回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm20208

第三讲 相关系数与偏相关系数

2020.2.26

内容

- 相关系数
- 相关系性检验
- 随机变量（向量）的不相关化
- 偏相关系数

Pearson相关系数

随机变量之间的关联性通常以Pearson相关系数度量，它实际上度量的是线性关联程度。相关系数的概念和初始定义由Galton提出，但深入的研究和推广属于K. Pearson。

样本$(x_1, y_1),\ldots,(x_n, y_n)$ 的 Pearson 样本相关系数：

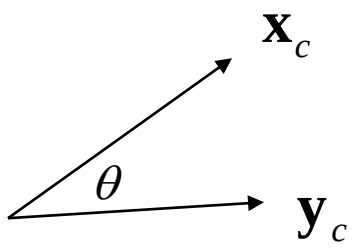
$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}.$$

随机变量x，y的(总体)相关系数：$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}}$

记$\mathbf{x}_c, \mathbf{y}_c$为中心化向量:

$$\mathbf{x}_c = (x_1 - \bar{x}, ..., x_n - \bar{x})^{\mathsf{T}} = \mathbf{x} - \mathbf{1}\bar{x},$$

$$\mathbf{y}_c = (y_1 - \bar{y}, ..., y_n - \bar{y})^{\mathsf{T}} = \mathbf{y} - \mathbf{1}\bar{y}.$$



Pearson相关系数度量了向量之间的相似性/角度:

$$r_{xy} = \frac{\mathbf{x}_c^{\mathsf{T}} \mathbf{y}_c}{\| \mathbf{x}_c \| \cdot \| \mathbf{y}_c \|} = \cos(\theta_{\mathbf{x}_c \mathbf{y}_c}), \quad -1 \le r_{xy} \le 1$$

相关系数的分布

以后将证明如下命题1,2的一般形式

命题1(正态总体,精确分布). 假设 $(x_1, y_1),\ldots,(x_n, y_n)$ iid，假设$y_i|x_i$服从一元正态分布，则当相关系数$\rho_{xy} = 0$时，有

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$$

注：若 (x_i, y_i)服从二元正态，则$y_i \mid x_i$服从一元正态

命题2(非正态总体,渐近分布). 假设 $(x_1, y_1),...,(x_n, y_n)$ iid, 设$\rho = \rho_{xy}$为总体相关系数, r为样本相关系数，则

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

注：符号$\xrightarrow{d}$表示依分布收敛,上式表明当n足够大时,

$\sqrt{n}(r-\rho)$近似地服从$N\left(0,(1-\rho^2)^2\right)$,或

近似地, $r \sim N\left(\rho,(1-\rho^2)^2/n\right)$

特别地，当$\rho = 0$时, $z = \sqrt{n} \times r \xrightarrow{\mathrm{d}} N(0,1)$, $n \to \infty$.

相关系数的小样本检验/精确检验

$$H_0 : \rho_{xy} = 0$$

假设样本服从正态的条件下，由命题1我们有如下$t-$检验

t-检验(小样本检验/精确检验):

若$(x_1, y_1), ..., (x_n, y_n)$ iid, 假设$y|x \sim$ 正态。原假设为$H_0 : \rho_{xy} = 0$。

取检验统计量 $T = \sqrt{n-2}\dfrac{r}{\sqrt{1-r^2}}$, 当$|T| \geq t_{n-2}(\alpha/2)$时否定原假设。

对于给定的T值，p值$= P(|T^*| \geq |T| \,|\, T)$, 其中r.v.$T^* \sim t_{n-2}$。

大样本检验

不假设样本服从正态的条件下，由命题2我们有如下大样本检验：

相关性的 z - 检验

对于零假设$H_0:\rho_{xy}=0$,检验统计量取为$z=\sqrt{n}\times r$,检验准则为：

当$|z|\geq z_{\alpha/2}$时在α水平下拒绝原假设

p值$\approx P(|Z^*|>|z|\,|z)=2(1-\Phi(|z|),Z^*\sim N(0,1).$

相关性的卡方检验

通常人们使用等价的卡方检验$z^2=nr^2$,原假设下z^2近似服从χ_1^2.

