# 多元统计复习

## 第一章 多元正态分布

- 相关系数恒满足：$|\rho_{X,Y}| \leq 1$
- 如果 $X, Y$ 之间存在线性函数关系，则$|\rho_{X,Y}| = 1$

  此时，称 $X, Y$**完全相关**

  当 $\rho = 1$ 时，称**完全正相关**

  当 $\rho = -1$ 时，称**完全负相关**
- 如果 $\rho_{X,Y} = 0$ ，则称 $X, Y$ **不相关。**
- 如果 $0 < |\rho_{X,Y}| < 1$ 则称 $X, Y$ 不完全相关，

  当 $\rho > 0$ 时，称为**正相关。**

  当 $\rho < 0$ 时，称为**负相关。**

从相关系数和协方差的定义可以知道：

独立 $\Rightarrow$ 不相关　　不相关 $\nRightarrow$ 独立

独立 $\Rightarrow$ 没有关系 $\Rightarrow$ 没有线性关系 $\Rightarrow$ 不相关.

不相关 $\Rightarrow$ 没有线性关系，但是可能存在非线性关系 $\nRightarrow$ 独立.

### 由标准正态随机向量线性组合得到：

$$
\begin{gathered}
\text{定义} 2.1 (\text{构造性定义}) \text{如果} y = \mu + Ax, \\
\text{其中} x = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}, x_1, x_2, \cdots, x_n \text{是独立同分布的标准正态分布变量}, \\
\mu \text{是} p \text{维常数向量}, A \text{为} p \times n \text{常数矩阵}, \text{则称} y \text{服从} p \text{元正态分布}, \\
\text{记为} y \sim N_p(\mu, \Sigma), \text{其中} \Sigma = AA^{\mathrm{T}}.
\end{gathered}
$$

## 第二章 多元线性模型

### 多元正态分布

**随机向量均值：** $\mu = E(X) = \sum_i x_i p_i$

**样本均值：** $x = \frac{1}{n} \sum_{i=1}^{n} x_i$

**协方差：协方差表示的是两个变量总体误差的期望**

$$
\begin{aligned}
cov(X, Y) &= E[(X - E(X))(Y - E(Y))] \\
&= E(XY) - 2E(Y)E(X) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}
$$

协方差性质：

$$
\begin{aligned}
&cov(X,X)=DX\\
&cov(X,Y)=cov(Y,X)\\
&cov(aX,bY)=ab\cdot cov(X,Y),a,b\text{ 为任意常数}\\
&cov(C,X)=0,C\text{ 为任意常数}\\
&cov\left(X_1+X_2,Y\right)=cov\left(X_1,Y\right)+cov\left(X_2,Y\right)
\end{aligned}
$$

如果$X,Y$相互独立，则$cov(X,Y)=0$。反过来不成立

如果$cov(X,Y)=0,X,Y$不一定相互独立。

对于方差存在的随机变量$X,Y$,有$D(X\pm Y)=DX+DY\pm 2cov(X,Y)$

当$X,Y$相互独文时,$D(X\pm Y)=DX+DY$

协方差阵：

$$
\begin{aligned}
\Sigma &= \begin{bmatrix}
D\left(X_1\right) & cov\left(X_1,X_2\right) & \cdots & cov\left(X_1,X_p\right)\\
cov\left(X_1,X_2\right) & D\left(x_2\right) & \cdots & cov\left(X_2,X_p\right)\\
\vdots & \vdots & & \\
cov\left(X_p,X_1\right) & cov\left(X_p,X_2\right) & \cdots & D\left(X_p\right)
\end{bmatrix}\\
&=\left(\sigma_{ij}\right)_{p\times p}
\end{aligned}
$$

**相关系数：用相关系数表示两变量间的线性关系并判断其密切程度**

$$
\rho=\frac{cov(x,y)}{\sqrt{var(x)var(y)}}=\frac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}}
$$

**密度函数：**

公式为：设 $y\sim N_p(\mu,\Sigma),\Sigma>0$ ,则 $y$ 的密度函数为：

$$
f(y)=\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}
$$

## 多元线性模型

模型一般形式：

$$
y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon
$$

对每一组观测值

$$
y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\cdots+\beta_kx_{ki}+\varepsilon_i
$$

随机误差:

随机误差也称为偶然误差和不定误差，是由于在测定过程中一系列有关因素微小的随机波动而形成的具有相互抵偿性的误差，是不可抗因素带来的影响

随机误差项 $\mu$ 的方差的无偏估计量为：

$$
\delta^2=\frac{\sum e_i^2}{n-k-1}=\frac{e'e}{n-k-1}
$$

经典假设：

假定1：解释变量是非随机的，在重复抽样中，解释变量取固定值，且相互之间互不相关，表明了模型中的解释变量和随机干扰项对被解释变量的影响是完全独立的。

假定2：随机干扰项与解释变量之间不相关。

$$cov\left(u_i, X_{ji}\right)=0, j=1,2, \ldots, k; i=1,2, \ldots, n$$

此假定说明解释变量与随机干扰项 $u_i$ 相互独立，互不相关，它们对解释变量 $Y_i$ 的影响同样也是独立的，用矩阵表示就是 $E(X'u)=0$

假定3：随机干扰项服从零均值，同方差，零协方差

$$\begin{aligned}
&E\left(u_i\right)=0 \\
&var\left(u_i\right)=E\left(u_i^2\right)=\sigma^2 \\
&cov\left(u_i, u_j\right)=E\left(u_i, u_j\right)=0
\end{aligned}$$

假定4：随机干扰项服从正态分布，即

$u_i \tilde{} N\left(0, \sigma^2\right) \quad i=1,2, \ldots, n$

矩阵形式表示为：$u \tilde{} N\left(0, \sigma^2 I\right)$

假定5：正确设定回归模型

与一元回归模型一样，多元回归模型的正确设定也有三个方面的要求：

1.选择正确的变量进入模型，

2.对模型的形式进行正确的设定，

3.对模型的解释变量被解释变量及随机干扰项做了正确的假定

## 最小二乘估计(OLS):

原理：根据被解释变量的所有观测值与估计值之差的平方和最小的原则求得参数估计量

对于随机抽取的 $n$ 组观测值 $\left(Y_i, X_{ji}\right), i=0,1,2, \ldots, k$

样本参数的参数估计值：$\hat{Y}=\hat{\beta}_0+\beta_1 \hat{x}_{1 i}+\beta_2 \hat{x}_{2 i}+\cdots+\beta_k \hat{x}_{k i} \quad i=1,2, \ldots, k$

根据最小二乘原理，参数估计值应该是下列方程组的解：

$$\begin{cases}\frac{\partial}{\partial \hat{\beta}_0} Q=0 \\ \frac{\partial}{\partial \hat{\beta}_1} Q=0 \\ \frac{\partial}{\partial \hat{\beta}_2} Q=0 \\ \cdots \\ \frac{\partial}{\partial \hat{\beta}_k} Q=0\end{cases}$$

其中，

$$\begin{aligned}
Q &=\sum_{i=1}^n e_i^2=\sum_{i=1}^n\left(Y_i-\hat{Y}_i\right)^2 \\
&=\sum_{i=1}^n\left(Y_i-\left(\hat{\beta}_0+\hat{\beta}_1 X_{1 i}+\hat{\beta}_2 X_{2 i}+\cdots+\hat{\beta}_k X_{k i}\right)\right)^2 .
\end{aligned}$$

**待估计参数的正规方程组：**

$$\begin{cases} \Sigma\left(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki}\right)=\Sigma Y_i \\ \Sigma\left(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki}\right)X_{1i}=\Sigma Y_iX_{1i} \\ \Sigma\left(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_{2i}X_{2i}+\cdots+\hat{\beta}_kX_{ki}\right)X_{2i}=\Sigma Y_iX_{2i} \\ \vdots \\ \Sigma\left(\hat{\beta}_0+\hat{\beta}_1X_{1i}+\hat{\beta}_2X_{2i}+\cdots+\hat{\beta}_kX_{ki}\right)X_{ki}=\Sigma Y_iX_{ki} \end{cases}$$

解该 $(k+1)$ 个方程组成的线性代数方程组，即可得到 $(k+1)$ 个待估参数的估计值 $\hat{\beta}_j$，$j=0,1,2,\ldots,k$。

正规方程组的矩阵形式：

$$\begin{pmatrix} n & \Sigma X_{1i} & \cdots & \Sigma X_{ki} \\ \Sigma X_{1i} & \Sigma X_{1i}^2 & \cdots & \Sigma X_{1i}X_{ki} \\ \cdots & \cdots & \cdots & \cdots \\ \Sigma X_{ki} & \Sigma X_{ki}X_{li} & \cdots & \Sigma X_{ki}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{pmatrix}$$

