第四章 相关分析

4.0 例子：家庭特征与家庭消费之间的关系

为了解家庭特征与消费模式之间的关系, 调查70个家庭的情况. 收集了下面两组变量:

消费模型变量

$$\begin{cases} x_1: \text{每年去餐馆就餐的频率} \\ x_2: \text{每年外出看电影的频率} \end{cases}$$

家庭特征变量

$$\begin{cases} y_1: \text{户主的年龄} \\ y_2: \text{家庭的年收入} \\ y_3: \text{户主的受教育程度} \end{cases}$$

目的:

1) 分析两组变量之间的关系;

2) 找出最能代表各组变量的特征量.

统计分析方法：典型相关分析

典型相关分析是研究两组变量之间相关性的一种统计分析方法，也是一种降维技术.

4.1 复相关系数

两个单变量之间的相关关系："一对一"

(简单)相关系数, 偏相关系数.

一个单变量与一个向量之间的相关关系："一对多"

复相关系数.

4.1.1 总体复相关系数

设 $Y \overset{d}{\sim} N_p(\mu, \Sigma)$, 其中 $\Sigma > 0$.

将 Y, μ 和 Σ 分别剖分为

$$Y = \begin{pmatrix} y_1 \\ Y_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

其中， $y_1, \mu_1 \in R^1$; $\sigma_{11} > 0$;
Y_2, μ_2, $\Sigma_{21} = \Sigma'_{12} \in R^{p-1}$;
Σ_{22} 是 $(p-1)$ 阶正定阵.

考虑 y_1 与 $a'Y_2$ 之间的简单相关系数, 其中 $a \in R^{p-1}$,

$$\begin{aligned}\rho_{y_1,a'Y_2} &= \frac{Cov(y_1, a'Y_2)}{\sqrt{Var(y_1)}\sqrt{Var(a'Y_2)}} = \frac{Cov(y_1, Y_2)a}{\sqrt{\sigma_{11}}\sqrt{a'Var(Y_2)a}} \\ &= \frac{\Sigma_{12}a}{\sqrt{\sigma_{11}}\sqrt{a'\Sigma_{22}a}}.\end{aligned}$$

则定义 y_1 与 Y_2 的复相关系数为

$$\rho_{y_1,Y_2} = \sup_{a\in R^{p-1}} \rho_{y_1,a'Y_2} = \frac{1}{\sqrt{\sigma_{11}}} \sup_{a\in R^{p-1}} \frac{\Sigma_{12}a}{\sqrt{a'\Sigma_{22}a}}.$$

由 ρ_{y_1,Y_2} 的非负性知

$$\rho_{y_1,Y_2} = \frac{1}{\sqrt{\sigma_{11}}}\sqrt{\sup_{a\in R^{p-1}} \frac{(\Sigma_{12}a)^2}{a'\Sigma_{22}a}} = \sqrt{\frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}},$$

并由 Cauchy-Schwarz不等式知上式在 $a = \Sigma_{22}^{-1}\Sigma_{21}$ 达最大.

定义1： 变量 y_1 与向量 Y_2 之间的**复相关系数**为

$$\rho_{y_1,Y_2} = \sqrt{\frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}},$$

其中, $\sigma_{11} = Var(y_1)$, $\Sigma_{22} = Cov(Y_2)$, $\Sigma_{12} = Cov(y_1, Y_2)$.

复相关系数 ρ_{y_1,Y_2} 的性质：

1) $0 \le \rho_{y_1,Y_2} \le 1$;
2) ρ_{y_1,Y_2} 越大则 y_1 与 Y_2 的相关性越强;
3) $\rho_{y_1,Y_2} = 0 \Longleftrightarrow \Sigma_{12} = 0$, 即 y_1 与 Y_2 独立.

定理1： 当 $a=\Sigma_{22}^{-1}\Sigma_{21}$ 时, y_1 与 $a'Y_2$ 的相关系数最大, **为复相关系数** ρ_{y_1,Y_2}, 且 $y_1-a'Y_2$ 的方差最小, 为 $\sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}=Var(y_1|Y_2)$.

证明： 对任意 $b\in R^{p-1}$, 有

$$\begin{aligned} Var(y_1-b'Y_2) &= Var[(y_1-a'Y_2)+(a-b)'Y_2] \\ &= Var(y_1-a'Y_2)+(a-b)'Var(Y_2)(a-b) \\ &+ 2Cov[(y_1-a'Y_2),(a-b)'Y_2]. \end{aligned}$$

由于 $a=\Sigma_{22}^{-1}\Sigma_{21}$, 则有

$$\begin{aligned} Cov[(y_1-a'Y_2),Y_2] &= Cov(y_1,Y_2)-a'Cov(Y_2,Y_2) \\ &= \Sigma_{12}-a'\Sigma_{22} \\ &= 0. \end{aligned}$$