检验准则：$z^2\geq\chi_1^2(\alpha)=(z_{\alpha/2})^2$,在$\alpha$水平下拒绝原假设。

例1. 样本相关系数$r_{xy}=0.1$, 是否可以认为x, y不相关？

不一定,需要考虑样本量大小。

如果$n=900,\ \sqrt{n}\times r=3>1.96$

如果$n=100,\ \sqrt{n}\times r=1<1.96$

判断相关性大小不能仅仅看相关系数，还应考虑到样本量：
以 $\sqrt{n}\,r$ 绝对值是否超过1.96做判断（水平0.05），或
以 nr^2 是否超过$3.84=1.96^2$做判断（水平0.05）。

2个评注

注1: 我们定义了两个检验统计量

$$T=\sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}\text{(正态总体)};\quad z=\sqrt{n}r\text{(一般总体)};$$

原假设下$\sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}\approx\sqrt{n}r$ (原假设下$\rho=0, r\to 0$).

注2. 2×2列联表的Pearson卡方、两样本t-检验都具有相关检验的形式。

两样本 t-检验是相关性检验(作业)

假设第一组$y_1,..,y_{n_0}\ iid \sim N(\mu_0,\sigma^2)$, 组号$x_i=0$

第二组$y_{n_0+1},..,y_{n_0+n_1}\ iid \sim N(\mu_1,\sigma^2)$, 组号$x_i=1$

$r=r_{xy}$为样本相关系数,则两样本t-检验 $T=\sqrt{n-2}\,\dfrac{r}{\sqrt{1-r^2}}$.

2×2列联表的Pearson卡方统计量$X^2=nr^2$(作业）

$(x_1,y_1),...,(x_n,y_n)$ iid, 如果x_i,y_i都是0-1伯努利随机变量.

则Pearson卡方统计量$X^2=nr^2$.

虚框内容略过

2×2表格的Pearson卡方$X^2 = nr^2$的证明：

记$a = \#\{i : x_i = y_i = 1\} = \sum x_i y_i$,

$n_1 = \#\{i : x_i = 1\} = \sum x_i$, $\quad n_0 = n - n_1$

$m_1 = \#\{i : y_i = 1\} = \sum y_i$, $\quad m_0 = n - m_1$

$b = n_1 - a = \#\{i : x_i = 1, y_i = 0\} = \sum x_i(1 - y_i)$.

$c = m_1 - a, d = n_0 - c = m_0 - b$.

计数a, b, c, d组成右表,.

x \ y	1	0	总计
1	a	b	n_1
0	c	d	n_0
总计	m_1	m_0	n

另外一方面，$r = \dfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}} = \dfrac{a - n\times n_1/n \times m_1/n}{\sqrt{n_1 - n_1^2/n}\sqrt{m_1 - m_1^2/n}}$,

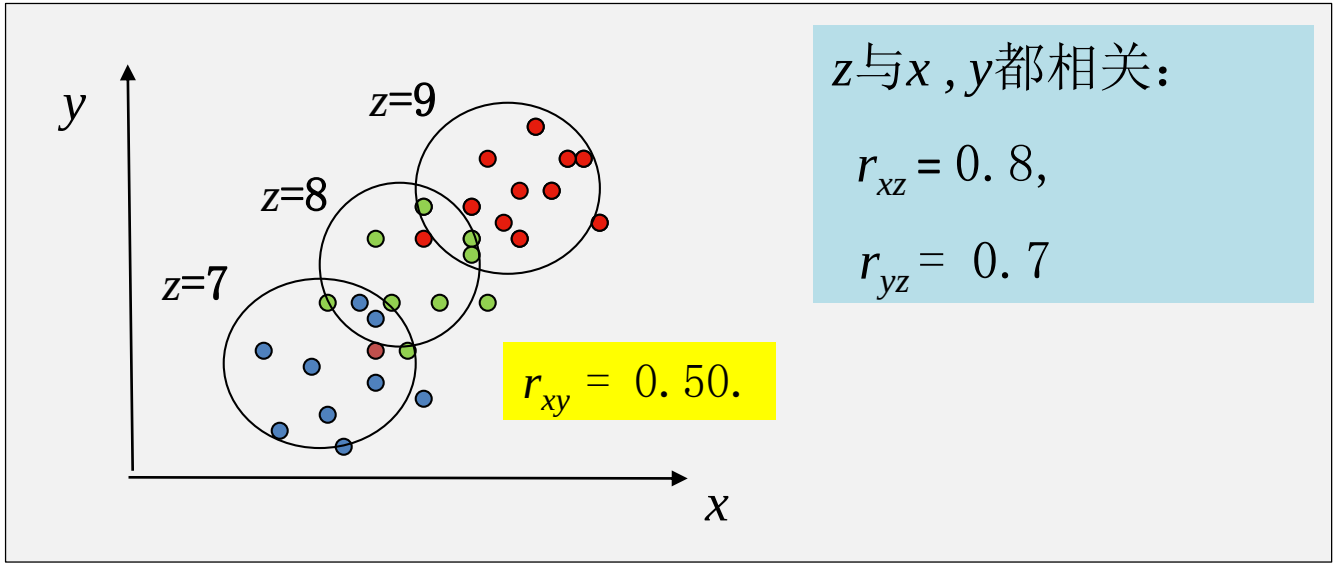
则可以验证$nr^2 = \dfrac{n(a - n\times n_1/n\times m_1/n)^2}{(n_1 - n_1^2/n)(m_1 - m_1^2/n)} = \ldots = \dfrac{n(ad - bc)^2}{n_1 n_0 m_1 m_0} = X^2$.