即

$$(X'X)\hat{\beta}=X'Y$$

由于 $X'X$ 满秩，故有

$$\hat{\beta}=(X'X)^{-1}X'Y$$

**OLS性质：**

1. 线性特征：$\hat{\beta}=(X'X)^{-1}X'Y$

   $\hat{\beta}$ 是 $Y$ 的线性函数，因 $(X'X)^{-1}X'$ 是非随机或取固定值的矩阵

2. 无偏特性 $E(\hat{\beta}_k)=\beta_k$

3. 最小方差特性（有效性）

   在 $\beta_k$ 所有的线性无偏估计中，OLS估计 $\hat{\beta}_k$ 具有最小方差

结论：在古典假定下，多元线性回归的OLS估计式是最佳线性无偏估计式

**统计检验**

拟合优度检验（$R^2$ 检验–离差平方和，回归平方和与残差平方和以及可决系数定义），显著性检验（模型检验（ $F$ 检验 ）与回归系数检验（ $t$ 检验））

**重点：**

## 拟合优度检验

目的是检验样本数据点聚集在回归线周围的密集程度，从而评价回归方程对样本数据的代表程度。评价对应的构建模型与样本适配程度的评价指标。

### $R^2$ 检验：

前提：

总离差平方和分解：

对于有 $k$ 个解释变量的多元线性回归模型：

$$e_i = Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

其对应的回归方程为：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

将 $Y_i$ 与其平均值 $\overline{Y}$ 之间的离差分解如下:

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

$$\text{总离差平方和：} \quad TSS = \sum \left(Y_i - \bar{Y}\right)^2$$

$$\text{回归平方和：} \quad ESS = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

$$\text{残差平方和：} \quad RSS = \sum \left(Y_i - \hat{Y}_i\right)^2$$

$$\begin{aligned} TSS &= \sum \left(Y_i - \bar{Y}\right)^2 = \sum \left[\left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)\right]^2 \\ &= \sum \left(Y_i - \hat{Y}_i\right)^2 + 2\sum e_i \left(\hat{Y}_i - \bar{Y}\right) + \sum \left(\hat{Y}_i - \bar{Y}\right)^2 \\ &= \sum \left(Y_i - \hat{Y}_i\right)^2 + 0 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2 \\ &= RSS + ESS \end{aligned}$$

总离差平方分解为回归平方和与残差平方和两部分
多元样本可决系数：

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

可由回归平方和占残差平方和的比重来衡量样本回归线对样本观测值的拟合程度

$$0 \le ESS \le TSS$$
$$0 \le R^2 \le 1$$

$R^2$ 数值越接近于1，说明拟合程度越高，否则越低.

修正样本可决系数

$$\bar{R}^2=1-\frac{RSS/(n-k-1)}{TSS/(n-1)}$$

$$\bar{R}^2=1-\frac{RSS}{TSS}\cdot\frac{n-1}{(n-k-1)}=1-\left(1-R^2\right)\frac{n-1}{n-k-1}$$

## 回归方程显著性检验（ $F$ 检验）

检验被解释变量与所有解释变量之间的线性关系是否显著，用线性模型来描述它们之间的关系是否恰当，是指在一定的显著性水平下，从总体进行解释
对于多元线性回归模型

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+u_i\quad i=1,2,\cdots,n$$

是必须进行显著性检验的
原假设：$H_0:\beta_1=0,\beta_2=0,\cdots,\beta_k=0$

备择假设：$H_1:\beta_j(j=1,2,\cdots,k)$ 不全为零

检验思想来自于 $TSS=ESS+RSS$

在 $H_0$ 成立时，构建以下统计量：

$$F=\frac{ESS/k}{RSS/(n-k-1)}$$

则该统计量服从自由度为 $(k,n-k-1)$ 的 $F$ 分布。

如果$F>F_\alpha(k,n-k-1)$，则在 $\alpha$ 显著性水平下拒绝原假设，即模型的线性关系显著成立，模型通过方程显著性检验。

在例$2-1$中

$$F=19462.76$$
$$n=32,n-k-1=32-2-1=29$$
$$\alpha=0.05,F_{0.05}(2,29)=3.33$$
$$F=19462.76>F_{0.05}(2,29)=3.33$$

故模型总体是显著的。

拟合优度与显著性检验的关系：

$$\overline{R^2}=1-\frac{n-1}{n-k-1+kF}$$

$$F=\frac{\overline{R^2}/k}{\left(1-\overline{R^2}\right)/(n-k-1)}$$

检验 $F$ 检验的原假设等价于检验 $\overline{R^2}=0$

## 回归系数的显著性检验（ $t$ 检验）

研究解释变量(X)能否有效解释被解释变量(Y)的线性变化，他们能否保留在线性回归方程中。
该检验的原假设与备择假设为：

$$H_0 : \beta_j = 0, j = 1, 2, \cdots, k$$
$$H_1 : \beta_j \neq 0, j = 1, 2, \cdots, k$$

构造的 $t$ 检验统计量为：$t = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$

原假设成立时，检验统计量 $t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{ij}}} \sim t(n-p-1)$

$c_{ij}$ 是 $C = (X^T X)^{-1}$ 的对角线上的第 $j+1 (j = 0, 1, \ldots, p)$ 个元素.

若 $|t| > t_{\alpha/2}(n-k-1)$ , 则在 $1-\alpha$ 水平下拒绝原假设, 即 $\beta_j$ 对应的解释变量 $\boldsymbol{X}_j$ 是显著的;
若 $|t| \leq t_{\alpha/2}(n-k-1)$ , 则在 $1-\alpha$ 水平下接受原假设，即 $\beta_i$ 对应的解释变量 $\boldsymbol{X}_j$ 是不显著的;

## 变量选择

**最优模型满足条件(2个)：**

1. 模型反映了变量间的真实关系
2. 模型包含的变量尽量少

**选择方法：**

1. 因素分析：因素分析是一种定性分析，预测时选择自变量的第一步（主观判断）
2. 简单相关分析：分别计算预测对象与各影响因素的简单相关系数，选择那些与预测对象相关程度高的作为自变量
3. 逐个剔除法（后退法）：首先将与预测对象有关的全部因素引入方程，建立模型，然后通过每个回归系数的t值大小，逐个剔除不显著变量，直到模型中包含的变量都是影响预测对象的显著因素为止。

**注意：**

1. 当不显著变量较多时，不能同时剔除，要从最小的哪个系数所对应的变量开始剔除
2. 删除后要观察，其他的统计量的变化，如果有所改善，剔除就是适宜的，否则应该保留在模型中
3. 逐步回归：有进有出，观察t统计量经检验是否显著，即每引入一个变量后，对已经选入的变量进行逐个检验，当原因如的变量由于后引入的变量变得不显著时，要进行剔除，直到守恒，无进无出。

## 回归诊断

- 通过残差向量 $e = y - \hat{y} = y - X\hat{\beta}$ 进行分析，$e$ 是 $\varepsilon$ 的估计，残差分析可以诊断模型的基本假定是否成立，通过残差去判断模型的拟合效果。
- 残差分析可以引导发现数据中的结构，也可能指出那些蕴含在数据中的，在只用一些概述性统计量分析时容易被疏漏的信息
- 可以通过残差分析，找出原始数据中的可疑数据即异常点
- 异常点检测时，一般将标准化残差的绝对值大于等于2的观测值认为是可以点，大于等于3的认为是异常点，标准化残差=(残差−残差均值)/残差标准差

### 回归预测

- 点预测：将自变量的预测值 $x$ 带入回归模型所得到的因变量 $y$ 的值，作为与相对应的预测值
- 区间预测：就是区间估计，即在给定的置信度下求出精确值的 $y_0$ 置信区间，成为 $y_0$ 的区间预测

公式如下：

1. **点预测**

   求出回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\ldots+\hat{\beta}_kx_k$ ,对于给定自变量的值 $x_1^*,\ldots,x_k^*$ ，用 $\hat{y}^*=\hat{\beta}_0+\hat{\beta}_1x_1^*+\ldots+\hat{\beta}_k{x_k}^*$ 来预测 $y^*=\beta_0+\beta_1x_1^*+\ldots+\beta_kx_k^*+\varepsilon$ ，称 $\hat{y}^*$ 为 $y^*$ 的点预测。

2. **区间预测**

   $y$ 的 $1-\alpha$ 的预测（置信）区间为 $(\hat{y_1},\hat{y_2})$ ，其中

$$\begin{cases}\hat{y}_1=\hat{y}-\hat{\sigma}_e\sqrt{1+\sum_{i=0}^k\sum_{j=0}^k c_{ij}x_ix_jt_{1-\alpha/2}(n-k-1)}\\\hat{y}_2=\hat{y}+\hat{\sigma}_e\sqrt{1+\sum_{i=0}^k\sum_{j=0}^k c_{ij}x_ix_jt_{1-\alpha/2}(n-k-1)}\end{cases}$$

$$C=L^{-1}=(c_{ij}),L=X^\top X$$

$$\hat{\sigma}_e=\sqrt{\frac{Q}{n-k-1}}$$

## 第三章 广义线性模型

### 线性回归的基本假设

- 因变量的观测值相互独立，且服从正态分布，即 $y_i\sim N(\mu_i,\sigma_i^2)$.