因此有

$$
\begin{aligned}
Var(y_1 - b'Y_2) &= Var(y_1 - a'Y_2) + (a-b)'Var(Y_2)(a-b) \\
&= Var(y_1 - a'Y_2) + (a-b)'\Sigma_{22}(a-b) \\
&\geq Var(y_1 - a'Y_2).
\end{aligned}
$$

同时有

$$
\begin{aligned}
Var(y_1 - a'Y_2) &= Var(y_1) + Var(a'Y_2) - 2Cov(y_1, a'Y_2) \\
&= \sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
&= Var(y_1|Y_2).
\end{aligned}
$$

#

由定理1知：$Var(y_1 - a'Y_2)$ 达最小意味着 $y_1 - \mu_1$ 与 $a'Y_2 - a'\mu_2$ 最接近, 即 y_1 与 $(\mu_1 - a'\mu_2) + a'Y_2$ 最接近.

因此可以用 $(p-1)$ 个预报因子 Y_2 的线性组合来预测单个因变量 y_1, 其最优斜率为 a, 最优截距为 $(\mu_1 - a'\mu_2)$.

注意到:

$$
\begin{aligned}
E(y_1|Y_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Y_2 - \mu_2) \\
&= \mu_1 + a'(Y_2 - \mu_2) \\
&= (\mu_1 - a'\mu_2) + a'Y_2.
\end{aligned}
$$

条件期望是最优(方差最小)的线性预测.

4.1.2 样本复相关系数

设总体 $X \overset{d}{\sim} N_p(\mu, \Sigma)$, 其样本为 $x_1, \cdots, x_n$. 考虑 X 的剖分 $X = (x^{(1)}, (X^{(2)})')'$. 记 $\bar{x}$, V 和 S 分别为样本均值, 样本离差阵和样本协差阵. 并对它们作相应剖分.

则由 $x^{(1)}$ 与 $X^{(2)}$ 的复相关系数

$$\rho_{x^{(1)}, X^{(2)}} = \sqrt{\frac{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}{\sigma_{11}}},$$

定义 $x^{(1)}$ 与 $X^{(2)}$ 的**样本复相关系数**为

$$r_{x^{(1)}, X^{(2)}} = \sqrt{\frac{V_{12}V_{22}^{-1}V_{21}}{v_{11}}},$$

以及 a 的估计为 $\hat{a} = V_{22}^{-1}V_{21}$. 不难知道, 它们分别是复相关系数 $\rho_{x^{(1)}, X^{(2)}}$ 和方向 a 的极大似然估计.

修正的样本复相关系数：

$$(r^*_{x^{(1)},X^{(2)}})^2 = r^2_{x^{(1)},X^{(2)}} - \frac{p-1}{n-p}(1 - r^2_{x^{(1)},X^{(2)}}).$$

用途：

当 $p = 1$ 时, 修正的样本复相关系数等于样本复相关系数, 但当 $p \geq 2$ 时, 修正的样本复相关系数小于样本复相关系数.

在对线性回归模型做模型选择时，常用修正的样本相关系数来判断是否再选入一个预报因子.

样本复相关系数的分布

1) $\boldsymbol{\Sigma_{12}=0}$ 的情形: $x^{(1)}$ 与 $X^{(2)}$ 独立.

由Wishart分布的独立分解性质知,

$$
\begin{aligned}
t_1 &= v_{11} - V_{12}V_{22}^{-1}V_{21} \overset{d}{\sim} \sigma_{11}\chi^2(n-p); \\
t_2 &= V_{22}^{-1/2}V_{21} \overset{d}{\sim} N_{p-1}(0, \sigma_{11}I_{p-1}),
\end{aligned}
$$

且 t_1 与 t_2 独立. 因此,

$$
\begin{aligned}
F &= \frac{n-p}{p-1}\cdot\frac{r^2_{x^{(1)},X^{(2)}}}{1-r^2_{x^{(1)},X^{(2)}}} = \frac{n-p}{p-1}\cdot\frac{V_{12}V_{22}^{-1}V_{21}}{v_{11}-V_{12}V_{22}^{-1}V_{21}} \\
&= \frac{t_2't_2/(p-1)}{t_1/(n-p)} \overset{d}{\sim} F(p-1, n-p).
\end{aligned}
$$

则由 F 可以检验 $x^{(1)}$ 与 $X^{(2)}$ 是否相互独立.

注：该检验与3.3.6中独立性检验在 m=2 的情形一致.

