6.4 相关分析

在相关分析中，所涉及的变量都是随机的，且处于平等的地位，故用 $X_1,\cdots,X_p$ 来记，而不用 Y.

6.4.1 相关系数的估计和检验

设(X_1,X_2)服从二维正态分布 $N(a,b,\sigma_1^2,\sigma_2^2,\rho)$，其概率密度函数见第二章(2.7)式. 在第三章指出：a,σ_1^2 分别是 X_1 的均值方差，b,σ_2^2 分别是 X_2 的均值方差，而 ρ 是X_1,X_2之间的相关系数. 在 3.3 节中仔细论述了相关系数的意义，尤其是指出了：当总体分布为正态时，相关系数确实是变量之间的相关性的合理指标，

而在非正态情况则只是线性相关程度的度量.

相关系数 ρ 的公式是

$$\rho = \mathrm{Cov}(X_1, X_2)/(\mathrm{Var}(X_1)\mathrm{Var}(X_2))^{1/2} \tag{4.1}$$

这个公式启发了 ρ 的一个估计方法,即矩估计法.设(X_{11}, X_{21}), $\cdots$, (X_{1n}, X_{2n})为(X_1, X_2)的 n 个独立同分布的观察值,按矩法,分别以 $\left(\overline{X}_j = \sum_{i=1}^{n} X_{ji}/n, j = 1,2\right)$

$$\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2/(n-1), \sum_{i=1}^{n}(X_{2i} - \overline{X}_2)^2/(n-1) \text{ 和}$$

$$\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)/(n-1)$$

去估计 $\mathrm{Var}(X_1)$, $\mathrm{Var}(X_2)$和 $\mathrm{Cov}(X_1, X_2)$.由此,按(4.1),得出 ρ 的估计为

$$r = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\left[\sum_{i=1}^{n}(X_{1i} - \overline{X}_1)^2 \sum_{i=1}^{n}(X_{2i} - \overline{X}_2)^2\right]^{1/2}} \tag{4.2}$$

r 称为“样本相关系数”.

对 ρ 的检验,最有兴趣的是原假设

$$H_0: \rho = 0 \tag{4.3}$$

对立假设为 $\rho \neq 0$. H_0 表示 X_1, X_2 独立(在第三章已指出这在非正态情况下不成立).一个显然的检验方法是:计算 r,

$$\text{当} |r| \leqslant C \text{ 时接受} H_0, \text{不然就否定 } H_0 \tag{4.4}$$

常数 C 与样本大小 n 及检验水平 α 有关.要决定 C,必须求出在 $\rho = 0$时样本相关系数 r 的分布.这分布不很复杂,但我们这里无法介绍推导过程了,只指出:当 $\rho = 0$ 时有*

$$\sqrt{n-2}\,r/\sqrt{1-r^2} \sim t_{n-2} \tag{4.5}$$

由于$|r| \leqslant C$ 等价于 $\left|\sqrt{n-2}\,r/\sqrt{1-r^2}\right| \leqslant \sqrt{n-2}\,C/\sqrt{1-C^2}$,

* 证明见习题8.

由(4.5)不难定出：当给定检验水平 α 时，(4.4)中的 C 应取为方程 $\sqrt{n-2}C\Big/\sqrt{1-C^2}=t_{n-2}(\alpha/2)$ 之解，即

$$C=t_{n-2}(\alpha/2)\Big/\sqrt{n-2+t_{n-2}^2(\alpha/2)} \tag{4.6}$$

对 $n=20,30,\cdots,100$，由(4.6)算出的 C，为($\alpha=0.05$)

n	20	30	40	50	60	70	80	90	100
C	0.441	0.360	0.328	0.290	0.254	0.235	0.220	0.207	0.197

当样本大小 n 为 20 时，即使样本相关系数 r 达到 ± 0.4，尚不足以推断 ρ 异于 0. 随着 n 增加，这个界限逐步下降，但即使 n 达到 100，这个界限也还大约在 0.2. 这说明：要发现两变量之间较微弱的相关，样本大小 n 必须很大才行. 同时也说明了：对较小的 n，r 的精度很差，意义不大.

当 $\rho\neq 0$ 时样本相关系数 r 的分布问题，在本世纪初曾是 K. 皮尔逊和 R. A. 费歇尔等统计学大师着力研究的对象，最后被费歇尔在 1915 年解决了，其形式极为复杂，在此不能细述了.