  线性回归模型中，如果使用最小二乘法估计模型参数，无需作正态分布假设。进行统计推断或应用极大似然法估计模型参数时，需要作正态性假设。
- 所有的观测值都具有相同的方差，即 $\sigma_i^2=\sigma^2$ 。
- 因变量的期望值可以直接表示为参数的线性组合 $\mu_i=x_i^T\beta$ .

  其中 $\beta=(\beta_0,\beta_1,\ldots,\beta k)$ , 表示参数向量：$x_i^T=(1,x_{1i},\ldots,x_{ki})$ ,表示第 $i$ 次观测的解释变量向量。

### 线性回归的局限性

- 因变量的正态假设不能满足实际应用要求。（实际应用中，因变量可能是二分类变量、多分类变量、计数
  型变量、大于零的右偏性变量等。）
- 因变量的方差常常随均值变化而变化，非常数方差。（如指数分布，方差是均值的平方。）
- 某些情况下自变量之间可能通过乘法关系对因变量产生影响。

## 广义线性模型的基本构成

- 随机成分:观测值 $y_i$ 是相互独立的随机变量，且服从指数分布族。(指数分布族的方差可随其均值的变化而变化)
- 系统成分:广义线性模型的线性预测值仍然为 $\eta_i = x_i^T\beta$.
- 连接函数:因变量的拟合值经过连接函数 $g$ 变换之后等于线性预测，即 $g(\mu_i) = x_i^T\beta$ .( $g$ 单调可导，$\mu_i = h(x_i^T\beta)$ ， $h$ 表示连接函数 $g$ 的逆函数)

## 广义线性模型的一般形式

上述假设可以简记为:

$$\begin{cases} y_i \sim \text{均值为}\mu_i\text{的指数分布族} \\ g(\mu_i) = x_i^T\beta \end{cases}, \ i = 1, 2, \cdots, n$$

## 指数分布族

假设随机变量 $Y$ 服从指数分布族，则其密度函数表示为:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$$

$b(\theta)$ 和 $c(y,\phi)$ 是已知函数， $b(\theta)$ 唯一确定分布的具体形式， $c(y,\phi)$ 起标准化作用。两个函数的不同形式决定了具体的分布类型。
$\theta$ 为自然参数，与分布的均值 $\mu$ 有关。

$\phi$ 为离散参数，与分布的均值无关，仅与方差有关。

**常用分布类型:**

正态分布、泊松分布、二项分布、伽马分布、逆高斯分布、Tweedie分布等。

**常见分布及其联系函数**

| 分布 | 联系函数 | |
|---|---|---|
| 正态分布 | $\eta = \mu$ | 普通线性模型 |
| 二项分布或多项分布 | $\eta = log_\mu$ | 对数线性模型 |
| ( $poisson$ 分布) | | ( $poisson$ 回归) |
| | $\eta = log\frac{P}{1-P}$ | $Logistic$ 回归模型 |
| | $\eta = log\frac{h(t)}{h_0(t)}$ | $COX$ 回归模型 |

# $Logistic$ 模型

1. **模型定义**

   设 $y_i$ 服从参数为 $p_i$ 的二项分布，则 $\mu_i = E(y_i) = p_i$

   采用逻辑连接函数，即

   $g(\mu_i) = \log\ \text{it}\ (p_i) = \log\frac{p_i}{1-p_i} = x_i^T\beta$

这个广义线性模型称为 $Logistic$ 模型

2. $Logistic$ **模型**

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \beta_m x_m$$

$g(x)$ 是对 $P$ 的变换，称为 $logit$ 变换：

$$g(x) = ln\left[\frac{P}{1-P}\right]$$

可以得到

$$P = \frac{exp[g(x)]}{1 + exp[g(x)]}$$

模型估计方法;

**最大似然法(Maximum Likelihood Method)** : 构造似然函数(Likelihood function)

$L = \Pi P(y=1|x)P(y=0|x)$，通过迭代法估计一组参数 $\beta_0, \beta_1, \beta_2 \ldots \beta_m$ 使 $L$ 达到最大。

3. **模型及自变量的统计检验**

**模型检验**

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_i = \cdots = \beta_m = 0$$
$$H_1 : \text{至少有一个} \beta \neq 0$$

采用似然比检验，当 $P \leq 0.05$ 时，拒绝 $H_0$ ，认为模型具有统计学意义。

**自变量检验**

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

采用Wald检验，当 $P \leq 0.05$ 时，拒绝 $H_0$ ，认为 $\beta_i$ 不为 $0$。

4. $Logistic$ **回归模型特点**

- Logistic function 取值 0–1，可描述/预测概率，Logistic 模型是概率模型
- Logistic function 呈S–形曲线，符合流行病学对危险因素与疾病风险关系的认识。

5. $Logistic$ **回归的自变量(影响因素)**

- 可以是连续变量，也可以是分类变量。
- 如果自变量中有分类变量，应以数字表示不同分类，如：“吸烟状况”为自变量 $X_1$ ，

  可以：

  $X_1 = 1$ 表示吸烟，

  $X_1 = 0$ 表示不吸烟。

6. **多项** $Logistic$ **模型**

前面介绍的模型因变量为二水平分类变量，当分类变量有两个以上的水平且这些水平为仅有的可能水平时，可以采用多项 $Logit$ 模型

假定对于第 $i$ 个观测，因变量$y_i$ 有 $M$ 个取值 $1, 2, \cdots, M$，自变量为 $x$，则多项 $Logit$ 回归模型为:

$$P\left(y_i=k\right)=\frac{\exp\left(x_i\beta_k\right)}{1+\sum_{j=2}^{M}\exp\left(x_i\beta_j\right)},k=2,3,\cdots,M$$

$$而P\left(y_i=1\right)=1-\sum_{j=2}^{M}P\left(y_i=j\right)=\frac{1}{1+\sum_{j=2}^{M}\exp\left(x_i\beta_j\right)}$$

# 第四章 聚类分析

> “物以类聚，人以群分”，科学研究在揭示对象特点及其相互作用的过程中，不惜花费时间和精力进行对象分类 ，以揭示其中相同和不相同的特征。

## 分类与聚类的区别

分类：用已知类别的样本训练集来设计分类器(有监督)

聚类:用事先不知样本的类别，而利用样本的先验知识来构造分类器(无监督学习)

有监督学习是让计算机去学习我们已经建立好的分类系统。

无监督学习看起来非常困难：目标是我们不告诉计算机怎么做，而是让它(计算机)自己去学习怎样做一些事情。

**聚类(Clustering):**

- 聚类是一个将数据集划分为若干组 (class)或类 (cluster)的过程，并使得同一个组内的数据对象具有较高的相似度而不同组中的数据对象是不相似的
- 相似或不相似是基于数据描述属性的取值来确定的，通常利用各数据对象间的距离来进行表示。
- 聚类分析尤其适合用来探讨样本间的相互关联关系从而对各个样本结构做一个初步的评价。
- 聚类(Clustering)就是将数据分组成为多个类(Cluster或译为簇)。
- 在同一个类内对象之间具有较高的相似度，不同类之间的对象差别较大
- 从机器学习的角度讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。

## 聚类分析

- 聚类分析是一种**建立分类**的多元统计分析方法，它能够将一批样本(或变量）数据根据其诸多特征，按照在**性质上的亲疏程度**(各变量取值上的总体差异程度）在**没有先验知识(没有事先指定的分类标准）**的情况下进行自动分类，产生多个分类结果。
- *聚类分析中大多数解为近似解，选择其中一个作为最优解

## 聚类分析的基本思想

- 假定研究对象之间存在不同程度的相似性(亲疏程度)
- 根据观测样本，找出并计算一些能够度量相似程度的统计最(相似系数、相关系数、距离等)
- 按照相似性统计量，将相似程度大的聚合到一类，关系疏远的聚合到另一类，直到把所有样本都聚合完毕，形成一个由小到大的分类系统。
- 最后将分类系统直观地用图形表示出来，即谱系图。

## 聚类分析的关键

- 亲疏关系的判别：相似性与距离（不相似性）
- 分类数量的确定：分多少类合适

## 判别方法好坏的标准

1. **内部有效性指标**：这些指标是根据聚类结果本身来评估的。例如，同一聚类内的对象应尽可能相似(内部紧密性，如SSE，WCSS等），不同聚类的对象应尽可能不同(间隔性，如BCSS等）。
2. **外部有效性指标**：如果有真实的类标签，那么可以根据真实的类标签来评估聚类结果的好坏。常用的外部指标包括调整兰德系数(Adjusted Rand Index, ARI）、互信息(Mutual Information, MI）、Fowlkes-Mallows指数等。
3. **稳定性和可解释性**：良好的聚类结果应该是稳定的，也就是说，对输入数据的小幅度改变不会导致聚类结果的大幅度改变。此外，聚类结果应具有一定的可解释性，使得人们可以理解每一簇代表的含义。
4. **运行时间和计算复杂性**：对于大规模数据集，聚类算法的运行时间和计算复杂性也是评价其好坏的重要指标。

