\Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{pmatrix},$$

其中 $\mathbf{x}_1$ 和 $\boldsymbol{\mu}_1$ 是 $k\times 1$ 向量, Σ_{11} 是 $k\times k$ 矩阵。若 $\Sigma_{12}=0$, 则 $\mathbf{x}_i\sim N(\boldsymbol{\mu}_i,\Sigma_{ii}), i=1,2$, 且 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 独立。

证明： $\Sigma_{12}=0$ 时，$\mathbf{x}=(\mathbf{x}_1^{\top},\mathbf{x}_2^{\top})^{\top}$ 的密度可分解为 $\mathbf{x}_1$ 的函数和 $\mathbf{x}_2$ 的函数的乘积，它们分别是 $N(\boldsymbol{\mu}_1,\Sigma_{11})$ 和 $N(\boldsymbol{\mu}_2,\Sigma_{22})$ 的概率密度。 ■

多元正态分布在线性变换下具有封闭性，多元正态随机向量的分量、线性组合、条件分布都依然服从正态分布。

定理 1.10 假设 $\mathbf{x}\sim N_p(\boldsymbol{\mu},\Sigma)$，划分

$$\mathbf{x}=\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix},\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix},\Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{pmatrix},$$

其中 $\mathbf{x}_1$ 和 $\boldsymbol{\mu}_1$ 是 $k\times 1$ 向量, Σ_{11} 是 $k\times k$ 矩阵。

(a) $\mathbf{x}_1\sim N_k(\boldsymbol{\mu}_1,\Sigma_{11}),\mathbf{x}_2\sim N_{p-k}(\boldsymbol{\mu}_2,\Sigma_{22})$
(b) 给定 $\mathbf{x}_2$ 条件下，$\mathbf{x}_1$ 服从正态分布：

$$\mathbf{x}_1|\mathbf{x}_2\sim N_k(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2),\Sigma_{11\bullet 2}).$$

证明： 令去相关化变换

$$\begin{pmatrix}\mathbf{x}_1^{\perp}\\ \mathbf{x}_2\end{pmatrix}=\begin{pmatrix}I_k&-\Sigma_{12}\Sigma_{22}^{-1}\\ 0&I_{p-k}\end{pmatrix}\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix}$$

由推论1.8,

$$\begin{pmatrix}\mathbf{x}_1^{\perp}\\ \mathbf{x}_2\end{pmatrix}\sim N_p\left(\begin{pmatrix}\boldsymbol{\mu}_1-\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2\\ \boldsymbol{\mu}_2\end{pmatrix},\begin{pmatrix}\Sigma_{11\bullet 2}&0\\ 0&\Sigma_{22}\end{pmatrix}\right),$$

由命题1.9,

$$\mathbf{x}_1^{\perp}\sim N_k(\boldsymbol{\mu}_1-\Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2,\Sigma_{11\bullet 2}),\mathbf{x}_2\sim N_{p-k}(\boldsymbol{\mu}_2,\Sigma_{22}),$$

且两者独立。所以给定 $\mathbf{x}_2$ 时，$\mathbf{x}_1=\mathbf{x}_1^{\perp}+\Sigma_{12}\Sigma_{22}^{-1}\mathbf{x}_2\sim N_k(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2),\Sigma_{11\bullet 2})$. ■

备注 1.1 定理1.10说明了多元正态分布的诸多事实，列举如下

(1) 均值/回归函数（条件期望）线性性：

$$E(\mathbf{x}_1|\mathbf{x}_2) = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2).$$

(2) 方差齐性：条件方差是常数

$$\mathrm{var}(\mathbf{x}_1|\mathbf{x}_2) = \Sigma_{11\bullet 2}.$$

(3) 独立性：去相关化

$$\mathbf{x}_1^{\perp} = \mathbf{x}_1 - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{x}_2 \sim N_k(\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2, \Sigma_{11\bullet 2})$$

与 $\mathbf{x}_2$ 独立。

(4) 对于多元正态分布来说，去相关化等价于条件化：

$$\mathbf{x}_1^{\perp} = \mathbf{x}_1 - E(\mathbf{x}_1|\mathbf{x}_2) + constant.$$

偏相关系数实际上是条件相关系数（参见定理1.12）。

其中性质 (a)-(c) 是非正态情形下线性模型的基本假设。

利用定理1.10, 我们可将推论1.8 (b) 的结论推广到一般情形。

定理 1.11 假设 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, A 为 $k \times p$ 行满秩矩阵，$\mathbf{b}$ 为 $k \times 1$ 常数向量，则

$$A\mathbf{x} + \mathbf{b} \sim N_k(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^{\top}).$$