(X^2为独立性的Pearson卡方检验统计量)

偏相关系数

偏相关系数与控制变量

例1. 调查100名 7-9岁儿童(z=7-9)，发现阅读能力y与身高x正相关，相关系数 $r_{xy} = 0.50$.



控制年龄，即给定年龄 z 时， 数据分布呈球状，x 与 y 不相关！

我们以偏相关系数度量这种“条件”相关性。

随机向量的协方差矩阵

向量：黑体小写字母

定义：随机向量$\mathbf{x}, \mathbf{y}$的协方差矩阵定义为

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = E(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{y} - \boldsymbol{\mu}_y)^{\mathsf{T}},$$

其中$\boldsymbol{\mu}_x = E(\mathbf{x}), \boldsymbol{\mu}_y = E(\mathbf{y}).$

特别地，随机向量$\mathbf{x}$的方差-协方差矩阵

$$\operatorname{var}(\mathbf{x}) \; or \; \operatorname{cov}(\mathbf{x}) = \operatorname{cov}(\mathbf{x}, \mathbf{x}) = E(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^{\mathsf{T}}.$$

命题3. 对于常数矩阵A,从常数向量$\mathbf{b},\mathbf{c}$

- $E\left(A\mathbf{x}+\mathbf{b}\right)=AE(\mathbf{x})+\mathbf{b},$
- $\operatorname{cov}(\mathbf{x}_1+\mathbf{x}_2,\mathbf{y}_1+\mathbf{y}_2)$
  $=\operatorname{cov}(\mathbf{x}_1,\mathbf{y}_1)+\operatorname{cov}(\mathbf{x}_1,\mathbf{y}_2)+\operatorname{cov}(\mathbf{x}_2,\mathbf{y}_1)+\operatorname{cov}(\mathbf{x}_2,\mathbf{y}_2).$
- $\operatorname{cov}\left(A\mathbf{x}+\mathbf{b},B\mathbf{y}+\mathbf{c}\right)=\ A\operatorname{cov}\left(\mathbf{x},\mathbf{y}\right)B^{\mathsf{T}}$
- $\operatorname{var}\left(A\mathbf{x}+\mathbf{b}\right)=A\operatorname{var}\left(\mathbf{x}\right)A^{\mathsf{T}}.$

$$\text{证}:\operatorname{cov}\left(A\mathbf{x}+\mathbf{b},B\mathbf{y}+\mathbf{c}\right)\overset{\text{定义}}{=}\mathrm{E}(A\mathbf{x}+\mathbf{b}-A\boldsymbol{\mu}_x-\mathbf{b})(B\mathbf{y}+\mathbf{c}-B\boldsymbol{\mu}_y-\mathbf{c})^{\mathsf{T}}$$

$$=E(A\mathbf{x}-A\boldsymbol{\mu}_x)(B\mathbf{y}-B\boldsymbol{\mu}_y)^{\mathsf{T}}$$

$$=A\left[E(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{y}-\boldsymbol{\mu}_y)^{\mathsf{T}}\right]B^{\mathsf{T}}=A\operatorname{cov}\left(\mathbf{x},\mathbf{y}\right)B^{\mathsf{T}}$$

随机向量的“不相关化”

(协方差矩阵的对角化)

假设任意$q\times 1$随机向量$\mathbf{y}$，$p\times 1$随机向量$\mathbf{x}$, 方差-协方差矩阵：

$$\Sigma=\operatorname{cov}\begin{pmatrix}\mathbf{y}\\ \mathbf{x}\end{pmatrix}=\begin{pmatrix}\operatorname{var}(\mathbf{y}) & \operatorname{cov}(\mathbf{y},\mathbf{x})\\ \operatorname{cov}(\mathbf{x},\mathbf{y}) & \operatorname{var}(\mathbf{x})\end{pmatrix}=\begin{pmatrix}\Sigma_{\mathbf{yy}} & \Sigma_{\mathbf{yx}}\\ \Sigma_{\mathbf{xy}} & \Sigma_{\mathbf{xx}}\end{pmatrix}$$

我们希望从$\mathbf{y}$中消去与$\mathbf{x}$有关的成分：求$A_{q\times p}$使得$\mathbf{y}-A\mathbf{x}$与$\mathbf{x}$不相关.