## 常用统计量

1. 用相似系数，比较相似的样本归为一类，不怎么相似的样本归为不同的类。
2. 另一种方法是将一个样本看作P维空间的一个点，并在空间定义距离，距离越近的点归为一类，距离较远的点归为不同的类。
3. 对样本进行聚类(Q型聚类)，常用的统计量为距离
4. 对变量进行聚类(R 型聚类)，常用的统计量为相似系数

## R型聚类(Row Clustering)

R型聚类主要针对行(变量）进行聚类，也就是说，它对于数据集中的各个样本进行分类。在许多实际问题中，我们主要关注的是如何将数据集中的样本分类，所以R型聚类被广泛应用。例如，假设我们有一组人口统计数据，其中的每一行代表一个人，每一列代表一个特征(如年龄、收入等），那么R型聚类就是根据这些特征将人群进行分类。

## Q型聚类(Column Clustering)

与R型聚类相反，Q型聚类主要针对列(样本）进行聚类，也就是对于数据集中的各个特征进行分类。这种聚类方式常常用于挖掘各个特征之间的相互关系。在上述人口统计数据的例子中，Q型聚类可能会将年龄和收入这两个特征归为一类，因为这两个特征可能在某种程度上相关。

## 相似性度量

- 聚类分析研究的主要内容如何度量事物之间的相似性？
- 怎样构造聚类的具体方法以达到分类的目的？

聚类分析中，个体之间的“亲疏程度”是极为重要的，它将直接影响最终的聚类结果。对“亲疏”程度的测度一般有两个角度

- 第一，个体间的相似程度
- 第二，个体间的差异程度。衡量个体间的**相似程度通常可采用简单相关系数**等，个体间的**差异程度通常通过某种距离来测度**。

## 距离及选择性质

**常见的距离:**

- 欧氏距离：$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$
- 绝对距离：$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$
- 切比雪夫：$d(x,y)=\max_i|x_i-y_i|$
- 明氏距离：$d(x,y)=\sqrt[i]{\sum_{i=1}^{p}|x_i-y_i|^k}$
- 马氏距离：$d(x,y)=\sqrt{(x-y)^TS^{-1}(x-y)}$
- 兰氏距离：$d(x,y)=\sum_{i=1}^{p}\frac{|x_i-y_i|}{x_i+y_i}$

  其中 $x_i>0,y_i>0\ (i=1,\cdots,p)$

**距离选择的原则:**

- 要考虑所选择的距离公式在实际应用中有明确的意义。如欧氏距离就有非常明确的空间距离概念，马氏距离有消除量纲影响的作用。
- 要综合考虑对样本观测数据的预处理和将要采用的聚类分析方法。如在进行聚类分析之前已经对变量作了标准化处理、则通常就可采用欧氏距离。
- 要考虑研究对象的特点和计算量的大小。样品间距离公式的选择是一个比较复杂且带有一定主观性的问题，我们应根据研究对象的特点不同做出具体分析。实际中，聚类分析前不妨试探性地多选择几个距离公式分别进行聚类，然后对聚类分析的结果进行对比分析以确定最合适的距离测度方法。

## 系统聚类法

1. 最短距离法
2. 最长距离法
3. 中间距离法
4. 重心法
5. 类平均法
6. 离差平方和法(Ward法)

上述6种方法归类的基本步骤一致，只是类与类之间的距离有不同的定义。

**基本思想**

先将 $n$ 个样品各自看成一类，然后规定样品之间的“距离”和类与类之间的距离。选择距离最近的两类合并成一个新类，计算新类和其它类(各当前类)的距离，再将距离最近的两类合并。这样，每次合并减少一类，直至所有的样品都归成一类为止。

**基本步骤**

1. 计算 $n$ 个样品两两间的距离 $d_{ij}$ ，记作 $D=\{d_{ij}\}$ 。
2. 构造 $n$ 个类，每个类只包含一个样品。
3. 合并距离最近的两类为一新类。
4. 计算新类与各当前类的距离。
5. 重复步骤3、4，合并距离最近的两类为新类，直到所有的类并为一类为止。

6. 画聚类谱系图。

7. 决定类的个数和类。

### 1.最短距离法

定义类与类之间的距离为两类最近样品的距离，即为

$$D_{ij}=\min_{x_i\in G_i,x_j\in G_j}d_{ij}$$

设类与类合并成为一个新类，则任一类与新类的距离为

$$\begin{aligned}D_{kr}&=\min_{X_i\in G_k,X_j\in G_r}d_{ij}\\&=\min\left\{\min_{X_i\in G_k,X_j\in G_p}d_{ij},\min_{x_i\in G_k,x_j\in G_q}d_{ij}\right\}\\&=\min\left\{D_{kp},D_{kq}\right\}\end{aligned}$$

设类 $p$ 与类 $q$ 合并成为一个新类，记为 $k$ ,则 $k$ 与任一类 $r$ 的距离是

$$d_{kr}=\min\left\{d_{pr},d_{qr}\right\}$$

### 2.最长距离法

定义类 $G_i$ 与 $G_j$ 之间的距离为两类最远样品的距离，即为

$$D_{pq}=\max_{x_i\in G_p,x_j\in G_q}d_{ij}$$

将类 $G_p$ 与 $G_q$ 合并成为 $G_r$ ，则任一类 $G_i$ 与 $G_j$ 的类间距离公式为

$$\begin{aligned}D_{kr}&=\max_{X_i\in G_k,X_j\in G_r}d_{ij}\\&=\max\left\{\max_{X_i\in G_k,X_j\in G_p}d_{ij},\max_{X_i\in G_k,X_j\in G_q}d_{ij}\right\}\\&=\max\left\{D_{kp},D_{kq}\right\}\end{aligned}$$

### 3.中间距离法

最短、最长距离定义表示都是极端情况，我们定义类间距离可以既不采用两类之间最近的距离也不采用两类之间最远的距离，而是采用介于两者之间的距离，称为中间距离法。中间距离法将类 $G_p$ 与 $G_q$ 类合并为类 $G_r$ ，则任意的类 $G_k$ 和 $G_r$ 的距离公式为

$$D_{kr}^2=\frac{1}{2}D_{kq}^2+\beta D^2pq\quad\left(-\frac{1}{4}\leq\beta\leq0\right)$$

设 $D_{kq}>D_{kp}$ ,如果采用最短距离法，则 $D_{kr}=D_{kp}$ ,如果采用最长距离法，则$D_{kr}=D_{kq}$ 。

**特别当 $\beta=-\frac{1}{4}$ ,它表示取中间点算距离，公式为**

$$D_{kr}=\sqrt{\frac{1}{2}D_{kr}^2+\frac{1}{2}D_{kp}^2-\frac{1}{4}D_{kq}^2}$$

### 4.重心法

重心法定义类间距离为两类重心(各类样品的均值)的距离。重心指标对类有很好的代表性，但利用各样本的信息
不充分。

设 $G_p$ 与 $G_q$ 分别有样品 $n_p$ ， $n_q$ ，个，其重心分别为 $\overline{X_p}$ 和 $\overline{X_q}$ ，则 $G_p$ 与 $G_q$ 之间的距离定义为 $\overline{X_p}$ 和 $\overline{X_q}$ 之间的距离，这里我们用欧氏距离来表示，即

$$D_{pq}^2=\left(\overline{X_p}-\overline{X_q}\right)'\left(\overline{X_p}-\overline{X_q}\right)$$

设将 $G_p$ 与 $G_q$ 合并为 $G_r$ ，则 $G_r$ 内样品个数为 $n_r=np+nq$ ，它的重心是 $\overline{X_r}=\frac{1}{n_r}\left(n_p\overline{X_p}+n_q\overline{X_q}\right)$ ，类 $G_k$ 的重心是 $\overline{X_i}$ ，它与新类的距离为

$$D_{kr}^2=\frac{n_p}{n_r}D_{kq}^2+\frac{n_q}{n_r}D_{kq}^2-\frac{n_pn_q}{n_r^2}D_{pq}^2$$