2) 一般情形:

考虑变换 $Y = diag(\sigma_{11}^{-1/2}, \Sigma_{22}^{-1/2})X$, 并对 Y 作相同剖分得两部分 $y^{(1)} = \sigma_{11}^{-1/2}x^{(1)}$, $Y^{(2)} = \Sigma_{22}^{-1/2}X^{(2)}$, 且

$$Y \overset{d}{\sim} N_p\left(\begin{pmatrix} \sigma_{11}^{-1/2}\mu_1 \\ \Sigma_{22}^{-1/2}\mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} \\ \sigma_{11}^{-1/2}\Sigma_{22}^{-1/2}\Sigma_{21} & I_{p-1} \end{pmatrix}\right),$$

则

$$\rho_{y^{(1)},Y^{(2)}} = \sqrt{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}/\sigma_{11}} = \rho_{x^{(1)},X^{(2)}}.$$

因此, 不失一般性, 可以假设 X 的协差阵为

$$\Sigma = \begin{pmatrix} 1 & \Sigma_{12} \\ \Sigma_{21} & I_{p-1} \end{pmatrix},$$

此时有 $\rho^2_{x^{(1)},X^{(2)}} = \Sigma_{12}\Sigma_{21}$. 简记 $\rho = \rho_{x^{(1)},X^{(2)}}$.

由Wishart分布的性质知:

(1) $V_{22} \overset{d}{\sim} W_{p-1}(n-1, I_{p-1})$;

(2) $t_1 = v_{11} - V_{12}V_{22}^{-1}V_{21} \overset{d}{\sim} \sigma_{1|2}\chi^2(n-p), \quad \sigma_{1|2} = 1-\rho^2$;

(3) 在 V_{22} 给定的条件下, $t_2 = V_{22}^{-1/2}V_{21} \overset{d}{\sim} N_{p-1}(V_{22}^{1/2}\Sigma_{21}, (1-\rho^2)I_{p-1})$;

(4) t_1 与 (t_2, V_{22}) 相互独立.

则在 V_{22} 给定的条件下, 有

$$u = t_2't_2 = V_{12}V_{22}^{-1}V_{21} \overset{d}{\sim} (1-\rho^2)\chi^2(p-1,\eta), \quad \eta = \frac{\Sigma_{12}V_{22}\Sigma_{21}}{1-\rho^2};$$

$$\eta \overset{d}{\sim} \tau\chi^2(n-1), \quad \tau = \frac{\Sigma_{12}\Sigma_{21}}{1-\rho^2} = \frac{\rho^2}{1-\rho^2},$$

由此可以导出 u 的密度函数.

又由于

$$z = \frac{r^2_{x^{(1)},X^{(2)}}}{1-r^2_{x^{(1)},X^{(2)}}} = \frac{V_{12}V_{22}^{-1}V_{21}}{v_{11} - V_{12}V_{22}^{-1}V_{21}} = \frac{u}{t_1},$$

可以导出 z 的分布, 进而推出 $R=r^2$ 的密度函数

$$\frac{(1-\rho^2)^{(n-1)/2}(1-R)^{(n-p-2)/2}}{\Gamma((n-1)/2)\Gamma((n-p)/2)}\sum_{k=0}^{\infty}\frac{\rho^{2k}R^{(p-1)/2+k-1}}{k!\Gamma((p-1)/2+k)}\Gamma^2((n-1)/2+k).$$

4.2 典型相关分析

复相关系数：“一对多”

一个单变量与一个向量之间的相关关系

典型相关系数：“多对多”

两个向量之间的相关关系

假定

$$\begin{pmatrix} X \\ Y \end{pmatrix} \overset{d}{\sim} N_{p+q}(\mu, \Sigma),\ \mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix},\ \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} > 0,$$

其中, $X \overset{d}{\sim} N_p(\mu_x, \Sigma_{xx})$, $Y \overset{d}{\sim} N_q(\mu_y, \Sigma_{yy})$.

分别考虑 X 和 Y 的线性组合: $a'X$ 与 $b'Y$, 其中, $a \in R^p, b \in R^q$.
同样地, 可用 $a'X$ 与 $b'Y$ 的相关系数最大值来描述 X 与 Y 的相关性,
即

$$\begin{aligned} \sup_{\substack{a\in R^p \\ b\in R^q}} \rho_{a'X,b'Y} &= \sup_{\substack{a\in R^p \\ b\in R^q}} \frac{Cov(a'X, b'Y)}{\sqrt{Var(a'X)}\sqrt{Var(b'Y)}} \\ &= \sup_{\substack{a\in R^p \\ b\in R^q}} \frac{a'\Sigma_{xy}b}{\sqrt{a'\Sigma_{xx}a}\sqrt{b'\Sigma_{yy}b}}. \end{aligned}$$

由极值的非负性转而考虑

$$\sup_{\substack{a\in R^p\\ b\in R^q}} (\rho_{a'X,b'Y})^2 = \sup_{\substack{a\in R^p\\ b\in R^q}} \frac{(a'\Sigma_{xy}b)^2}{(a'\Sigma_{xx}a)(b'\Sigma_{yy}b)}.$$

若令

$$\tilde{a} = \frac{a}{\sqrt{a'\Sigma_{xx}a}}, \quad \tilde{b} = \frac{b}{\sqrt{b'\Sigma_{yy}b}},$$