解：$0=\operatorname{cov}(\mathbf{y}-A\mathbf{x},\mathbf{x})=\operatorname{cov}(\mathbf{y},\mathbf{x})-A\operatorname{cov}(\mathbf{x},\mathbf{x})$

$=\Sigma_{\mathbf{yx}}-A\Sigma_{\mathbf{xx}}\Rightarrow A=\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Rightarrow$ 所求的不相关向量为$\mathbf{y}-\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x}$

命题4.

(1) $\mathbf{y}^{\perp} \widehat{=} \mathbf{y} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x}$是$\mathbf{y}$的关于$\mathbf{x}$ 的不相关化, 即$\mathrm{cov}(\mathbf{y}^{\perp}, \mathbf{x}) = 0.$

(2) $\mathrm{var}(\mathbf{y}^{\perp}) = \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}} \overset{\text{记为}}{=} \Sigma_{\mathbf{yy}\bullet\mathbf{x}} \geq 0$ (半正定).

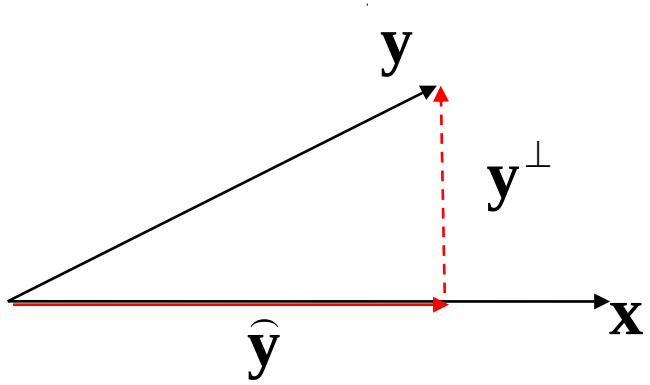
验证(2): $\mathrm{var}(\mathbf{y}^{\perp}) = \mathrm{cov}(\mathbf{y} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x},\ \ \mathbf{y} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathbf{x})$

$$= \mathrm{cov}(\mathbf{y},\mathbf{y}) - \mathrm{cov}(\mathbf{y},\mathbf{x})\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathrm{cov}(\mathbf{x},\mathbf{y}) + \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\mathrm{cov}(\mathbf{x},\mathbf{x})\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}$$

$$= \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{yy}\bullet\mathbf{x}}$$

记$\hat{\mathbf{y}} = \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}$，由命题4我们有"正交"分解：

$$\mathbf{y} = \mathbf{y}^{\perp} + \hat{\mathbf{y}}, \quad \text{其中} \operatorname{cov}(\mathbf{y}^{\perp}, \hat{\mathbf{y}}) = 0.$$



正交分解的两边同时求方差，得方差分解/勾股定理：

$$\operatorname{var}(\mathbf{y}) = \operatorname{var}(\mathbf{y}^{\perp}) + \operatorname{var}(\hat{\mathbf{y}})$$

$$\Sigma_{yy} = \Sigma_{yy\bullet x} + \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$$

关于方差分解/勾股定理：$\Sigma_{yy} = \Sigma_{yy\bullet x} + \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$

$\hat{\mathbf{y}}$看作是$\mathbf{y}$中能被$\mathbf{x}$解释的部分（$\mathbf{y}$在$\mathbf{x}$上的投影).

$\mathrm{var}(\hat{\mathbf{y}}) = \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ 相对于Σ_{yy}的大小反映了$\mathbf{x}$对$\mathbf{y}$的解释能力，但两者都是矩阵，我们用矩阵$\Sigma_{yy}^{-1/2}\left[\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\right]\Sigma_{yy}^{-1/2}$的数字特征，比如行列式、迹、特征根等度量$\mathbf{x},\mathbf{y}$的相关程度。

$\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}$的最大特征根

$$\lambda_{\max}^{1/2}\left(\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}\right),$$

称为第一典则相关系数,度量了两个随机向量$\mathbf{x},\mathbf{y}$的相关程度。