### 5.类平均法 ★

类平均法定义类同距离平方为这两类元素之间距离平方的平均数，即为

$$D_{pq}^2=\frac{1}{n_pn_q}\sum_{X_i\in G_p}\sum_{X_j\in G_q}d_q^2$$

该聚类的某一步将 $G_p$ 与 $G_q$合并为 $G_r$ ，则任一类 $G_k$ 与 $G_r$ 的距离为

$$\begin{aligned}D_{kr}^2&=\frac{1}{n_kn_r}\sum_{X_i\in G_k}\sum_{X_j\in G_r}d_{ij}^2\\&=\frac{1}{n_kn_r}\left(\sum_{X_i\in G_k}\sum_{X_j\in G_p}d_p^2+\sum_{X_i\in G_k}\sum_{X_j\in G_q}d_q^2\right)\\&=\frac{n_p}{n_r}D_{kp}^2+\frac{n_q}{n_r}D_{kq}^2\end{aligned}$$

### 6.离差平方和法 ★

该方法是Ward提出来的，所以又称为Ward法。该方法的基本思想来自方差分析，如果分类正确，同类样品的离差平方和应当较小，类与类的离差平方和较大。具体做法是先将 $n$ 个样品各自成一类，然后每次缩小一类，每缩小一类，离差平方和就要增大，选择使得方差增加最小的两类合并，直到所有的样品归为一类为止。

设将 $n$ 个样品分成 $k$ 类 $G_1,G_2,\ldots,G_k$ 用 $X_{ij}$ 表示 $G_l$ 中的第 $l$ 个样品，$n_l$ 表示样品 $G_l$ 中样品的个数，$\overline{X_l}$ 是 $G_l$ 的重心，则 $G_l$ 的样品离差平方和为

$$S_t=\sum_{t=1}^{n_t}\left(X_{it}-\overline{X_t}\right)'\left(X_{it}-\overline{X_t}\right)$$

如果 $G_p$ 与 $G_q$ 合并为新类 $G_r$ ,类内离差平方和分别为

$$S_p=\sum_{i=1}^{n}\left(X_{ip}-\overline{X}_p\right)'\left(X_{ip}-\overline{X}_p\right)$$

$$S_q=\sum_{i=1}^{n}\left(X_{iq}-\overline{X}_q\right)'\left(X_{iq}-\overline{X}_q\right)$$

$$S_r=\sum_{i=1}^{n}\left(X_{ir}-\overline{X}_r\right)'\left(X_{ir}-\overline{X}_r\right)$$

它们反映了各自类内样品的分散程度，如果 $G_p$ 与 $G_q$ 这两类相距较近，则合并后所增加的离散平方和 $S_r-S_p-Sq$ 应较小；否则，应较大。于是定义 $G_p$ 与 $G_q$ 之间的平方距离为：

$$D_{pq}^2=S_r-S_p-S_q$$

其中 $G_r=G_p\cup G_q$ ，可以证明类间距离的递推公式为

$$D_{kr}^2=\frac{n_k+n_p}{n_r+n_k}D_{kp}^2+\frac{n_k+n_q}{n_r+n_k}D_{kq}^2-\frac{n_k}{n_r+n_k}D_{pq}^2$$

**对异常值很敏感，对较大的类倾向产生较大的距离，从而不易合并，较符合实际需要**

## K-means聚类

**基本思想**

1. 首先输入 $k$ 的值，即我们指定希望通过聚类得到 $k$ 个分组；
2. 从数据集中随机选取 $k$ 个数据点作为初始大佬（质心）；
3. 对集合中每一个小弟，计算与每一个大佬的距离，离哪个大佬距离近，就跟定哪个大佬；
4. 这时每一个大佬手下都聚集了一票小弟，这时候召开选举大会，每一群选出新的大佬（即通过算法选出新的质心）；
5. 如果新大佬和老大佬之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），可以认为我们进行的聚类已经达到期望的结果，算法终止；
6. 如果新大佬和老大佬距离变化很大，需要迭代3~5步骤。

**决定性因素**

1. 数据的采集和抽象（个人认为是抽样）
2. 初始的中心选择
3. 最大迭代次数
4. 收敛值
5. $k$ 值的选定
6. 度量距离的手段

**主要因素**

1. 初始中心点
2. 输入的数据及 $k$ 值的选择
3. 距离度量

**特点**

- 事先确定分类数
- 计算过程无须存储激据，因此能处理更大的数据量，也称快速聚类
- 样品的最终聚类在某种程度上依赖于最初的划分或种子点

**计算机实现**

1. 从 $D$ 中随机取k个元素，作为 $k$ 个簇的各自的中心。
2. 分别计算剩下的元素到 $k$ 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。
3. 根据聚类结果，重新计算 $k$ 个簇各自的中心，计算方法是取族中所有元素各自维度的算术平均数
4. 将 $D$ 中全部元素按照新的中心重新聚类。
5. 重复第 $4$ 步，直到聚类结果不再变化。
6. 将结果输出

# 第五章 判别分析

## 距离判别法

- 基本思想：样本与哪一类总体的距离最近，就判别它属于哪一类总体。
- 判别准则：对于任给一次观测值，若它与第i类的重心距离最近，就认为它来自于第i类。
- 马氏距离：不受单位影响，是一个无单位的数值。

$$d^2(X,Y)=(X-Y)'\Sigma^{-1}(X-Y)$$
$$d^2(X,G)=(X-\mu)'\Sigma^{-1}(X-\mu)$$

设 $G$ 是 $p$ 维总体，数学期望(均值向量）为 $\mu$ ，协方差矩阵为$\Sigma$。

采用马氏距离的原因：马氏距离综合考虑了样品数据之间的依存关系，从绝对和相对两个角度考察样品，消除了变量单位不一致的影响，更具合理性。

两总体 $G_1$ 和 $G_2$ 的均值分别为$\mu_1$和$\mu_2$ ，有相同的协方差矩阵 $\Sigma$ ，按欧氏距离： $x$ 应判入总体 $G_2$ ;按马氏距离： $x$ 应判入总体 $G_1$ (更合理！）

**一.两个总体的距离判别**

设 $p$ 维总体 $G_1$ 和 $G_2$ 的数学期望(均值向量）分别为 $\mu_1$ 和 $\mu_2$，协方差矩阵分别为 $\sum_1$ 和 $\sum_2$。$x$ 是一个新样品( $p$ 维）。先要判断 $x$ 来自哪个总体。距离判别是分别计算$x$到$G_1$和$G_2$的马氏距离 $d(x，G_1)$和 $d(x，G_2)$，并按如下的判别准则进行判别：

$$\begin{cases} x\epsilon G_1，d(x，G_1)\le d(x，G_2) \\ x\epsilon G_2，d(x，G_1)>d(x，G_2) \end{cases}$$

(当等号成立时，可将判给两总体中的任一个。对正态总体来说，等号成立的概率为零)

下面对总体协方差矩阵相等和不相等两种情况，分别讨论判别准则。

1. $\sum_1 = \sum_2 = \sum$

这时，平方马氏距离之差

$$
\begin{aligned}
& d^2\left(x, G_2\right)-d^2\left(x, G_1\right) \\
=&\left(x-\mu_2\right)^T \Sigma^{-1}\left(x-\mu_2\right)-\left(x-\mu_1\right)^T \Sigma^{-1}\left(x-\mu_1\right) \\
=&x^T \Sigma^{-1} x-2 x^T \Sigma^{-1} \mu_2+\mu_2^T \Sigma^{-1} \mu_2-x^T \Sigma^{-1} x+2 x^T \Sigma^{-1} \mu_1-\mu_1^T \Sigma^{-1} \mu_1 \\
=&2 x^T \Sigma^{-1}\left(\mu_1-\mu_2\right)+\mu_2^T \Sigma^{-1} \mu_2-\mu_1^T \Sigma^{-1} \mu_1+\mu_1^T \Sigma^{-1} \mu_2-\mu_2^T \Sigma^{-1} \mu_1 \\
=&2 x^T \Sigma^{-1}\left(\mu_1-\mu_2\right)-\left(\mu_1+\mu_2\right)^T \Sigma^{-1}\left(\mu_1-\mu_2\right) \\
=&2\left[x-\frac{1}{2}\left(\mu_1+\mu_2\right)\right]^T \Sigma^{-1}\left(\mu_1-\mu_2\right) \\
=&2(x-\bar{\mu})^T \Sigma^{-1}\left(\mu_1-\mu_2\right)
\end{aligned}
$$

其中 $\bar{\mu}=\frac{1}{2}\left(\mu_1+\mu_2\right)$。

令

$$W(x)=(x-\bar{\mu})^T \Sigma^{-1}\left(\mu_1-\mu_2\right)$$

称 $W(x)$ 为判别函数。则判别准则可简化为

$$
\begin{cases}x \epsilon G_1, W(x) \geq 0 \\ x \epsilon G_2, W(x)<0\end{cases}
$$