易得

$$Var(\tilde{a}'X) = 1, \quad Var(\tilde{b}'Y) = 1,$$
$$\rho_{\tilde{a}'X,\tilde{b}'Y} = Cov(\tilde{a}'X, \tilde{b}'Y) = \rho_{a'X,b'Y}.$$

因此上述条件极值问题化为

$$\sup_{\substack{a\in R^p\\ b\in R^q}} (\rho_{a'X,b'Y})^2 = \sup_{\substack{a\in R^p,\, a'\Sigma_{xx}a=1\\ b\in R^q,\, b'\Sigma_{yy}b=1}} (a'\Sigma_{xy}b)^2.$$

4.2.1 总体典型相关分析

矩阵二次型极值的一些性质

性质1. 设 C 是 $p \times q$ 的非零矩阵, A 和 B 分别是 p 阶和 q 阶的正定矩阵, 则

$$\sup_{\substack{x \in R^p \\ y \in R^q}} \frac{(x'Cy)^2}{(x'Ax)(y'By)} = \lambda_1^2 > 0,$$

其中, λ_1^2 是 $A^{-1}CB^{-1}C'$, $A^{-1/2}CB^{-1}C'A^{-1/2}$, $B^{-1}C'A^{-1}C$ 或 $B^{-1/2}C'A^{-1}CB^{-1/2}$ 的最大特征根, 且 $\lambda_1 > 0$.

上式在 $x = A^{-1/2}\alpha_1$, $y = B^{-1/2}\beta_1$ 时达极大, 且 $x'Cy = \lambda_1$, 其中, α_1 和 $\beta_1 = B^{-1/2}C'A^{-1/2}\alpha_1/\lambda_1$ 分别是 λ_1^2 作为 $A^{-1/2}CB^{-1}C'A^{-1/2}$ 和 $B^{-1/2}C'A^{-1}CB^{-1/2}$ 的最大特征根所对应的正则特征向量.

设 C 是秩为 k 的 $p \times q$ 的非零矩阵, $p \leq q$.

则 p 阶矩阵 CC' 的 p 个非负特征根中有 k 个正的特征根 $\lambda_1^2 \geq \cdots \geq \lambda_k^2 > 0$, 其余 $p-k$ 个特征根为 $\lambda_{k+1}^2 = \cdots = \lambda_p^2 = 0$.
同理, q 阶矩阵 $C'C$ 的 q 个非负特征根中有 $\lambda_1^2 \geq \cdots \geq \lambda_k^2 > 0$, 其余 $q-k$ 个特征根为 $\lambda_{k+1}^2 = \cdots = \lambda_q^2 = 0$.

记 $\alpha_1, \cdots, \alpha_p$ 为 CC' 的特征根所对应的正则正交特征向量,
$\beta_1, \cdots, \beta_q$ 为 $C'C$ 的特征根所对应的正则正交特征向量.

性质2. 在上述假设条件下，有

(1) 设 $x \in R^p, y \in R^q$, 则在 $x'x=1$, $y'y=1$ 的正则化约束条件下,

$\sup(x'Cy)^2=\lambda_1^2$, 当 $x=\alpha_1$, $y=\beta_1$ 时取最大值, 并且 $x'Cy=\lambda_1$;

(2) 对给定的 $1 \le m \le (p-1), x \in R^p, y \in R^q$ 满足如下约束条件:

正则化约束： $x'x=1$, $y'y=1$;

正交化约束：对任意 $1 \le i \le m$, 都有

$$\alpha_i'x=0,\ \beta_i'y=0,\ \alpha_i'Cy=0,\ \beta_i'C'x=0.$$

则在正则正交约束下有

(i) 当 $1 \le m \le (k-1)$ 时， $\sup(x'Cy)^2=\lambda_{m+1}^2$,

当 $x=\alpha_{m+1}$, $y=\beta_{m+1}$ 时取最大值, 且 $x'Cy=\lambda_{m+1}>0$;

(ii) 当 $k \le m \le (p-1)$ 时， $x'Cy=0$.

因此, 约束极值问题

$$\sup_{\substack{a\in R^p\\ b\in R^q}} (\rho_{a'X,b'Y})^2 = \sup_{\substack{a\in R^p,\, a'\Sigma_{xx}a=1\\ b\in R^q,\, b'\Sigma_{yy}b=1}} (a'\Sigma_{xy}b)^2,$$

的解为: $(a'\Sigma_{xy}b)^2$ 的最大值为 λ_1^2, λ_1^2 是 $\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$, $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2}$, $\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ 或 $\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的最大特征根, 在 $a_1 = \Sigma_{xx}^{-1/2}\alpha_1, b_1 = \Sigma_{yy}^{-1/2}\beta_1$ 时取最大值, 其中, α_1 和 $\beta_1 = \Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1/2}\alpha_1/\lambda_1$ 分别是 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2}$ 和 $\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的最大特征根 λ_1^2 所对应的正则特征向量.

