Lab-1: 相关分析 **2023** 年 **10** 月 **8** 日

内容: 1. 安装 R 包 2. 相关分析
任务: 阅读下面的材料, 重复代码命令（手工输入!），并做练习 1-3，提交结果。

1 安装程序包 (packages)

我们需要如下 R 包：corrplot (相关系数可视化)；alr4 (Weisberg: Linear Regression, 4th ed. 一书的数据集)。在 R 环境中，安装命令如下：

```
> install.packages( c("corrplot","alr4") ) #安装
> library(corrplot) #载入corrplot
```

2 相关系数

下面我们学习与相关系数有关的几个函数，包括计算相关系数的函数 cor, 相关检验函数 cor.test, 相关系数矩阵可视化函数 corrplot。另外，我们将以偏相关系数矩阵的计算为例，学习 R 函数的写法。

相关系数有关的函数

cor, cor.test, corrplot (package:corrplot), r2rp(自编),

我们以 R 自带数据集 state.x77 为例演示 R 函数。state.x77 给出了美国 50 个州 1977 年的如下信息:

`Population`(人口), `Income`(人均收入), `Illiteracy`(文盲率), `Life Exp`(平均寿命), `Murder`(凶杀案数目, 每 10 万人), `HS Grad`(高中学历比率), `Frost`(寒冷天气数), `Area`（面积）

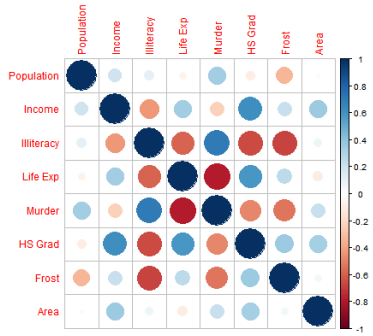
2.1 相关系数及其可视化

相关系数矩阵如下

```
> cor(state.x77)
Population Income Illiteracy Life Exp Murder HS Grad Frost  Area
Population       1.00   0.21       0.11   -0.07   0.34  -0.10 -0.33  0.02
Income           0.21   1.00      -0.44    0.34  -0.23   0.62  0.23  0.36
Illiteracy       0.11  -0.44       1.00   -0.59   0.70  -0.66 -0.67  0.08
Life Exp        -0.07   0.34      -0.59    1.00  -0.78   0.58  0.26 -0.11
Murder           0.34  -0.23       0.70   -0.78   1.00  -0.49 -0.54  0.23
HS Grad         -0.10   0.62      -0.66    0.58  -0.49   1.00  0.37  0.33
Frost           -0.33   0.23      -0.67    0.26  -0.54   0.37  1.00  0.06
Area             0.02   0.36       0.08   -0.11   0.23   0.33  0.06  1.00
```

使用程序包 corrplot 中的函数 corrplot 将上述相关系数矩阵画图表示（相关系数绝对值越大，圆圈的越大，红色代表正数，蓝色代表负数）:

```
> (R=cor(state.x77))     #Pearson correlation coefficients
> corrplot(R, diag=F)  # plot correlation coefficients
```



2.2 相关性检验

通常认为温度越高的地区，犯罪率越高。Murder 与 Frost 的相关系数等于 −0.5388834, 下面我们检验 Murder 与 Frost 是否显著相关：

```
> r=R["Murder","Frost"]
[1] -0.5388834
> cor.test(state.x77[, "Murder"], state.x77[, "Frost"]) # or
> cor.test(~ Murder+Frost,data=state.x77)

Pearson's product-moment correlation

data:  state.x77[, "Murder"] and state.x77[, "Frost"]
t = -4.4321, df = 48, p-value = 5.405e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.7106377 -0.3065115
sample estimates:
cor
-0.5388834
```

上述检验用 Pearson 相关系数度量相关程度并假设数据来自于正态总体，检验统计量为

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

原假设下 $t \sim t_{n-2}$. 输出结果给出了 t 检验的值 $t = -4.4321$, 自由度 $df = 48$, p 值 =5e-5. 最后还给出了相关系数值 −0.538, 95% 置信区间 $[-0.71, -0.31]$