6.4.2 偏相关

在统计学上，相关系数作为随机变量之间相关程度的刻画，用得很多，但在其解释上则应注意几点：一是统计相关不能等同于因果关系，这一点我们在第三章中已指出过了. 例如，分别以 X_1，X_2 记一个人的饮食和衣着消费，则 X_1，X_2 有较强的相关. 但很难说这二者有何因果关系：说好吃的人多半好穿，或者好穿的人多半好吃，未见得可信. 但既然如此，为什么在观察结果上又会显示出较强的相关呢？这就涉及到另一个需要注意之点：所考虑的变量(如此处的 X_1，X_2)并非孤立的，它们除彼此可能有的影响外，还受到一大批其他变量(不妨暂称为 $X_3,\cdots,X_p$ 等)的影响. 由于这个原因，相关系数有时被称为“完全相关系数”. 意思是说，在其中总结了由一切影响带来的相关性. 这个说法解释了上面提出的那个问

题:为何看来彼此并无密切因果关系的变量,在观察结果上会显示出较强的相关.这原因就在于被其他因素带动起来了.拿上例来说,如以 X_3 记人的收入,则一般说来,收入大的人各方面消费都倾向于高,它带动了 X_1(吃)和 X_2(穿)增长,以致使二者显示出较强的正相关.可以设想,如果能用某种方式把 X_3 的影响消去,则 X_1,X_2 可能显示很不一样的相关性质.例如它可以转为负相关.因为在一定收入的人中,在吃、穿中的一个方面消费大的人,一般会导致另一方面消费的减少.

一般,设有 p 个随机变量 $X_1,X_2,X_3,\cdots,X_p$.把 $X_3,X_4,\cdots,X_p$ 的影响从 X_1,X_2 中消去,剩余的部分分别记为 X_1' 和 X_2'.则 X_1',X_2' 的相关系数称为 X_1,X_2 对 $(X_3\cdots,X_p)$ 的偏相关系数,并记为 $\rho_{12\cdot(34\cdots p)}$.在以上论述中,"消去"一词的含义并未严格界定,但一般是在最小二乘法的意义下.例如,从 X_1 中消去 $X_3,\cdots,X_p$ 的影响,指的是找一个线性式

$$L_1(X_3,\cdots,X_p) = c_0 + c_3X_3 + \cdots + c_pX_p$$

使 $E[X_1 - L_1(X_3,\cdots,X_p)]^2$ 达到最小,剩余就是

$$X_1' = X_1 - L_1(X_3,\cdots,X_p)$$

同理找线性式 $L_2(X_3,\cdots,X_p)=d_0+d_3X_3+\cdots+d_pX_p$,使 $E[X_2 - L_2(X_3,\cdots,X_p)]^2$ 最小,剩余是

$$X_2' = X_2 - L_2(X_3,\cdots,X_p)$$

X_1,X_2 对 $(X_3,\cdots,X_p)$ 的偏相关系数 $\rho_{12\cdot(34\cdots p)}$ 就是 X_1',X_2' 的相关系数.要算出其表达式,就需要算出上文的线性式 L_1 和 L_2.下面我们对 $p=3$ 这个简单情况来计算一下.分别以 $a_1,a_2,a_3;\sigma_1^2,\sigma_2^2,\sigma_3^2$ 记 X_1,X_2 和 X_3 的均值和方差,以 $\rho_{12},\rho_{13},\rho_{23}$ 分别记 X_1,X_2 之间,X_1,X_3 之间,和 X_2,X_3 之间的相关系数.

关于找一个线性式 $L_1(X_3)$ 使 $E(X_1-L_1(X_3))^2$ 达到最小的问题,已在 3.3 节中讨论过了,按该章的(3.5)式,用此处的记号,有

$$L_1(X_3) = a_1 + \sigma_1\sigma_3^{-1}\rho_{13}(X_3 - a_3)$$

同理有

$$L_2(X_3) = a_2 + \sigma_2\sigma_3^{-1}\rho_{23}(X_3 - a_3)$$

故有

$$X'_1 = X_1 - a_1 - \sigma_1\sigma_3^{-1}\rho_{13}(X_3 - a_3)$$
$$X'_2 = X_2 - a_2 - \sigma_2\sigma_3^{-1}\rho_{23}(X_3 - a_3)$$

显然，$E(X'_1) = E(X'_2) = 0$，而按第三章(3.6)式，用此处的记号，有

$$\mathrm{Var}(X'_1) = \sigma_1^2(1 - \rho_{13}^2), \mathrm{Var}(X'_2) = \sigma_2^2(1 - \rho_{23}^2) \quad (4.7)$$