令

$$a^T=\left(\mu_1-\mu_2\right)^T \Sigma^{-1}$$

则判别函数可写成

$$W(x)=a^T(x-\bar{\mu})$$

即当$\mu_1, \mu_2$和 $\Sigma$ 已知时，判别函数 $W(x)$ 是$x$的线性函数。线性判别函数使用起来最方便，在实际应用中也最广泛。

$\mu_1, \mu_2$ 和 $\Sigma$ 的估计

在实际问题中，$\mu_1$，$\mu_2$和 $\Sigma$ 通常是未知的，这时可通过训练样本对$\mu_1$，$\mu_2$和 $\Sigma$ 作估计。设

$$x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$$

是来自 $G_1$ 的样本，样本容量为 $n_1$ 。而

$$x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$$

是来自 $G_2$ 的样本，样本容量为 $n_2$ 。$x_i^{(k)}$为 $p$ 维向量。

记

$$\hat{\mu}=\bar{x}^{(k)}=\frac{1}{n_k} \sum_{i=1}^{n_k} x_i^{(k)}, k=1,2$$

则样本均值向量 $\hat{\mu_1}$ 和 $\hat{\mu_2}$分别可作为 $\mu_1$ 和 $\mu_2$ 的估计。而样本协方差矩阵为

$\Sigma$ 的估计

$$
\begin{aligned}
S_k &= \frac{1}{n_k-1}\sum_{i=1}^{n_k}\left(x_i^{(k)}-\bar{x}^{(k)}\right)\left(x_i^{(k)}-\bar{x}^{(k)}\right)^T \\
&= \frac{1}{n_k-1}\begin{vmatrix} L_{11}^{(k)} & L_{12}^{(k)} & \cdots & L_{1p}^{(k)} \\ L_{21}^{(k)} & L_{22}^{(k)} & \cdots & L_{2p}^{(k)} \\ \vdots & \vdots & \cdots & \vdots \\ L_{p1}^{(k)} & L_{p2}^{(k)} & \cdots & L_{pp}^{(k)} \end{vmatrix}, k=1,2
\end{aligned}
$$

其中

$$
L_{ts}^{(k)}=\sum_{i=1}^{n_i}(x_{ti}^{(k)}-x_t^{-(k)})(x_{si}^{(k)}-x_s^{-(k)}), x_t^{-(k)}=\frac{1}{n_k}\sum_{i=1}^{n_k}x_{ti}^{(k)}, t,s=1,2,\ldots,p
$$

则 $\Sigma$ 的联合无偏估计为

$$
\hat{\sum}=\frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}
$$

判别函数的估计

于是，判别函数的估计为

$$
\hat{W}(x)=(x-\hat{\bar{\mu}})^T\hat{\sum}^{-1}(\hat{\mu_1}-\hat{\mu_2})
$$

其中 $\hat{\bar{\mu}}=\frac{1}{2}(\hat{\mu_1}+\hat{\mu_2})$。而判别准则为

$$
\begin{cases} x\in G_1, \hat{W(x)}\geq 0 \\ x\in G_2, \hat{W(x)}<0 \end{cases}
$$

2. $\sum_1 \neq \sum_2$

选择判别函数为

$$
\begin{aligned}
W(x) &= d^2(x,G_2)-d^2(x,G_1) \\
&= (x-\mu_2)^T(x-\mu_2)-(x-\mu_1)^T(x-\mu_1)
\end{aligned}
$$

则

$$
\begin{cases} x\in G_1, W(x)\geq 0 \\ x\in G_2, W(x)<0 \end{cases}
$$

这时，判别函数$W(x)$ 为 $x$ 的二次函数。

判别函数的估计

当 $\mu_1$ ， $\mu_2$， $\sum_1$ 和 $\sum_2$未知时，可由训练样本的 $\hat{\mu_1},\hat{\mu_2}$,$S_1$ 和 $S_2$ 作出估计。从而得到判别函数 $W(x)$ 的估计为

$$
\hat{W(x)}=(x-\hat{\mu_2})^TS_2^{-1}(x-\hat{\mu_2})-(x-\hat{\mu_1})^TS_1^{-1}(x-\hat{\mu_1})
$$

**错判概率**

由上面的分析可以看出，马氏距离判别法是合理的，但是这并不意味着不会发生误判。

利用判别函数进行判断，一般总会出现错判。一个判别准则的优劣，通常可以用它的误判概率来衡量。

例如，对两总体 $G_1$ 和 $G_2$ ，误判概率就是当 $x$ 属 $G_1$ 但误判为 $G_2$ ，或 $x$ 属 $G_2$ 却误判给 $G_1$ 的概率。只有当总体分布完全已知时，才能精确地计算误判概率。在实际应用中，总体分布通常未知，这时可用训练样本来评价判别准则的优劣。

★贝叶斯判别法

■贝叶斯统计的思想是：

□假定对研究对象已有一定的认识，常用先验概率分布来描述这种认识。

□取得一个样本，用样本来修正已有的认识(先验概率分布），得到后验概率分布。

□各种统计推断都通过了后验概率分布来进行。

■将贝叶斯思想用于判别分析，就得到了贝叶斯判别。

**贝叶斯公式是一个我们熟知的公式**

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum P(A|B_i)P(B_i)}$$

Bayes判别准则

以个体归属于某类的概率(或判别值）最大或错判总平均损失最小为标准。

举例：

> 办公室新来了一个雇员小王，大家都在猜测小王是好人还是坏人。按人们的主观意识，一个人是好人或坏人的概率均为0.5.坏人总是要做坏事，好人总是做好事，偶尔也会做一件坏事，一般好人做好事的概率为0.9，坏人做好事的概率为0.2。一天，小王做了一件好事，问小王是好人的概率有多大，你现在把小王判为何种人。

$$\begin{aligned}&P(\text{好人}|\text{做好事})\\&=\frac{P(\text{好人})P(\text{做好事}|\text{好人})}{P(\text{好人})P(\text{做好事}|\text{好人})+P(\text{坏人})P(\text{做好事}|\text{坏人})}\\&=\frac{0.5*0.9}{0.5*0.9+0.5*0.2}\\&=0.82\end{aligned}$$

$$\begin{aligned}&P(\text{坏人}|\text{做好事})\\&=\frac{P(\text{坏人})P(\text{做好事}|\text{坏人})}{P(\text{好人})P(\text{做好事}|\text{好人})+P(\text{坏人})P(\text{做好事}|\text{坏人})}\\&=\frac{0.5*0.2}{0.5*0.9+0.5*0.2}\\&=0.18\end{aligned}$$

一、概率判别

> $k$ 个总体的先验概率 $q_1,q_2,\ldots q_k$

密度函数分别为$p_1(x),p_2(x),\ldots p_k(x)$

$x$ 来自第 $j$ 类的后验概率为(Bayes公式)

$$p(j/x)=\frac{q_jp_j(x)}{\sum_{i=1}^{k}q_ip_i(x)},j=1,2,\dots k$$

当$p(j/x)=\underset{1\le j\le k}{max}p(j/x)$时，判x来自第j总体。

## 二、损失判别

$x$错判为第$g$总体的平均损失

$$E(g/x)=\sum_{j\neq i}\frac{q_jp_j(x)}{\sum_{i=1}^{k}q_ip_i(x)}L(g/j)$$

如$E(g/x)=\underset{1\le j\le k}{min}E(j/x)$时，判$x$来自第$g$总体。

$$L(g/j)=\begin{cases}0 & \mathrm{g}=j\\1 & \mathrm{g}\neq j\end{cases}$$

$$P(g/x)\overset{g}{\to}max\Leftrightarrow E(g/x)\overset{g}{\to}min$$

**两总体的贝叶斯判别**

$G_1$ 和 $G_2$ 代表两个总体，各自的先验概率为$p_1$和$p_2(p_1+p_2=1)$

$f_1(y)$ 和 $f_2(y)$ 分别是总体 $G_1$ 和 $G_2$ 中y的概率密度函数

$R_1$和 $R_2$ 代表按分类规则划分的两组区域。例如，如果一个新观测对象分到 $R_k$ ,那么我们声明该样本来自总体$G_k$,$k=1,2$。$R_1$和$R_2$是整个空间的分割