证明： 只需将 A 补全 $p-k$ 行成为 $p \times p$ 可逆阵 C: $C = \begin{pmatrix} A \\ B \end{pmatrix}$, 则由推论1.8,

$$C\mathbf{x} = \begin{pmatrix} A\mathbf{x} \\ B\mathbf{x} \end{pmatrix} \sim N_p(C\boldsymbol{\mu}, C\Sigma C^{\top}) = N_p\left(\begin{pmatrix} A\boldsymbol{\mu} \\ B\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} A\Sigma A^{\top} & A\Sigma B^{\top} \\ B\Sigma A^{\top} & B\Sigma B^{\top} \end{pmatrix}\right),$$

由定理1.10(b), 边际 $A\mathbf{x} \sim N_p(A\boldsymbol{\mu}, A\Sigma A^{\top})$。 ■

如前所述，偏相关系数 $\rho_{xy\bullet\mathbf{z}}$ 可理解为消除了 $\mathbf{z}$ 的 (线性) 影响之后，x, y 之间的相关系数，有时也解释为给定或控制 $\mathbf{z}$ 不变的条件下，x, y 之间的相关系数。这两种说法一般认为意思相同且不严格加以区分，但实际上，只有所有变量联合服从多元正态的情形下，偏相关系数才是真正意义上的条件相关系数。

定理 1.12 假设随机向量 $(x, y, \mathbf{z}^{\top})^{\top}$ 服从多元正态分布，则给定 $\mathbf{z}$ 的条件下，随机变量 x, y 的条件相关系数等于偏相关系

数。$\rho_{xy\bullet\mathbf{z}} = 0 \Leftrightarrow x, y$ 在给定 $\mathbf{z}$ 时条件独立。

证明： 假设 $(\mathbf{w}^\top, \mathbf{z}^\top)^\top = (x, y, \mathbf{z}^\top)^\top \sim N(\boldsymbol{\mu}, \Sigma)$, 其中 $\mathbf{w} = (x, y)^\top$ 是二维随机向量，$\mathbf{z}$ 是随机向量。划分

$$\Sigma = \begin{pmatrix} \Sigma_{\mathbf{ww}} & \Sigma_{\mathbf{wz}} \\ \Sigma_{\mathbf{zw}} & \Sigma_{\mathbf{zz}} \end{pmatrix}$$

则由多元正态的条件分布 (定理1.10) 知

$$\mathbf{w}|\mathbf{z} \sim N(*, \Sigma_{\mathbf{ww}\bullet\mathbf{z}}),$$

其中 2×2 条件协方差矩阵

$$\Sigma_{\mathbf{ww}\bullet\mathbf{z}} = \begin{pmatrix} \Sigma_{xx\bullet\mathbf{z}} & \Sigma_{xy\bullet\mathbf{z}} \\ \Sigma_{yx\bullet\mathbf{z}} & \Sigma_{yy\bullet\mathbf{z}} \end{pmatrix} \triangleq \begin{pmatrix} a & c \\ c & b \end{pmatrix},$$

因此给定 $\mathbf{z}$ 条件下，x, y 的条件相关系数

$$\rho_{xy|\mathbf{z}} = \frac{c}{\sqrt{ab}} = \frac{\Sigma_{xx\bullet\mathbf{z}}}{\sqrt{\Sigma_{xx\bullet\mathbf{z}}\Sigma_{yy\bullet\mathbf{z}}}},$$

最后一式正是偏相关系数的定义。 ■

高斯图模型以图 (graph) 表示多元正态各个分量之间的条件相关性，细节参见1.4.4。

1.4 补充

1.4.1 非线性相关性度量

常见的非线性关联度量包括 Spearman's rho、Kendall's tau 属性变量的 odds ratio、Pearson 卡方 χ^2 或其它类似的散度 (divergence) 度量。

1.4.2 相关系数的渐近分布

定理 A 1.1 假设 $(x_i, y_i), i = 1, ..., n$ iid 服从均值为 0、相关系数为 ρ 且四阶矩存在的二元分布，则当 $n \to \infty$ 时

$$\sqrt{n}(r - \rho) \xrightarrow{d} N(0, \gamma^2).$$

其中 $\gamma^2 = \frac{1}{4}\rho^2 c_1 - \rho c_2 + c_3$, $c_1 = \frac{\mathrm{var}(x^2)}{\sigma_x^4} + 2\frac{\mathrm{cov}(x^2,y^2)}{\sigma_x^2\sigma_y^2} + \frac{\mathrm{var}(y^2)}{\sigma_y^4}$, $c_2 = \frac{\mathrm{cov}(x^2,xy)}{\sigma_x^3\sigma_y} + \frac{\mathrm{cov}(y^2,xy)}{\sigma_x\sigma_y^3}$, $c_3 = \frac{\mathrm{cov}(xy,xy)}{\sigma_x^2\sigma_y^2}$.