因此, $a_1'X$ 与 $b_1'Y$ 的相关系数 $\rho_{a_1'X,b_1'Y} = \lambda_1 > 0$.
此时, a_1 和 b_1 分别是 $\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ 和 $\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ 的最大特征根 λ_1^2 所对应的特征向量.

当 $\Sigma_{xy} = 0$ 时, X 与 Y 相互独立. 因此, X 与 Y 的任意线性组合都相互独立, 相关系数的最大值也为零.

总体典型相关分析

(1) 设 $Cov(X,Y)=\Sigma_{xy}$ 的秩为 k, 则向量 X 与 Y 一共有 k 组(对)典型相关变量和 k 个典型相关系数.

(2) 设 $\alpha_1,\cdots,\alpha_k$ 和 $\beta_1,\cdots,\beta_k$ 分别是 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2}$ 和 $\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的 k 个非零特征根 $\lambda_1^2\geq\cdots\geq\lambda_k^2>0$ 所对应的正则正交特征向量, 其中 $\beta_i=\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1/2}\alpha_i/\lambda_i$, $1\leq i\leq k$.

令 $a_i = \Sigma_{xx}^{-1/2}\alpha_i$, $b_i = \Sigma_{yy}^{-1/2}\beta_i$, $1 \le i \le k$,

则称 $(a_i'X, b_i'Y)$ 为第 i 组(对)**典型相关变量**,

$a_i'X$ 与 $b_i'Y$ 的相关系数为 $\lambda_i > 0$,

称 λ_i 为第 i 个**典型相关系数**, $1 \le i \le k$.

a_i 和 b_i 分别是 $\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ 和 $\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ 的特征根 λ_i^2 所对应的特征向量, $1 \le i \le k$.

(3) 第 i 组(对)典型相关变量 $(a_i'X, b_i'Y)$ 是正则的, 即

$$Var(a_i'X) = a_i'\Sigma_{xx}a_i = 1,\ Var(b_i'Y) = b_i'\Sigma_{yy}b_i = 1,\ 1 \le i \le k.$$

(4) 典型相关变量 $\{(a_i'X, b_i'Y)\}_{i=1}^k$ 之间是正交的,

即对 $1 \le i < j \le k$, 有

$$Cov(a_i'X, a_j'X) = a_i'\Sigma_{xx}a_j = 0,\quad Cov(a_i'X, b_j'Y) = a_i'\Sigma_{xy}b_j = 0,$$
$$Cov(b_i'Y, b_j'Y) = b_i'\Sigma_{yy}b_j = 0,\quad Cov(b_i'Y, a_j'X) = b_i'\Sigma_{yx}a_j = 0.$$

(5) 第1组(对)典型相关变量 $(a_1'X, b_1'Y)$ 是下述条件极值问题的解:

在 $a'\Sigma_{xx}a = 1$, $b'\Sigma_{yy}b = 1$ 的条件下使得 $a'\Sigma_{xy}b$ 达到最大值,

其最大值为 $\lambda_1 > 0$.

(6) 对 $2 \le i \le k$, 第 i 组(对)**典型相关变量** $(a_i'X, b_i'Y)$ 是

下述条件极值问题的解：

在 $a'\Sigma_{xx}a = 1$, $b'\Sigma_{yy}b = 1$ 且对任意 $1 \le j \le (i-1)$ 都有

$a'\Sigma_{xx}a_j = a'\Sigma_{xy}b_j = b'\Sigma_{yx}a_j = b'\Sigma_{yy}b_j = 0$ 的条件下使得

$a'\Sigma_{xy}b$ 达到最大值, 其最大值为 $\lambda_i > 0$.

典型相关分析的作用

设 Σ_{xy} 的秩为 k, $k \le p \le q$.

则 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1/2}$ 的特征根

$\lambda_1^2 \ge \cdots \ge \lambda_k^2 > 0 = \lambda_{k+1}^2 = \cdots = \lambda_p^2$ 所对应的

正则正交特征向量为 $\alpha_1, \cdots, \alpha_p$.

相应地, $\Sigma_{yy}^{-1/2}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的特征根

$\lambda_1^2 \ge \cdots \ge \lambda_k^2 > 0 = \lambda_{k+1}^2 = \cdots = \lambda_q^2$ 所对应的

正则正交特征向量为 $\beta_1, \cdots, \beta_q$.

令

$$\begin{aligned}
&U = A'\Sigma_{xx}^{-1/2}X, \quad V = B'\Sigma_{yy}^{-1/2}Y, \\
&W_1 = C_1'\Sigma_{xx}^{-1/2}X, \quad W_2 = C_2'\Sigma_{yy}^{-1/2}Y, \\
&W = (W_1', W_2')',
\end{aligned}$$

其中, $A = (\alpha_1, \cdots, \alpha_k), \qquad B = (\beta_1, \cdots, \beta_k),$

$$C_1 = (\alpha_{k+1}, \cdots, \alpha_p), \quad C_2 = (\beta_{k+1}, \cdots, \beta_q).$$

则有

$$Cov\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \begin{pmatrix} I_k & \Lambda & 0 \\ \Lambda & I_k & 0 \\ 0 & 0 & I_{p+q-2k} \end{pmatrix},$$

其中 $\Lambda = diag(\lambda_1, \cdots, \lambda_k)$.