而

$$\begin{aligned}\mathrm{Cov}(X'_1, X'_2) &= E(X'_1, X'_2) = E[(X_1 - a_1)(X_2 - a_2)] \\ &\quad - \sigma_1\sigma_3^{-1}\rho_{13}E[(X_3 - a_3)(X_2 - a_2)] \\ &\quad - \sigma_2\sigma_3^{-1}\rho_{23}E[(X_1 - a_1)(X_3 - a_3)] \\ &\quad + \sigma_1\sigma_3^{-1}\rho_{13}\sigma_2\sigma_3^{-1}\rho_{23}E[(X_3 - a_3)]^2 \\ &= \sigma_1\sigma_2\rho_{12} - \sigma_1\sigma_3^{-1}\rho_{13}\sigma_2\sigma_3\rho_{23} \\ &\quad - \sigma_2\sigma_3^{-1}\rho_{23}\sigma_1\sigma_3\rho_{13} + \sigma_1\sigma_2\sigma_3^{-2}\rho_{13}\rho_{23}\sigma_3^2 \\ &= \sigma_1\sigma_2\rho_{12} - \sigma_1\sigma_2\rho_{13}\rho_{23} = \sigma_1\sigma_2(\rho_{12} - \rho_{13}\rho_{23}) \end{aligned} \quad (4.8)$$

由(4.7)，(4.8)，得

$$\begin{aligned}\rho_{12\cdot(3)} &= \mathrm{Corr}(X'_1, X'_2) \\ &= \mathrm{Cov}(X'_1, X'_2)/(\mathrm{Var}(X'_1)\mathrm{Var}(X'_2))^{1/2} \\ &= (\rho_{12} - \rho_{13}\rho_{23})/[(1 - \rho_{13}^2)(1 - \rho_{23}^2)]^{1/2} \end{aligned} \quad (4.9)$$

细察表达式(4.9)，有如下的构造：把 X_1, X_2, X_3 之间的相关系数，连同 X_i 与 X_i 之间的相关系数 $\rho_{ii} = 1$ 也在内，排列成一个三阶方阵(称为 X_1, X_2, X_3 的“相关阵”)

$$P = \begin{pmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{21} & \rho_{22} & \rho_{23} \\ \rho_{31} & \rho_{32} & \rho_{33} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

此处用了 $\rho_{ii} = 1$，$\rho_{ij} = \rho_{ji}$. 则其(1,1)元的子式，即划掉 P 的第一行第一列所剩下的行列式，等于 $P_{11} = 1 - \rho_{23}^2$. 同样，(2,2)元的子

式为 $P_{22}=1-\rho_{13}^2$,(1,2)元的子式为 $P_{12}=\rho_{12}-\rho_{13}\rho_{23}$.因此

$$\rho_{12\cdot(3)} = P_{12}/\sqrt{P_{11}P_{22}}$$

这个表达式,可以证明,能推广到 p 个自变量 X_1,X_2,X_3,X_p 的情况.仍以 ρ_{ij} 记 X_i,X_j 之间的相关系数($\rho_{ii}=1,\rho_{ij}=\rho_{ji}$),$P$ 记其相关阵:

$$P = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{pmatrix} \tag{4.10}$$

而以 P_{uv} 记 P 的(u,v)元的子式,即从 P 中划去第 u 行第 v 列所成的行列式,则

$$\rho_{12\cdot(34\cdots p)} = P_{12}/\sqrt{P_{11}P_{22}} \tag{4.11}$$

从表达式(4.9)看出一个现象.设 $\rho_{12}>0$,但不太接近于 1.即 X_1,X_2 为正相关,但相关程度不是非常密切.又 ρ_{13},ρ_{23}都很接近 1,则(4.9)式之分子将小于 0,即 $\rho_{12\cdot(3)}<0$.就是说,尽管 X_1,X_2 的通常相关系数为正,其偏相关系数可以为负.这拿前面举的那个 X_1 = 吃的支出,X_2 = 穿的支出,X_3 = 收入的例子可作一个印证.X_1,X_2 的(完全)相关 ρ_{12} 大于 0,但 ρ_{13},ρ_{23} 看来都为正且很大,故 $\rho_{12\cdot(3)}$当小于 0:从吃穿支出中消去收入的影响,等于在固定收入的情况下考虑二者的关系,其相关为负就不难理解了.当然,反过来也可能:即 $\rho_{12}<0$ 但 $\rho_{12\cdot(3)}>0$.