证明细节参见 [2]。

1.4.3 置换检验

大样本检验对总体分布没有要求，使用范围广泛，但要求样本量较大；精确检验对样本量没有要求，但总体分布的假设较强。置换检验综合了两者的优势，对总体分布和样本量都没有要求。

假设样本 $(x_i, y_i), i = 1, ..., n$, 样本相关系数为 r。零假设 $H_0 : x, y$ 独立。H_0 成立时，我们没有理由相信 x_i 一定与 y_i 搭配在一起，即给定所有样本数值，对于任一 $\{1, 2, ..., n\}$ 的置换 σ，所有搭配 $\{(x_i, y_{\sigma(i)}), i = 1, ..., n : \sigma\}$ 都是等可能的（概率为 $1/n!$)。为了求样本相关系数的零分布，我们随机置换所有 y's, 对每一个置换 σ, 计算置换样本 $(x_i, y_{\sigma(i)}), i = 1, ..., n)$ 的样本相关系数 r_σ, 只要置换次数足够大，那么我们可以足够精确地求出零分布。假设 $_1, ..., _N$ 为 N 个随机置换，N 足够大，对每次置换 $\sigma_k, 1 \le k \le N$, 计算得到置换数据的样本相关系数 r_k, 原假设成立时，我们认为 $r_{\sigma_k}, k = 1, ..., N$ 是 r 的 N 次实现（随机样本），那么 r 的 p 值可以如下计算

$$p = \#\{|r_{\sigma_k}| > |r| : k = 1, ..., N\}/N.$$

需要说明的是，Pearson 相关系数的理论分布较容易得到，所以人们一般不对 Pearson 相关性检验应用置换方法。置换检验方法更适合应用于理论分布难以建立的其它关联性度量。

1.4.4 高斯图模型

定理1.12表明，对于多元正态分布，偏相关系数等于 0 等价于条件独立。高斯图模型以节点代表多元正态随机向量 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ 的各个分量 $x_1, ..., x_p$，条件不独立的节点之间连线（称为边），条件独立的节点之间不连线，由此得到的图表示多元正态分布随机向量分量之间的条件独立或相依关系，称为高斯图模型。

计算每两个分量之间的条件相关系数或偏相关系数是一个繁杂的过程，幸运的是所有偏相关系数可由协方差矩阵的逆（称为精度矩阵）Σ^{-1} 方便地求得。

引理 1.13 (分块矩阵的逆) 假设正定矩阵 Σ 划分为

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

则

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11\bullet 2}^{-1} & -\Sigma_{11\bullet 2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22\bullet 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22\bullet 1}^{-1} \end{pmatrix}$$

证明：记 $\Omega=\Sigma^{-1}$ 并按与 Σ 同样的方式分块

$$\Omega=\begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{pmatrix},$$

由

$$\Sigma\Omega=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{pmatrix}=\begin{pmatrix}I_1 & 0\\ 0 & I_2\end{pmatrix}$$

得

$$\Sigma_{11}\Omega_{11}+\Sigma_{12}\Omega_{21}]=I_1,\Sigma_{21}\Omega_{11}+\Sigma_{22}\Omega_{21}=0$$

由第二式得 $\Omega_{21}=-\Sigma_{22}^{-1}\Sigma_{21}\Omega_{11}$, 代入第一式得

$$\Omega_{11}=(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}=_{11\bullet 2}^{-1}.$$

其它类似可证。 ■

定理 A 1.2 假设 $p\times 1$ 随机向量 $\mathbf{x}=(x_1,...,x_p)^\top$ 的方差矩阵 $\mathrm{var}(\mathbf{x})=\Sigma=(\sigma_{ij}),\Omega=\Sigma^{-1}=(\omega_{ij})$, 则对任何 $1\le i,j\le p$, 控制其它分量时 x_i,x_j 的偏相关系数

$$\rho_{ij\bullet-(ij)}=\begin{cases}-\omega_{ij}/\sqrt{\omega_{ii},\omega_{jj}} & i\neq j,\\ 1, & i=j,\end{cases}$$