不难看出,U 与 V 的相关性等价于 X 与 Y 的相关性.

由于 $k \leq min(p,q)$, 因此用 U 与 V 分别代表 X 与 Y

可以起到数据降维的作用.

事实上, 若记 $U = (U_1, \cdots, U_k)', V = (V_1, \cdots, V_k)'$,

则 (U_i, V_i) 就是 X 与 Y 的第 i 组(对)典型相关变量,

它们的相关系数就是第 i 个典型相关系数 $\lambda_i > 0$, $1 \leq i \leq k$.

4.2.2 样本典型相关分析

设正态总体 (X,Y) 有样本 $\{(x_i,y_i)\}_{i=1}^n$, $n>p+q$, $p\le q$.

则 Σ_{xx},Σ_{yy} 和 Σ_{xy} 极大似然估计为

$$\begin{aligned}
\hat\Sigma_{xx} &= n^{-1}\sum_{i=1}^n (x_i-\bar x_n)(x_i-\bar x_n)',\\
\hat\Sigma_{yy} &= n^{-1}\sum_{i=1}^n (y_i-\bar y_n)(y_i-\bar y_n)',\\
\hat\Sigma_{xy} &= n^{-1}\sum_{i=1}^n (x_i-\bar x_n)(y_i-\bar y_n)'.
\end{aligned}$$

性质:

由于特征根是矩阵中各元素的非退化连续函数, 因此由正态随机向量组成的矩阵 $\hat\Sigma_{xx}^{-1/2}\hat\Sigma_{xy}\hat\Sigma_{yy}^{-1}\hat\Sigma_{yx}\hat\Sigma_{xx}^{-1/2}$ 的特征根是连续型随机变量. 那么, 该矩阵的特征根以概率1满足:

$$1>\hat\lambda_1^2>\cdots>\hat\lambda_p^2>0.$$

记 $\hat{\alpha}_1, \cdots, \hat{\alpha}_p$ 和 $\hat{\beta}_1, \cdots, \hat{\beta}_p$ 分别是 $\hat{\Sigma}_{xx}^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1/2}$ 和 $\hat{\Sigma}_{yy}^{-1/2}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2}$ 的特征根 $\hat{\lambda}_1^2, \cdots, \hat{\lambda}_p^2$ 所对应的正则正交特征向量.

令 $\hat{a}_i = \hat{\Sigma}_{xx}^{-1/2}\hat{\alpha}_i$, $\hat{b}_i = \hat{\Sigma}_{yy}^{-1/2}\hat{\beta}_i$, $1 \le i \le p$.

则称 $(\hat{a}_i'X, \hat{b}_i'Y)$ 为第 i 组(对)**样本**典型相关变量,

称 $\hat{\lambda}_i$ 为第 i 个**样本**典型相关系数, $1 \le i \le p$.

不难知道, $\hat{a}_i$ 和 $\hat{b}_i$ 分别是 $\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1}\hat{\Sigma}_{yx}$ 和 $\hat{\Sigma}_{yy}^{-1}\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}$ 的特征根 $\hat{\lambda}_i^2$ 所对应的特征向量, $1 \le i \le p$.

$(\hat{a}_i'X, \hat{b}_i'Y)$ 和 $\hat{\lambda}_i$ 分别是**总体典型相关变量** $(a_i'X, b_i'Y)$ 和**总体典型相关系数** λ_i 的**极大似然估计**, $1 \le i \le p$.

问题:

由于 $1 > \hat{\lambda}_1^2 > \cdots > \hat{\lambda}_p^2 > 0$, 如何判断有意义的典型相关变量?

即给出一个估计 $k \le p$, 认为

$$\lambda_1^2 > \cdots > \lambda_k^2 > 0,$$
$$\lambda_{k+1}^2 = \cdots = \lambda_p^2 = 0.$$

4.2.3 典型相关变量个数的检验

设正态总体 (X, Y) 有样本 $\{(x_i, y_i)\}_{i=1}^n$, $n > p+q$, $p \le q$.

记 V 为样本离差阵,

$$V = \begin{pmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{pmatrix}, \quad V_{xx} = \sum_{i=1}^n (x_i - \bar{x}_n)(x_i - \bar{x}_n)',$$

$$V_{yy} = \sum_{i=1}^n (y_i - \bar{y}_n)(y_i - \bar{y}_n)', \quad V_{xy} = \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)'.$$

1) 典型相关变量的个数等于0 $v.s.$ **大于0的检验问题**

典型相关变量个数 $k = 0 \Longleftrightarrow \Sigma_{xy} = 0$, 即 X 与 Y 独立.