我们可以推导总错分率(TPM)：

$$P(\text{观测对象被错分到}G_1)=P(y\in R_1|G_2)P(G_2)=P(1|2)p_2$$

$$P(\text{观测对象被错分到}G_2)=P(y\in R_2|G_1)P(G_1)=P(2|1)p_1$$

贝叶斯分类法则目标是最小化错分的期望代价(Expected cost of misclassification,ECM)：

$$ECM=c(2|1)P(2|1)p_1+c(1|2)P(1|2)p_2$$

定理(贝叶斯分类法则)：

基于最小化ECM的贝叶斯分类法则为：

$$R_1:\frac{f_1(y)}{f_2(y)}\ge(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$$

$$R_2:\frac{f_2(y)}{f_1(y)}\ge(\frac{c(2|1)}{c(1|2)})(\frac{p_1}{p_2})$$

(a) $p_2/p_1=1$(先验概率相同)

$$R_1:\frac{f_1(y)}{f_2(y)}\ge\frac{c(1|2)}{c(2|1)}$$

$$R_2:\frac{f_1(y)}{f_2(y)}<\frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2)/c(2|1)=1$(错分成本相同)

$$R_1:\frac{f_1(y)}{f_2(y)}\geq\frac{p_2}{p_1}$$

$$R_2:\frac{f_1(y)}{f_2(y)}<\frac{p_2}{p_1}$$

(c) $p_2/p_1=c(1|2)/c(2|1)=1$ 或 $p_2/p_1=1/c(1|2)/c(2|1)$

$$R_1:\frac{f_1(y)}{f_2(y)}\geq 1$$

$$R_2:\frac{f_1(y)}{f_2(y)}<1$$

·多总体的贝叶斯判别

设有 $k$ 个总体 $G_1,G_2,\dots G_k$ 的概率密度为 $f_j(x)$ 各总体出现的先验概率为

$$P_j=P(G_j),j=1,2,\dots,k,\text{满足}\sum_{j=1}^{k}P_j=1$$

样品 $x$ 属于总体 $G_i$ 的后验概率

$$P(G_i|x)=\frac{p_if_i(x)}{\sum_{j=1}^{k}p_jf_j(x)}$$

Bayes判别准则：若

$$P(G_i|x)=\underset{1\leq j\leq k}{max}\{P(G_j|x)\}(i=1,2,\dots,k)$$

则判样本 $x\in G_i$

注：当达到最大后验概率的 $G_i$ 不止一个时，可判为达到最大后验概率的总体的任何一个。

# 第六章 主成分分析

## 主成分分析

**基本思想：“降维”和“简化数据结构**

主成分分析就是把原有的多个指标转化成少数几个代表性较好的综合指标，这少数几个指标能够反映原来指标**大部分**的信息(**85%以上**),并且各个指标之间保持独立，避免出现重叠信息。

**要讨论的问题**

1. 基于**相关系数矩阵**还是基于**协方差矩阵**做主成分分析。

   选相关系数矩阵：当分析中所选择的经济变量具有不同的量纲，变量水平差异很大

2. 选择**几个主成分**?

   主成分分析的目的是简化变量。

3. 如何解释主成分所包含的实际意义?

## 总体主成分

### 主成分的含义

样本点之间的差异是由 $x_1$ 和 $x_2$ 的变化引起的，两者变动的相差不大但如果用新坐标 $y_1$ 和 $y_2$ 来代替，易见，这些样本点的差异主要体现在 $y_1$ 轴上，$n$ 个点在 $y_1$ 轴方向上的方差达到最大，即在此方向上包含了有关 $n$ 个样品的最多的信息.

将这些点投影到 $y_1$ 轴方向能使信息的损失最小，如果 $y_1$ 轴方向的差异占了全部样本点差异的绝大部分，那么将 $y_2$ 忽略是合理的，这样就把两个变量简化为一个，显然这里的 $y_1$ 轴代表了数据变化最大的方向，称之为第一主成分，$y_2$ 称为第二主成分，并要求已经包含在 $y_1$ 中的信息不出现在 $y_2$ 中

### 旋转变换

注意两个主成分 $y_1$ 和 $y_2$ 都是 $x_1$ 和 $x_2$ 的线性组合:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{P}^{\mathrm{T}}\boldsymbol{X}$$

其中 $P$ 为旋转变换矩阵，它是正交矩阵.

旋转变换的目的是使 $n$ 各样品点在y轴方向上的离散度最大，即 $y_1$ 的方差最大。变量 $y_1$ 代表了原始数据的绝大部分信息。在研究某经济问题时，即使不考虑以也无损大局。经过上述旋转变换，原始数据的大部分信息集中到轴上，对数据中包含的信息起到了浓缩作用。

### 主成分的性质

1. $Var(y) = \Lambda = diag\left(\lambda_1, \lambda_2, \cdots, \lambda_p\right)$
2. $\sum_{i=1}^{p} \operatorname{Var}\left(y_i\right) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \sigma_{it} = \sum_{i=1}^{p} \operatorname{Var}\left(x_i\right).$
3. $\rho\left(y_k, x_t\right) = \frac{\operatorname{Cov}(y_k, x_i)}{\sqrt{\operatorname{Iar}(y_k)\operatorname{Iar}(x_i)}} == \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{tt}}} e_k.$

### 主成分个数的确定

- 第 $i$ 个主成分 $y_i$ 的方差贡献率为:

$$\omega_t = \lambda_i / \sum_{j=1}^{p} \lambda_j, i = 1, \cdots, p$$

- 通常取使得累计贡献率满足

$$\sum_{i=1}^{k} \omega_i = \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{p} \lambda_j \gg 80\%$$

  的最小的 $k$ 为主成分个数。

- 通常取累计贡献率首次超过85%的 $k$ （80%亦可）

### 变量标准化及意义

从总体协方差矩阵出发做主成分分析倾向于反映方差大的量的信息会出现“大数吃小数”的现象。为了均等地对待一个原始变量，常常将各原始变量作标准化处理:

$$x_i^* = \frac{x_i - E\left(x_i\right)}{\sqrt{\operatorname{Var}\left(x_i\right)}}, \quad i = 1, 2, \cdots, p.$$

标准化后的总体 $\boldsymbol{X}^* = \left(x_1^*, \cdots, x_p^*\right)^{\mathrm{T}}$ 的协方差矩阵恰好是原总体 $X$ 的相关系数矩阵 $\rho$ .

综上讨论，既可从 $\Sigma$ 出发，也可以从 $\rho$ 出发做主成分分析，考虑到现实经济意义，后者用得更多.

### 样本主成分

协方差矩阵与相关性矩阵，样本主成分特征值与方差贡献率的计算

实际问题中 $\Sigma$ 和 $\rho$ 往往是未知的，需要用样本的协方差矩阵 $S$ 和样本的相关系数矩阵 $R$ 来估计：

$$\boldsymbol{S}=\frac{1}{n-1}\sum_{k=1}^{n}\left(\boldsymbol{X}_{(k)}-\overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_{(k)}-\overline{\boldsymbol{X}}\right)^{\mathrm{T}}=\left(s_{ij}\right)_{p\times p}.$$

$$\boldsymbol{R}=\frac{1}{n-1}\sum_{k=1}^{n}\boldsymbol{X}_{(k)}^{*}{\boldsymbol{X}_{(k)}^{*}}^{T}=\left(r_{ij}\right)_{p\times p}.$$

$$\boldsymbol{X}_{(k)}^{*}=\left[\frac{x_{k1}-\bar{x}_1}{\sqrt{s_{11}}},\frac{x_{k2}-\bar{x}_2}{\sqrt{s_{22}}},\cdots,\frac{x_{pp}-\bar{x}_p}{\sqrt{s_{pp}}}\right]^{\mathrm{T}},r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}s_{ij}}},i,j=1,\cdots,p.$$

### 主成分分析的步骤

1. 将原始样本标准化 $x^{*}=\left(x_1^{*},\cdots,x_p^{*}\right)^{\mathrm{T}},\quad x_i^{*}=\frac{x_i-\bar{x}_i}{\sqrt{s_{ii}}}(i=1,\cdots,p)$ .
2. 求样本的相关系数矩阵 $R$ .
3. 求 $R$ 的特征值 $\lambda_1^{*},\lambda_2^{*},\cdots,\lambda_p^{*}\left(\lambda_1^{*}\geqslant\lambda_2^{*}\geqslant\cdots\geqslant\lambda_p^{*}\geqslant 0\right)$ 以及对应的单位正交特征向量 $a_1^{*},a_2^{*},\cdots,a_p^{*}$ .
4. 按主成分累积贡献率超过85%确定主成分的个数 $k$ ，并写出主成分表达式为 $z_i^{*}={\boldsymbol{a}_i^{*}}^{\mathrm{T}}\boldsymbol{x}^{*},\quad i=1,\cdots,k$ .
5. 对分析结果做统计意义和实际意义两方面的解释.