如果不假设正态总体，那么可采用大样本检验

$$z = \sqrt{n-2}r \text{ 或 } \sqrt{n}r \text{ 原假设下近似 } z \sim N(0,1)$$

函数 cor.test 并没提供该检验，手工计算如下

```
r=R["Murder","Frost"]
#r=cor(state.x77[, "Murder"], state.x77[, "Frost"])
# r= -0.5388834
z= sqrt(50-2)*r
pvalue=2*(1-pnorm(abs(z)))
pvalue
[1]  0.0001888419
```

2.3 置换检验

t 检验中总体的正态假设无法验证，而大样本 z 检验需要较大的样本量。所以上述两个检验都不一定适用于当前数据。为了给出一个更为合理的结论，我们使用置换检验方法计算检验统计量在原假设下的（精

确）分布，计算精确的 p 值。原假设为 x,y 独立。数据为 $(x_i,y_i), i=1,...,n$。取检验统计量 $T=T(r)$ 为 r 的某个函数, 比如 $T=r$ (或 t, 或 z，不影响下面得到的 p 值），

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

这是基于原始数据计算得到的相关系数。原假设成立时，x,y 独立，置换数据 $(x_{\sigma(i)},y_i), i=1,...,n$ 与原始数据出现的可能性相同，其中 $(\sigma(1),...,\sigma(n))$ 是 $(1,2,...,n)$ 的一个置换。基于置换数据计算相关系数

$$r_{per}=\frac{\sum_{i=1}^n(x_{\sigma(i)}-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n(x_{\sigma(i)}-\bar{x})^2}\sqrt{\sum_{i=1}^n(y_i-\bar{y})^2}}$$

假设 N 次随机置换得到 N 个检验统计量 $r_{per}^{(1)},...,r_{per}^{(N)}$, 我们认为这些是原假设成立的时候从总体 T 得到的一批随机样本，因而可以用来估计 T 在原假设下的分布。特别地

$$p=\frac{1}{N}\sum_{k=1}^{N}\{|r_{per}^{(k)}|\geq|r|\}. \tag{1}$$

是原假设下随机置换计算得到的相关系数绝对值超过原始数据相关系数的概率 (N 很大)。

由于置换数据相关性系数公式中的分母不依赖于置换 $\sum_{i=1}^n(x_{\sigma(i)}^{(k)}-\bar{x})^2=\sum_{i=1}^n(x_i-\bar{x})^2$, 从 p 值计算公式 (1) 来看，不等式 $|r_{per}^{(k)}|\geq|r^{(0)}|$ 两边的分母相同，故检验统计量可取为

$$T=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}$$

而置换数据的版本为

$$T_{per}=\sum_{i=1}^{n}x_{\sigma(i)}y_i-n\bar{x}\bar{y}$$

```
x=state.x77[, "Frost" ]
y=state.x77[, "Murder" ]
n=length(x)
#r0=cor (x,y )
t0=sum(x*y) -n*mean(x)*mean(y)
R_per=NULL
t_per=NULL
N=1000000
for (i in 1:N){
        x_per=sample(x) # 置换 x
        #R_per[i]=cor(x_per,y) # 置换后的相关系数
        t_per[i]=sum(x_per*y) -n*mean(x)*mean(y) # 置换后的 t
}
#p1=mean(abs(R_per)>= abs(r0) )
p2=mean(abs(t_per)>= abs(t0) )
p2
[1] 6.4e-05
```

练习 1. **我们可随机产生数据，检查** t-**检验（函数** cor.test）**与置换检验的结果（**p **值）几乎相同。**

```
n=20
x=rnorm(n)
y=rnorm(n)
r0=cor (x,y)
R_per=NULL
N=100000
for (i in 1:N){
        x_per=sample(x) # 置换 x
        R_per[i]=cor(x_per,y) # 置换后的相关系数
}
pvalue.per=mean(abs(R_per)>= abs(r0) )
 cor.test(x,y)->tmp
 pvalue.ttest = tmp$p.value
 pvalue.ttest
pvalue.per
```

2.4 非参数检验

非参数型的相关系数：Kendall's tau, Spearman's rho，在函数 cor, cor.test 中指定 method="kendall" 或"spearman" (缺省为"pearson"). 以 Spearman's rho 为例，假设数据为 $(x_i, y_i), i = 1,...,n$, 假设 x_i 在所有 x 中的秩（排名）为 R_i，y_i 在所有 y 中的秩（排名）为 S_i，Spearman's rho 定义为 $(R_i, S_i), i = 1,...,n$ 的 Pearson 相关系数