因此,在涉及多个变量相互影响的问题中,不仅考虑完全相关系数,而且考虑种种有意义的偏相关系数(在全部 p 个自变量中,可任选出 $k\geqslant 3$ 个:$X_{i_1},\cdots,X_{i_k}$,而考虑 X_{i_1},X_{i_2} 对$(X_{i_3},\cdots,X_{i_k})$的偏相关系数.其计算仍按(4.11),只是在 P 中要把不是 $i_1,\cdots,i_k$ 那些行列都划去),这样对整个相关的图景就可获得深入一层的了解.

读者也不要误以为偏相关系数高于完全相关系数,这二者各说明“相关”这个概念的一个侧面,其含义不同.在什么情况下哪一

种相关更为贴切,要看问题的性质.

如果对$(X_1,\cdots,X_p)$进行了 n 次观察,得样本

$$(X_{1i},\cdots,X_{pi}),i=1,\cdots,n$$

则可以用前面的方法(见(4.2)式)估计 X_u 与X_v 的相关系数,即计算样本相关系数 r_{uv}:

$$r_{uv}=\sum_{i=1}^{n}(X_{ui}-\overline{X}_u)(X_{vi}-\overline{X}_v)\Big/\Big[\sum_{i=1}^{n}(X_{ui}-\overline{X}_u)^2\cdot\sum_{i=1}^{n}(X_{vi}-\overline{X}_v)^2\Big]^{1/2}$$

其中 $\overline{X}_k=(X_{k1}+\cdots+X_{kn})/n,k=1,\cdots,p$. 有 $r_{uu}=1,r_{uv}=r_{vu}$. 以 r_{uv}代替P 中的ρ_{uv}得样本相关阵

$$R=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p}\\ r_{21} & r_{22} & \cdots & r_{2p}\\ \vdots & \vdots & & \vdots\\ r_{p1} & r_{p2} & \cdots & r_{pp}\end{pmatrix}\qquad(4.12)$$

然后用

$$r_{12\cdot(34\cdots p)}=R_{12}/\sqrt{R_{11}R_{22}}\qquad(4.13)$$

去估计 $r_{12\cdot(34\cdots p)}$. 它称为样本偏相关系数.

如果要检验有关 $\rho_{12\cdot(34\cdots p)}$的假设,则必须假定变量服从正态分布. 在这种假定下,可以证明:原假设

$$H_0:\rho_{12\cdot(34\cdots p)}=0\qquad(4.14)$$

的一个水平 α 的检验为

$$\begin{cases}|r_{12\cdot(34\cdots p)}|\leqslant t_{n-p}(\alpha/2)/[n-p+t_{n-p}^2(\alpha/2)]^{1/2}, \text{接受 } H_0\\ |r_{12\cdot(34\cdots p)}|>t_{n-p}(\alpha/2)/[n-p+t_{n-p}^2(\alpha/2)]^{1/2}, \text{否定 } H_0\end{cases}\qquad(4.15)$$

此检验与前述相关系数为 0 的检验之差别仅在于,把(4.6)式中的 $n-2$ 换为 $n-p$.

例 4.1 随机抽取 1000 人调查其(每年)吃的支出(X_1),衣着支出(X_2)和收入(X_3),算出的样本相关系数分别为 $r_{12}=0.57$,

$r_{13}=0.82, r_{23}=0.80$. 对 $n=1000, \alpha=0.05, t_{n-2}(\alpha/2)$ 和 $t_{n-3}(\alpha/2)$ 都可取为 1.96. 于是易算得 $|r_{12}|>t_{n-2}(\alpha/2)/\sqrt{n-2+t_{n-2}^2(\alpha/2)}$,因而 X_1, X_2 的(完全)相关在 $\alpha=0.05$ 的水平上为显著的且为正相关. 按公式(4.9),算出

$$r_{12\cdot(3)}=(r_{12}-r_{13}r_{23})/\sqrt{(1-r_{13}^2)(1-r_{23}^2)}=-0.73$$

它在水平 $\alpha=0.05$ 时为高度的负相关.

6.4.3 复相关

设有若干个随机变量 $X_1,\cdots,X_p$. 可能有这种情况:X_1 对每个 $X_j(j\geqslant 2)$的相关性不一定很显著,但全体 $X_2,\cdots,X_p$ 合起来,则与 X_1 有较显著的相关. 例如,设 X_1 为某种水田农作物的产量,$X_2,\cdots,X_p$ 为该作物生长期那几个月的各月降雨量(例如3、4、5、6月),亩产与指定一月的降雨量肯定有关,但不一定十分大,而全体这几个月的降雨情况,则肯定与亩产有更大的相关. 这种以 X_1 为一方,$X_2,\cdots,X_p$ 全体为一方之间的相关,称为 X_1 与$(X_2,\cdots,X_p)$的"复相关".

