第四讲. 相关系数（续）

2014.2.27

非正态情形下的大样本相关性检验和置信区间

非正态假设下，或即使假设正态分布但ρ非零情形下，样本相关系数分布如何？

样本相关系数的大样本分布

性质1. 样本：$(x_1,y_1),\ldots,(x_n,y_n)$ iid，设ρ为总体相关系数，r为样本相关系数，则(可参看有关数理统计教程)

$$\sqrt{n}\,(r-\rho)\xrightarrow{d}N\left(0,(1-\rho^2)^2\right),\quad \text{当}\, n\to\infty$$

(1). 假设检验

相关性的大样本检验（不假设联合正态分布）：

H_0成立时，当$n\to\infty$时，

$$z=\sqrt{n}\times r\xrightarrow{\mathrm{d}}N(0,1)$$

则p值$\approx P(|N(0,1)|>|z|)=2(1-\Phi(|z|))$，$\Phi$是$\mathrm{N}(0,1)$累积分布函数。

注：通常使用 $z=\sqrt{n-2}\times r\sim N(0,1)$

基于性质1，可以如下构造ρ的95%置信区间：$\left\{\rho:\left|\dfrac{\sqrt{n}\,(r-\rho)}{1-\rho^2}\right|\le 1.96\right\}$，

但通常使用相关系数的Fisher变换构造置信区间。

(2). 置信区间

方差稳定化

$$\operatorname{atanh}(r)=\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$称为相关系数r的Fisher变换。

基于*Fisher*变换的95%置信区间：

$$\left\{\rho:\left|\sqrt{n}\left\{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)-\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)\right\}\right|\le 1.96\right\}$$

该区间构造依赖于下述事实：

性质2.设$(x_1,y_1)\ldots,(x_n,y_n)iid$，设$\rho=\mathrm{cor}(x,y)$,则当$n\to\infty$时

$$\sqrt{n}\left\{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)-\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)\right\}\xrightarrow{d}N(0,1)$$

证明：根据性质1中r的极限分布和delta 方法

注：基于Fisher变换比直接基于r的渐近分布所构造的置 信区间具有更好的精确度 。

非参数相关性检验/置信区间

$H_0:\rho=0\leftrightarrow H_1:\rho\neq 0$

(1) 置换检验(permutation test): 不假设正态分布、且/或样本量较小的情形下，可使用如下置换方法

For k $=1,2,\ldots,N$ （N : 置换次数）

(1) 置换$x_1,x_2,\ldots,x_n$,得序列$x_{i_1},\ldots,x_{i_n}$

(2) 计算置换后的样本对$(x_{i_1},y_1),\ldots,(x_{i_n},y_n)$的相关系数$r_k$

$$p值=\frac{\#\{k:|r_k|\geq|r|\}}{N}$$

(2). 自助法（bootstrap)构造置信区间

无联合正态分布假设下，$\rho=\mathrm{corr}(x,y)$的95%置信区间可由自助法（*bootstrap*）构建：

从$(x_1,y_1),\ldots,(x_n,y_n)$中有放回地抽取 n对，计算其样本相关系数 r^*。

重复N次. 得到N个 r^* 。

计算N个 r^*的 2.5%和97.5%分位点，分别作为置信区间的下 界和上界。

3. Spearman’s rho, Kendall’s tau

- Spearman’s rho

样本:$(x_1,y_1),\ldots,(x_n,y_n)$. 记$R_i=x_i$在所有$x$中的秩（排名）;$S_i=y_i$在所有$y$中的秩

Spearman's rho

$$\hat{\rho}=r_{RS}=\frac{\sum(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum(R_i-\bar{R})^2\sum(S_i-\bar{S})^2}}$$

例：x=(23, 67, 0.2, 99)
Rank(x) =(2, 3, 1, 4)

- Kendall’s tau

$$\hat{\tau}=\frac{\mathrm{C-D}}{\binom{n}{2}},\ \text{其中}\mathrm{C}=\#\{(i,j):(x_i-x_j)(y_i-y_j)>0,i<j\},\mathrm{D}=\binom{n}{2}-C$$

统计推断：置换或自助法

4. 随机向量（中文课本第二章）

1. 均值：设$\mathrm{x}=\begin{pmatrix}\mathrm{x}_1\\ \vdots\\ \mathrm{x}_k\end{pmatrix}$是随机向量，其均值定义为$\mathrm{E(x)}=\begin{pmatrix}\mathrm{E}(\mathrm{x}_1)\\ \vdots\\ \mathrm{E}(\mathrm{x}_k)\end{pmatrix}$

2. 协方差矩阵

设$\mathrm{x}=\begin{pmatrix}\mathrm{x}_1\\ \vdots\\ \mathrm{x}_k\end{pmatrix}$，$\mathrm{y}=\begin{pmatrix}\mathrm{y}_1\\ \vdots\\ \mathrm{y}_m\end{pmatrix}$是两个随机向量,记$\mu_{\mathrm{x}}=\mathrm{E(x)}$，$\mu_{\mathrm{y}}=\mathrm{E(y)}$.

x，y的协方差矩阵($k\times m$)定义为:

$$\mathrm{cov(x,y)}=\mathrm{E}(\mathrm{x}-\mu_{\mathrm{x}})(\mathrm{y}-\mu_{\mathrm{y}})'=\left(\mathrm{cov}(x_i,y_j)\right)_{i=1,\dots,k;\ j=1,\dots,m}$$

3. 方差矩阵(或方差 - 协方差矩阵）:

$$\mathrm{var(x)}\text{或}\mathrm{cov(x)}=\mathrm{E}(\mathrm{x}-\mu)(\mathrm{x}-\mu)'=\mathrm{cov(x,x)}$$