因此上述检验问题等价于 X 与 Y 是否相互独立的检验问题.

似然比统计量为

$$\begin{aligned}\lambda &= \left(\frac{|V|}{|V_{xx}||V_{yy}|}\right)^{n/2} = \left(\frac{|V_{xx} - V_{xy}V_{yy}^{-1}V_{yx}|}{|V_{xx}|}\right)^{n/2} \\ &= |I_p - V_{xx}^{-1}V_{xy}V_{yy}^{-1}V_{yx}|^{n/2} = \left[\prod_{i=1}^{p}(1-\hat{\lambda}_i^2)\right]^{n/2}.\end{aligned}$$

令

$$T_0 = \prod_{i=1}^{p}(1-\hat{\lambda}_i^2) = \frac{|V_{xx} - V_{xy}V_{yy}^{-1}V_{yx}|}{|(V_{xx} - V_{xy}V_{yy}^{-1}V_{yx}) + V_{xy}V_{yy}^{-1}V_{yx}|},$$

在零假设 $\Sigma_{xy} = 0$ 下, 有 $T_0 \overset{d}{\sim} \Lambda_{p,n-1-q,q}$. 因此可以用零分布 $\Lambda_{p,n-1-q,q}$ 构造检验方案. 也可由似然比检验统计量的渐近分布

$$-2\log(\lambda) = -n\sum_{i=1}^{p}\log(1-\hat{\lambda}_i^2) \xrightarrow{d} \chi^2(pq)$$

构造检验方案.

2) 典型相关变量的个数等于 k $\quad v.s.\;$ 大于 k 的检验问题

等价于检验问题:

$$\mathbf{H}_0 : rank(\Sigma_{xy}) = k \quad v.s. \quad \mathbf{H}_1 : rank(\Sigma_{xy}) > k.$$

也等价于检验问题:

$$\mathbf{H}_0 : \lambda_k^2 > 0,\, \lambda_{k+1}^2 = 0 \quad v.s. \quad \mathbf{H}_1 : \lambda_{k+1}^2 > 0.$$

也等价于检验问题:

$$\mathbf{H}_0 : \text{存在}\, p \times (p-k)\, \text{的列满秩矩阵}\, C,\, \text{使得}\, \Sigma_{yy}^{-1/2}\Sigma_{yx}C = 0.$$

似然比统计量为

$$
\begin{aligned}
\lambda &= \sup_{C} \frac{|C'(V_{xx} - X_{xy}V_{yy}^{-1}V_{yx})C|}{|C'V_{xx}C|} \\
&= \sup_{C'C=I_{p-k}} |C'(I_p - V_{xx}^{-1/2}V_{xy}V_{yy}^{-1}V_{yx}V_{xx}^{-1/2})C| \\
&= \sup_{D'D=I_{p-k}} |D'diag(1-\hat{\lambda}_1^2, \cdots, 1-\hat{\lambda}_p^2)D| \\
&= \Big[\prod_{i=k+1}^{p} (1-\hat{\lambda}_i^2)\Big]^{n/2},
\end{aligned}
$$

其中 C 是任意 $p \times (p-k)$ 的列满秩矩阵.

由于 $rank(\Sigma_{xy}) = k$, 因此有

$$\Sigma_{xy} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix},$$

其中 C_1 是 $k \times q$ 的满秩阵, 而 $C_2 = QC_1$, Q 是$(p-k) \times k$ 的矩阵.
因此, 有

$$\begin{aligned} &dim(\Theta_0) = p + q + p(p+1)/2 + q(q+1)/2 + kq + (p-k)k, \\ &dim(\Theta) - dim(\Theta_0) = (p-k)(q-k). \end{aligned}$$

由Wilks定理知

$$-2\log(\lambda) = -n \sum_{i=k+1}^{p} \log(1-\hat{\lambda}_i^2) \xrightarrow{d} \chi^2((p-k)(q-k)),$$

进而构造检验方案. 也可以采用修正的统计量

$$-\left(n - 1 - k - \frac{p+q+1}{2} + \sum_{i=1}^{k} \hat{\lambda}_i^2\right) \sum_{i=k+1}^{p} \log(1-\hat{\lambda}_i^2).$$

4.3 广义相关系数

设随机向量 $X_{p\times 1}$ 和 $Y_{q\times 1}$，记

$$\Sigma = Cov\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} > 0.$$

称 $M_{yx} = V_{yy}^{-1}V_{yx}V_{xx}^{-1}V_{xy}$ 为 X 和 Y 的**线性关联阵**.

记 $k = rank(M_{yx})$, 称 k 为 X, Y 的**相关秩**.

记线性关联阵 M_{yx} 的非零特征根为 $\lambda_1^2 \geq \cdots \geq \lambda_k^2 > 0$.