其中 ρ 下标中的 $-(ij)$ 代表 $\mathbf{x}$ 的除了第 i,j 分量之外所有其它的分量。

证明：显然，按照定义，当 $i=j$ 时，$\rho_{ii\bullet-(ii)}=1$。下面我们不妨只证 $i=1,j=2$ 的情形。记 $\mathbf{w}=(x_1,x_2)^\top,\mathbf{z}=(x_3,...,x_p)^\top$, 则 $\mathbf{x}=(\mathbf{w}^\top,\mathbf{z}^\top)^\top$。划分方差矩阵

$$\Sigma=\begin{pmatrix}\Sigma_{\mathbf{ww}} & \Sigma_{\mathbf{wz}}\\ \Sigma_{\mathbf{zw}} & \Sigma_{\mathbf{zz}}\end{pmatrix},$$

其中

$$\Sigma_{\mathbf{ww}}=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$$

是 Σ 的前两行两列组成的 2×2 子矩阵, 即 $\mathbf{w}=(x_1,x_2)^\top$ 的方差矩阵，而

$$\Sigma_{\mathbf{wz}}=cov(\mathbf{w},\mathbf{z})=\begin{pmatrix}\Sigma_{1\mathbf{z}}\\ \Sigma_{2\mathbf{z}}\end{pmatrix}$$

是 $2\times(p-2)$ 阶矩阵。所以

$$\Sigma_{\mathbf{ww}\bullet\mathbf{z}}=\Sigma_{\mathbf{ww}}-\Sigma_{\mathbf{wz}}\Sigma_{\mathbf{zz}}^{-1}\Sigma_{\mathbf{zw}}=\begin{pmatrix}\sigma_{11}&\sigma_{12}\\\sigma_{21}&\sigma_{22}\end{pmatrix}-\begin{pmatrix}\Sigma_{1\mathbf{z}}\\\Sigma_{2\mathbf{z}}\end{pmatrix}\Sigma_{\mathbf{zz}}^{-1}(\Sigma_{\mathbf{z}1},\Sigma_{\mathbf{z}2})$$
$$=\begin{pmatrix}\sigma_{11\bullet\mathbf{z}}&\sigma_{12\bullet\mathbf{z}}\\\sigma_{21\bullet\mathbf{z}}&\sigma_{22\bullet\mathbf{z}}\end{pmatrix}\triangleq\begin{pmatrix}a&b\\b&c\end{pmatrix}$$

由偏相关系数定义

$$\rho_{12\bullet\mathbf{z}}=b/\sqrt{ac}.$$

另一方面，由分块矩阵的逆的公式,

$$\Omega=\Sigma^{-1}=\begin{pmatrix}\Sigma_{\mathbf{ww}}&\Sigma_{\mathbf{wz}}\\\Sigma_{\mathbf{zw}}&\Sigma_{\mathbf{zz}}\end{pmatrix}^{-1}=\begin{pmatrix}\Sigma_{\mathbf{ww}\bullet\mathbf{z}}^{-1}&*\\*&*\end{pmatrix}$$

其左上角 2×2 矩阵

$$\Sigma_{\mathbf{ww}\bullet\mathbf{z}}^{-1}=\begin{pmatrix}a&b\\b&c\end{pmatrix}^{-1}=\frac{1}{ac-b^2}\begin{pmatrix}c&-b\\-b&a\end{pmatrix}=\begin{pmatrix}\omega_{11}&\omega_{12}\\\omega_{21}&\omega_{22}\end{pmatrix}$$

所以

$$\rho_{12\bullet\mathbf{z}}=b/\sqrt{ac}=-\omega_{12}/\sqrt{\omega_{11}\omega_{22}}.$$

■

偏相关系数矩阵

定理1.2给出了偏相关系数矩阵的如下简单算法。若 $p\times1$ 随机向量 $\mathbf{x}$ 的协方差矩阵为 Σ, 精度矩阵 $\Omega=\Sigma^{-1}$, 则所有分量之间的偏相关系数组成的矩阵

$$R_{pc}=\left(\rho_{ij\bullet-(ij)}\right)=2I_p-C^{-1/2}\Omega C^{-1/2},\ \ C=\mathrm{diag}(\Omega).$$

高斯图模型

若 $\mathbf{x}\sim N_p(\boldsymbol{\mu},\Sigma)$, $\Omega=\Sigma^{-1}=(\omega_{ij})$, 若 $\omega_{ij}\neq0$, 则 x_i,x_j 之间连线；若 $\omega_{ij}=0$, 即 x_i,x_j 条件独立，则 x_i,x_j 之间不连线。

参考文献

Here are the references in citation order.

[1] J. Bickel, E. A. Hammel, and J. W. O'connell. 'Sex Bias in Graduate Admissions: Data From Berkeley'. In: *Science* 187.4175 (1975), pp. 398–404 (cited on page 6).

[2] T.S. Ferguson. *A Course in Large Sample Theory*. Springer, 1996 (cited on page 13).