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

函数选项 exact 用于选择 p 值的计算方式是精确（缺省）还是近似。

```
> x=c(2 , -2,-11, 3, 4 )
> y=c(0,-1,-3, 99,7)
> rankx=rank(x)
> ranky=rank(y)
> rankx
[1] 3 2 1 4 5
> ranky
[1] 3 2 1 5 4
> pearson=cor(x,y)
> pearson
[1] 0.407719
> spearman=cor(rankx,ranky)
> spearman
[1] 0.9
> cor.test (x,y,method = "spearman")

Spearman's rank correlation rho

data:  x and y
S = 2, p-value = 0.08333
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.9
```

练习 2. **上述例子中** Spearman **检验的精确** p **值为** 0.08333，**它是基于** spearman **系数的原假设下的精确分布计算得到的，该精确** p **值可用置换检验方法逼近。试分别进行** 1000，10000，100000 **次置换，按照公式** (1) **分别计算置换检验** p **值，观察这些** p **值是否接近** 0.08333。

3 偏相关系数

3.1 偏相关系数的计算及检验

偏相关系数检验可以使用大样本结果：

$$\sqrt{n} r_{xy\bullet z} \xrightarrow{d} N(0,1)$$

或更精确地

$$\sqrt{n-p} r_{xy\bullet z} \xrightarrow{d} N(0,1)$$

其中 p 是变量个数，n 是样本量。

注：（1）也可使用正态假设下的 t 检验（参见第三讲最后 P22 页）。但一般情况下不能运用置换方法，除非控制变量 z 是属性变量，比如 $z = 0,1$, 则置换须在 $z = 1$ 的组内和 $z = 0$ 的组内分别进行。（2）R 包 ppcor 中的函数 pcor 可用于计算偏相关系数，pcor.test 用于检验。你可自行研究其用法，以及搜索是否有更好的关于偏相关系数的程序包。

3.2 偏相关系数矩阵的计算

第 4 讲给出了偏相关系数的计算公式：

给定一个相关系数矩阵或协方差矩阵 $\Sigma_{k\times k}$, 记 $\Omega = \Sigma^{-1} = (\omega_{ij})$, 则变量 i, j 的偏相关系数

$$\rho_{ij\bullet_{\text{other}}} = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$$

记 $D = diag(\Omega)$, 则偏相关系数矩阵 $R_{partial} = (\rho_{ij\bullet_{\text{other}}})$

$$R_{partial} = -D^{-1/2}\Omega D^{-1/2} + 2I_k$$

编写下述函数 r2rp 用于计算偏相关系数矩阵:

```
r2rp =function(R){  #R: correlation matrix or covariance matrix
Omega=solve(R)
d=diag(Omega)
D0.5=diag(1/sqrt(d))
Rp=- D0.5%*%Omega%*%D0.5
diag(Rp)=1
return(Rp)
} #end

#run
r2rp(R=cor(state.x77))
```

4 典则相关系数和决定系数

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

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

令 $q\times q$ 矩阵

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

当 $q > 1$ 时，Φ 是一个 $q\times q$ 矩阵（$0 \le \Phi \le I_q$），该矩阵的最大特征根的平方根 $\lambda_{\max}^{1/2}(\Phi)$ 为第一典则相关系数，第二大特征根的平方根称为第二典则相关系数，等等。$q = 1$ 时，

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

是介于 0 和 1 之间的实数，称为决定系数。

练习 3. R 数据集 ability.cov 给出了 112 个儿童的 6 项测试成绩的协方差矩阵，6 个科目分别是 general, picture, blocks , maze, reading, vocab （综、绘画、积木、迷宫、阅读、词汇量）。

(a) 求 picture 与 reading 的偏相关系数

(b) 求 general 与其余 5 个变量 (picture, blocks , maze, reading, vocab) 之间的相关性大小 (决定系数）

(c) 求 (picture, blocks , maze）与（reading, vocab) 之间的相关性大小（第一典则相关系数）。