## 第七章 因子分析

### 因子分析

基本思想：利用降维思想，将每个研究变量分解为几个影响因素变量，将每个原始变量分解为两部分因素，一部分是由所有变量共同具有的少数几个公共因子组成的，另一部分是每个变量独自具有的因素为特殊因子

**目的**：

1.简化变量维数，使结构简单化，希望以最少的共同因素对总变量做最大解释，因为，抽取的因子越少越好，累计解释的**变异量(方差)**越大越好。

2.在因子分析的公共因子中，应最先抽取特征值最大的公共因子，按特征值从大至小抽取。

**区别与联系**：

**区别**：

- 主成分分析通过线性组合将原变量综合成几个主成分
- 因子分析通过构筑若干意义较为明确的公因子
- 主成分分析是“变异数”导向的方法(主成分是可测的)
- 因子分析是“共变异数”导向的方法(因子是不可测的)

**联系**：

因子分析是主成分分析的推广。

## 正交因子模型

因子模型，载荷矩阵，共同度与特殊度，公因子

因子模型：用少数不可测的公共因子的线性函数来描述原观测的每一分量

又分Q型与R型因子分析

**样本**间的因子分析是Q型因子分析

**变量**间的因子分析是R型因子分析

**因子模型：**

$$
\begin{aligned}
x_1 &= \mu_1 + a_{11}f_1 + a_{12}f_1 + \cdots + a_{1m}f_m + \varepsilon_1 \\
x_2 &= \mu_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\
&\vdots \\
x_p &= \mu_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p
\end{aligned}
$$

$$
\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}, \boldsymbol{F} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}
$$

**正交因子模型：可写为**

$$X = \mu + AF + \varepsilon$$

载荷矩阵：上图中 $A$ 为载荷矩阵

若 $Cov(F)$ 不是单位阵则，该模型为斜交因子模型，否则为正交因子模型

可由正交因子模型求得 $X$ 的协方差

$$
\begin{aligned}
\Sigma &= \operatorname{Cov}(X) \\
&= E(X-\mu)(X-\mu)^T \\
&= E(AF+\varepsilon)(AF+\varepsilon)^{\mathrm{T}} \\
&= E\left(AFF^{\mathrm{T}}A^{\mathrm{T}} + \varepsilon F^{\mathrm{T}}A^{\mathrm{T}} + AF\varepsilon^{\mathrm{T}} + \varepsilon\varepsilon^{\mathrm{T}}\right) \\
&= AE\left(FF^{\mathrm{T}}\right)A^{\mathrm{T}} + E\left(\varepsilon F^{\mathrm{T}}\right)A^{\mathrm{T}} + AE\left(F\varepsilon^{\mathrm{T}}\right) + E\left(\varepsilon\varepsilon^{\mathrm{T}}\right) \\
&= AA^{\mathrm{T}} + \Phi
\end{aligned}
$$

特殊度与共同度：

$$\mathrm{Cov(X,F)} = \mathrm{E}(\mathrm{X}-\mu)F^T = \mathrm{E}(\mathrm{AF}+\varepsilon)F^T = \mathrm{A}$$

特殊度：载荷矩阵 $A$ 的第 $j$ 列的载荷平方

求解公式

$$g_i^2 = \operatorname{Var}(x_i) = \sigma_{ii} = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \phi_i \quad i = 1, \ldots, p$$

共同度：载荷矩阵 $A$ 的第 $i$ 行的载荷平方和

求解公式:

$$h_i^2 = a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2, \quad h_i^2 = \sum_{j=1}^m a_{ij}^2 = \sum_{j=1}^m a_{ij}^{*2}$$

其表示$m$个公因子对变量$x_i$的方差贡献和，称之为第 $i$ 个共同度

正交变换不会改变公因子的共同度

公因子：**代表每组数据基本结构的新变量称为公共因子**

## 因子模型估计

**主成分估计：**

其方差–协方差结构形式为：

$$\Sigma = \mathrm{AA}^{\mathrm{T}} + 0 = \mathrm{AA}^{\mathrm{T}}$$
$$\Sigma = \lambda_1 \mathrm{e}_1 \mathrm{e}_1^{\mathrm{T}} + \lambda_2 \mathrm{e}_2 \mathrm{e}_2^{\mathrm{T}} + \ldots + \lambda_{\mathrm{p}} \mathrm{e}_{\mathrm{p}} \mathrm{e}_{\mathrm{p}}^{\mathrm{T}}$$

因子分析的目的是寻找少数几个公因子解释原来p个变量的协方差结构

若最后p–m个特征值很小，则忽略$\lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \ldots + \lambda_p e_p e_p^T$对$\Sigma$的贡献

于是得：

$$\Sigma \approx \left(\sqrt{\lambda_1} e_1, \sqrt{\lambda_2} \boldsymbol{e}_2, \cdots, \sqrt{\lambda_m} \boldsymbol{e}_m\right) \begin{pmatrix} \sqrt{\lambda_1} \boldsymbol{e}_1^T \\ \sqrt{\lambda_2} e_2^T \\ \vdots \\ \sqrt{\lambda_m} e_m^T \end{pmatrix} = \boldsymbol{A} \boldsymbol{A}^T$$

在主成分估计中，对公因子个数确定也是看贡献率

**最大似然方法**

最大似然方法要求样本 $y_1, y_2 \ldots y_n$ 独立并服从正态分布 $N_p(\mu, \Sigma)$

其中 $\Sigma = LL' + \Psi$

通过最大化似然函数 $L(\Sigma) \Leftrightarrow |\Sigma|^{-\frac{n}{2}} e^{-\frac{1}{2} tr(\Sigma^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})')}$ 可以得到 $\hat{L}$ 的最大似然估计。

不过这个方法要求 $y$ 服从正态分布，而现实中这个要求不一定能满足，因此应用也受限。

## 因子正交旋转

**旋转目的：**

1. 寻找每个公因子的实际意义
2. 如果各主因子的典型代表变量不突出，就需要进行旋转
3. 使因子载荷矩阵中载荷的绝对值向0，1两个方向变化，即主因子越大越好，而特殊因子越接近0越好

**正交旋转：**

也叫最大方差正交旋转法，有因子分析模型为 $X = AF + \varepsilon$ ,设 $\Gamma = (\gamma_{ij})$ 为一正交矩阵，

作正交变换 $B = A\Gamma$ ，且 $h_i^2(B) = h_i^2(A), \quad g_j^2(B) = \sum_{k=1}^{p} \gamma_{kj} g_k^2(A)$

如何进行旋转

$$A=\begin{bmatrix}a_{11} & a_{12}\\ \cdots & \cdots\\ a_{p1} & a_{p2}\end{bmatrix},\Gamma=\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix},$$

$$B=A\Gamma=\begin{bmatrix}a_{11}\cos\theta+a_{12}\sin\theta & -a_{11}\sin\theta+a_{12}\cos\theta\\ \cdots & \cdots\\ a_{p1}\cos\theta+a_{p2}\sin\theta & -a_{p1}\sin\theta+a_{p2}\cos\theta\end{bmatrix}=\begin{bmatrix}b_{11} & b_{12}\\ \cdots & \cdots\\ b_{p1} & b_{p2}\end{bmatrix}$$

所谓**最大方差旋转法**就是选择正交矩阵 $\Gamma$ ，使得 $B$ 所有列元素平方的相对方差之和达到最大。

## 因子得分

**加权最小二乘法:**

1. 将因子模型$X=\mu+AF+\varepsilon$

   改写为 $X-\mu=AF+\varepsilon$
2. 两边同时成 $\phi^{(-\frac{1}{2})}$
3. 可写作 $X^*=A^*F+\varepsilon^*$
4. 但 $A,\phi,\mu$ 无法得到准确数值，要通过估计值进行代替，$\hat{\phi}=diag(1-\hat{h_i^2})$ 和样本均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 采用正交旋转后的载荷矩阵 $A$ 的估计 $\hat{A}$ 进行代替
5. 最后得到对应 $x_j$ 的因子得分(也叫巴特莱特因子得分)

$$\hat{f}_j=\left(\hat{A}^{\mathrm{T}}\hat{\boldsymbol{\Phi}}^{-1}\hat{\boldsymbol{A}}\right)^{-1}\hat{\boldsymbol{A}}^{\mathrm{T}}\hat{\boldsymbol{\Phi}}^{-1}\left(\boldsymbol{x}_j-\overline{\boldsymbol{X}}\right)$$

**回归法:**

在正交因子模型中，假设$\begin{pmatrix}F\\X\end{pmatrix}$服从 $(m+p)$ 元正态分布，用回归预测方法可将 $F=(f_1,f_2,\cdots,f_m)^T$ 估计为 $F=A^T\Sigma^{-1}(X-\mu)$

在实际应用中，可用 $\overline{X},\overline{A}$ 和 $S$ 分别代表上式中的 $\mu,A$ 和 $\Sigma$ 来得到因子得分。样品 $x_j$ 的因子得分(称为汤姆森 $(Thompson)$ 因子得分)

$$\hat{f}_j=\hat{A}^{\mathrm{T}}S^{-1}\left(x_j-\bar{x}\right),\quad j=1,2,\cdots,n$$

## 因子分析基本步骤

1. 确认数据是否适合做因子分析

   **一般使用KMO和Bartlett’s进行验证**

$$KMO=\frac{\sum\sum_{i\neq j}r_{ij}^2}{\sum\sum_{i\neq j}r_{ij}^2+\sum\sum_{i=j}p_{ij}^2}$$

2. 构造因子变量
3. 旋转因子使其更具可解释性
4. 计算因子得分并做因子图