这种复相关的定义,与偏相关有其相似之处,就是也要找 $X_2,\cdots,X_p$ 的一个线性式 $L(X_2,\cdots,X_p)=c_0+c_2X_2+\cdots+c_pX_p$,使 $E[X_1-L(X_2,\cdots,X_p)]^2$达到最小. 然后,$X_1$ 与 $L(X_2,\cdots,X_p)$的通常相关系数,就定义为 X_1 和$(X_2,\cdots,X_p)$之间的"复相关系数",并记为 $\rho_{1(23\cdots p)}$.

求 $L(X_2,\cdots,X_p)$的方法,与3.3节所用方法相似(那里解决了 $p=2$ 的情况). 仔细推导过程不在此写出了,我们只给出最后的结果为

$$\rho_{1(23\cdots p)}=\sqrt{1-|P|/P_{11}} \tag{4.16}$$

这里$|P|$为(4.10)所定义的方阵 P 的行列式,P_{11}如前,是方阵 P 的(1,1)元的子式.

如果对$(X_1,X_2,\cdots,X_p)$进行了 n 次观察,得样本$(X_{1i},X_{2i},X_{pi}),i=1,\cdots,n$,则由之计算出样本相关阵 R(见(4.12)式),以

R 取代(4.16)中之 P,得样本复相关系数

$$r_{1(23\cdots p)} = \sqrt{1 - |R|/R_{11}} \tag{4.17}$$

它可作为 $\rho_{1(23\cdots p)}$ 的估计.

关于复相关系数的检验,实用上有兴趣的是

$$H_0:\rho_{1(23\cdots p)} = 0 \tag{4.18}$$

直观上看,一个显然的检验方法是

$$\text{当 } r_{1(23\cdots p)} \leqslant C \text{ 时接受 } H_0\text{,不然就否定 } H_0 \tag{4.19}$$

要依据检验水平 α 去决定(4.19)中的常数 C,就必须求出当 H_0 成立时,$r_{1(23\cdots p)}$ 的分布.可以证明:当正态假定成立且 H_0 为真时,$r^2_{1(23\cdots p)}$ 的分布为所谓"β 分布",其密度函数 $f(x)$ 为

$$f(x) = \begin{cases} \dfrac{1}{\beta\left(\dfrac{p-1}{2}, \dfrac{n-p}{2}\right)} x^{\frac{p-3}{2}}(1-x)^{\frac{n-p-2}{2}}, & 0 < x < 1 \\ 0, & \text{其他 } x \end{cases} \tag{4.20}$$

其中 $\beta\left(\dfrac{p-1}{2}, \dfrac{n-p}{2}\right)$ 曾在第二章的附录中定义过.用这个分布去决定(4.19)中的 C,可以通过 F 分布表.因为,在(4.20)的基础上可以证明:在 H_0 成立时有

$$\frac{n-p}{p-1}\,\frac{r^2_{1(23\cdots p)}}{1 - r^2_{1(23\cdots p)}} \sim F_{(p-1)/2,(n-p)/2} \tag{4.21}$$

$F_{a,b}$ 为自由度 a,b 的 F 分布(见第 2 章例 4.11).由(4.21),定出在给定水平 α 时,(4.19)式中的 C 为

$$C = \left[\left(\frac{p-1}{n-p}F_{(p-1)/2,(n-p)/2}(\alpha)\right)\bigg/\left(1 + \frac{p-1}{n-p}F_{(p-1)/2,(n-p)/2}(\alpha)\right)\right]^{1/2} \tag{4.22}$$

在以上的叙述中,$X_1, \cdots, X_p$ 也可以只是考察的全部变量中的一部分.例如,X_1 代表亩产量,$X_2, \cdots, X_p$ 代表所考察的全部气象因子,如有关各月的降水量,月平均气温等,而 $X_{p+1}, \cdots, X_q$ 等

则代表与田间管理有关的因子,另外还可以有别的因子.我们可以考虑 X_1 与$(X_2,\cdots,X_p)$的复相关,以看看亩产量与气象因子相关的程度如何,可以考虑 X_1 与$(X_{p+1},\cdots,X_q)$的复相关,以看看亩产量与管理因子相关的程度如何,等等.上面所说的估计和检验方法当然仍然适用.