性质：

1. 设x, y分别为$n\times 1$的随机向量，z, w是$n\times 1$随机向量，则

(1) $\mathrm{E(x+y)}=\mathrm{E(x)}+\mathrm{E(y)}$

(2) $\mathrm{cov(x+y,z+w)}=\mathrm{cov(x,z)}+\mathrm{cov(x,w)}+\mathrm{cov(y,z)}+\mathrm{cov(y,w)}$

2. 设x为$n\times 1$随机向量，A为$m\times n$常数矩阵，则

(1) $\mathrm{E}(A\mathrm{x})=A\mathrm{E(x)}$

(2) $\mathrm{cov}(A\mathrm{x})=A\,\mathrm{cov(x)}A'$

证明：（2）$\mathrm{cov}(A\mathrm{x})=\mathrm{E}\left(A\mathrm{x}-A\mathrm{E(x)}\right)\left(A\mathrm{x}-A\mathrm{E(x)}\right)'$

$=\mathrm{E}A\left(\mathrm{x}-\mathrm{E(x)}\right)\left(\mathrm{x}-\mathrm{E(x)}\right)'A'=A\left[\mathrm{E}\left(\mathrm{x}-\mathrm{E(x)}\right)\left(\mathrm{x}-\mathrm{E(x)}\right)'\right]A'$

$=A\,\mathrm{cov(x)}A'$

3. 设x, y分别为$n\times 1$和$m\times 1$随机向量，A, B分别为$p\times n, q\times n$常数矩阵，则 $\mathrm{cov}(A\mathrm{x}, B\mathrm{y})=A\mathrm{cov(x,y)}B'$

4. 设x为$n\times 1$随机向量，则方差 - 协方差矩阵$\mathrm{cov(x)}\geq 0$ (半正定).

证明：记$\Sigma=\mathrm{cov(x)}$, 对于任何常数向量$\mathrm{c}\in\mathrm{R}^n$,

由性质2(2)，$\mathrm{c}'\Sigma\mathrm{c}=\mathrm{c}'\mathrm{cov(x)c}=\mathrm{cov(c'x)}$

注意cov(c'x)是随机变量c'x的方差，非负，所以$\mathrm{c}'\Sigma\mathrm{c}\geq 0$.

5. 设x为$n\times 1$随机向量, $\mathrm{E(x)}=\mu$，$\mathrm{cov(x)}=\Sigma$, A为$n\times n$常数矩阵，则 $\mathrm{E}\left(\mathrm{x'Ax}\right)=\mu'\mathrm{A}\mu+\mathrm{tr}(\mathrm{A}\Sigma)$

证明：$\mathrm{x'Ax}=\mathrm{tr(x'Ax)}=\mathrm{tr(Axx')}\Rightarrow\mathrm{E(x'Ax)}=\mathrm{tr(AE(xx'))}$

而$\Sigma=\mathrm{E(xx')}-\mu\mu'$

$\mathrm{tr(AE(xx'))}=\mathrm{tr}(\mathrm{A}[\Sigma+\mu\mu'])=\mathrm{tr}(\mathrm{A}\Sigma)+\mathrm{tr}(\mathrm{A}\mu r(\,)=\mathrm{tr}(\mathrm{A}\Sigma)+\mu'\mathrm{A}\mu$

“不相关化”（“正交”化/对角化）

任意随机向量x_1，x_2,记$\Sigma=\mathrm{cov}\begin{pmatrix}\mathrm{x}_1\\ \mathrm{x}_2\end{pmatrix}=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$

令$\mathrm{x}_1^{\perp}=\mathrm{x}_1-\Sigma_{12}\Sigma_{22}^{-1}\mathrm{x}_2$，则$\mathrm{x}_1^{\perp}$与$\mathrm{x}_2$不相关,且

$\mathrm{var}(\mathrm{x}_1^{\perp})=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\overset{\text{记为}}{=}\Sigma_{11\bullet 2}$

注：

$\mathrm{x}_1^{\perp}=\mathrm{x}_1$中消除掉与$\mathrm{x}_2$线性相关的部分$\Sigma_{12}\Sigma_{22}^{-1}\mathrm{x}_2$,

后者可以认为是x_1在x_2上的正交投影

总结一下：

$$\begin{pmatrix} x_1^{\perp} \\ x_2 \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \Rightarrow \operatorname{cov}\begin{pmatrix} x_1^{\perp} \\ x_2 \end{pmatrix} = \begin{pmatrix} \Sigma_{11\bullet 2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

$$\Leftrightarrow$$

方差矩阵对角化：

$$\begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} = \begin{pmatrix} \Sigma_{11\bullet 2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

引理1（分块矩阵的逆）： $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} > 0$ （正定）

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

其中 $\Sigma_{11\bullet 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$， $\Sigma_{22\bullet 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$，

证明：由对角化公式（Σ为n阶，Σ_{11}为k阶）：

$$\begin{pmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n-k} \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} I_k & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_{n-k} \end{pmatrix} = \begin{pmatrix} \Sigma_{11\bullet 2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

得 $\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n-k} \end{pmatrix}^{-1}\begin{pmatrix} \Sigma_{11\bullet 2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}\begin{pmatrix} I_k & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_{n-k} \end{pmatrix}^{-1}$

两边求逆 $\Rightarrow$

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I_k & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_{n-k} \end{pmatrix}\begin{pmatrix} \Sigma_{11\bullet 2}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix}\begin{pmatrix} I_k & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{n-k} \end{pmatrix}$$

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

注意：由对称性知右下角的 $\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\bullet 2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}$
实际上等于 $\Sigma_{22\bullet 1}^{-1}$，同样左下角 $-\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\bullet 2}^{-1} = -\Sigma_{22\bullet 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}$