则下面定义的每个量都称为 X 与 Y 的**广义相关系数**:

$$\begin{aligned}
\rho_{xy}^{(1)} &= (\prod_{i=1}^{k} \lambda_i^2)^{1/k}; \\
\rho_{xy}^{(2)} &= k^{-1}(\sum_{i=1}^{k} \lambda_i^2); \\
\rho_{xy}^{(3)} &= \lambda_1^2; \\
\rho_{xy}^{(4)} &= \lambda_k^2; \\
\rho_{xy}^{(5)} &= (k^{-1} \sum_{i=1}^{k} \lambda_i^{-2})^{-1}.
\end{aligned}$$

4.4 实例分析

家庭特征与家庭消费之间的关系

为了了解家庭的特征与其消费模式之间的关系，调查了70个家庭的下面两组变量：

消费模型变量

$$\begin{cases} X_1\text{：每年去餐馆就餐的频率} \\ X_2\text{：每年外出看电影频率} \end{cases}$$

家庭特征变量

$$\begin{cases} Y_1\text{：户主的年龄} \\ Y_2\text{：家庭的年收入} \\ Y_3\text{：户主受教育程度} \end{cases}$$

目的：分析两组变量之间的关系.

变量间的相关系数矩阵

	X1	X2	y1	y2	y3
X1	1.00	0.80	0.26	0.67	0.34
X2	0.80	1.00	0.33	0.59	0.34
y1	0.26	0.33	1.00	0.37	0.21
y2	0.67	0.59	0.37	1.00	0.35
y3	0.34	0.34	0.21	0.35	1.00

典型相关分析		
	典型相关系数	典型相关系数的平方
1	0.687948	0.473272
2	0.186865	0.034919

X组典型变量的系数		
	U1	U2
X1(就餐)	0.7689	-1.4787
X2(电影)	0.2721	1.6443
Y组典型变量的系数		
	V1	V2
Y1(年龄)	0.0491	1.0003
Y2(收入)	0.8975	-0.5837
Y3(文化)	0.1900	0.2956

$U_1 = 0.7689X_1 + 0.2721X_2$; $V_1 = 0.0491Y_1 + 0.8975Y_2 + 0.1900Y_3$;

$U_2 = -1.4787X_1 + 1.6443X_2$; $V_2 = 1.0003Y_1 - 0.5837Y_2 + 0.2956Y_3$.

典型变量的结构（相关系数）

	U1	U2
X1	0.9866	-0.1632
X2	0.8872	0.4614
	V1	V2
Y1	0.4211	0.8464
Y2	0.9822	-0.1101
Y3	0.5145	0.3013

典型变量的结构（相关系数）		
	V1	V2
X1	0.6787	-0.0305
X2	0.6104	0.0862
	U1	U2
Y1	0.2897	0.1582
Y2	0.6757	-0.0206
Y3	0.3539	0.0563

1）两个反映消费的指标与第一对典型变量中U_1的相关系数分别为0.9866和0.8872，可以看出U_1可以作为消费特性的指标；

2）第一对典型变量中V_1与Y_2之间的相关系数为0.9822，可见典型变量V_1主要代表了家庭收入；

3）U_1和V_1的相关系数为0.6879，这就说明家庭的消费与一个家庭的收入之间其关系是很密切的.

检验典型相关变量的个数: $k=0 \quad v.s. \quad k>0$.

这时, 样本量 $n=70$, $p=2$, $q=3$.

检验统计量为:

$$-2\log(\lambda) = -n\sum_{i=k+1}^{p}\log(1-\hat{\lambda}_i^2) = -70[\log(1-\hat{\lambda}_1^2)+\log(1-\hat{\lambda}_2^2)]$$
$$= -70[\log(1-0.473272)+\log(1-0.034919)] = 47.363.$$

计算检验的 p值:

$$P\{\chi^2(pq) > -2\log(\lambda)\} = P\{\chi^2(6) > 47.363\}$$
$$= 1.584\times 10^{-8} < 0.05.$$

结论: 拒绝 $k=0$ 的假设, 即不能认为两组变量不相关.

检验典型相关变量的个数: $k=1 \quad v.s. \quad k>1$.

检验统计量为:

$$
\begin{aligned}
-2\log(\lambda) &= -n\sum_{i=k+1}^{p}\log(1-\hat{\lambda}_i^2) = -70\log(1-\hat{\lambda}_2^2) \\
&= -70\log(1-0.034919) = 2.488.
\end{aligned}
$$

计算检验的p值:

$$
\begin{aligned}
P\{\chi^2((p-k)(q-k)) > -2\log(\lambda)\} &= P\{\chi^2(2) > 2.488\}, \\
&= 0.288 > 0.05.
\end{aligned}
$$

结论: 没有足够证据拒绝零假设.

可以认为典型相关变量的个数为1.