典型相关分析

《美国数学建模竞赛》

完整课程请长按下方二维码

典型相关分析

现实生活中两组变量间的相关关系的问题很多，例如家庭的特征（如户主的年龄、家庭的年收入、户主的受教育程度等）与消费模式（如每年去餐馆就餐的频率、每年外出看电影的频率等）等等。为此，1936年由Hulling提出了典型相关分析，揭示了两组多元随机变量之间的关系。

典型相关分析基本思想

通常情况下，为了研究两组变量

$(x_1, x_2, \cdots, x_p)$ $(y_1, y_2, \cdots, y_q)$

的相关关系，可以用最原始的方法，分别计算两组变量之间的全部相关系数，一共有pq个简单相关系数，这样又烦琐又不能抓住问题的本质。如果分别找出两组变量的各自的某个线性组合，讨论线性组合之间的相关关系，则更简捷。

首先分别在每组变量中找出第一对线性组合，使其具有最大相关性，

$$\begin{cases} u_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ v_1 = b_{11}y_1 + b_{21}y_2 + \cdots + b_{q1}y_q \end{cases}$$

然后再在每组变量中再找出第二对线性组合，使其分别与本组内的第一线性组合不相关，第二对本身具有次大的相关性，即u_2和v_2与u_1和v_1相互独立，但u_2和v_2相关，

$$\begin{cases} u_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ v_2 = b_{12}y_1 + b_{22}y_2 + \cdots + b_{q2}y_q \end{cases}$$

如此下去，直至两组变量的相关性被提取完为止。

例：蔬菜产出水平主要体现在蔬菜总产量(Y1)、人均蔬菜占有量(Y2)、蔬菜总产增长速度(Y3)三个方面，并称作因变量组（简称“产出组”）。问题：因变量组与自变量X1(市场经济综合因素)、X2(劳动力动力因素)、X3(气候因素)(简称“影响组”)的关系如何？

```
data ex;
input y1-y3  x1-x3 @@;
cards;
/*数据省略*/
proc cancorr data=ex all;var y1-y3; with  x1-x3;
run;
```

程序运行结果如下：

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.982193	0.977396	0.010189	0.964703
2	0.810462	0.784388	0.099059	0.656848
3	0.439231	.	0.232983	0.192924

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

Test of H0: The canonical correlations in the current row and all that follow are zero

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	27.3309	25.4168	0.9270	0.9270	0.00977553	10.88	9	17.187	<.0001
2	1.9142	1.6751	0.0649	0.9919	0.27694975	3.60	4	16	0.0282
3	0.2390		0.0081	1.0000	0.80707608	2.15	1	9	0.1765

整理得到蔬菜产出水平与影响因素的三个自变量的典型相关系数及特征值

序号	典型相关系数	标准误差	特征值	方差比率	累计方差比率
1	0.982193	0.010189	27.3309	0.9270	0.9270
2	0.810462	0.099059	1.9142	0.0649	0.9919
3	0.439231	0.232983	0.2390	0.0081	1.0000

结果表明：前两个典型相关系数较高，表明相应典型变量之间密切相关。

序号	F计算值	自由度	F检验的显著性概率
1	10.88	9	0.0001
2	3.60	4	0.0282
3	2.15	1	0.1765

结果表明：只有前两对典型变量通过了统计量检验，表明相应典型变量之间相关关系显著，能够用三个自变量影响变量来解释产出变量。

冗余度分析的结果

Canonical Redundancy Analysis

Canonical Variable Number	Standardized Variance of the VAR Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Standardized Variance of the VAR Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.6481	0.6481	0.9647	0.6253	0.6253
2	0.2054	0.8535	0.6568	0.1349	0.7602
3	0.1465	1.0000	0.1929	0.0283	0.7884

Canonical Variable Number	Standardized Variance of the WITH Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Standardized Variance of the WITH Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.3335	0.3335	0.9647	0.3217	0.3217
2	0.3338	0.6672	0.6568	0.2192	0.5409
3	0.3328	1.0000	0.1929	0.0642	0.6051

典型变量的解释能力

序号	产出组与影响组典型相关系数平方	对产出组解释能力	产出组方差被影响组典型变量解释比例	对影响组解释能力	影响组方差被产出租典型变量解释比例
1	0.9647	0.6481	0.6253	0.3335	0.3217
2	0.6568	0.2054	0.1349	0.3338	0.2192
3	0.1929	0.1465	0.0283	0.3328	0.0642

可以看出：①前两对典型变量的解释能力均较强；②第一、第二对典型变量具有较高的解释百分比，典型相关系数的平方表明，产出变量中分别有96.47%和65.68%的信息可以由相应的影响变量予以解释；③前两对典型变量的重叠系数较大，产出组的方差被影响组典型变量解释的比例分别为62.53%、13.49%。由于第三对典型变量在上述②、③项指标中的数值均较小，且未能通过F检验．因此舍弃第三对典型变量，只选定前两对典型变量进行分析。

典型相关模型结果如下：

The CANCORR Procedure

Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables

	V1	V2	V3
y1	6.1649	14.7443	-19.9180
y2	-5.2034	-15.0750	19.9861
y3	0.0696	0.9105	0.4820

Standardized Canonical Coefficients for the WITH Variables

	W1	W2	W3
x1	0.9953	-0.0132	-0.0962
x2	-0.0054	0.9591	-0.2831
x3	-0.0948	-0.2804	-0.9552

序号	典型相关模型
1	$v1=6.1649\ Y_1-5.2034\ Y_2+0.0696\ Y_3$ $w1=0.9953X_1-0.0054\ X_2-0.0948X_3$
2	$v2=14.7443Y_1-15.0750Y_2+0.9105Y_3$ $w2=-\ 0.0132\ X_1+0.9591\ X_2-0.2804\ X_3$

结果分析：自变量X1即市场经济综合因素对中国蔬菜产出水平起根本性作用。市场经济综合因素与蔬菜总产出的关系体现在第一对典型变量v1和w1中，v1是中国蔬菜产出水平各指标的线性组合，其中，蔬菜总产出（Y1）的载荷为6.164，是各产出水平指标中最大的。w1是影响因素指标的线性组合，其中市场经济综合因素(X1)的载荷为0.9953，远远超过w1内其它指标的数值。考虑到第一对典型相关变量的相关系数几乎接近于1，可以认为，市场经济综合因素对蔬菜总产出水平起根本性作用。自变量X2即劳动力动力因素是决定人均蔬菜占有量的关键因素。

第二对典型变量中．人均蔬菜占有量(Y2)在典型变量v2中的载荷为一15.075，是各产出水平指标中最大的，而自变量X2则在典型变量w2中载荷最大，为0.9591。这一对典型相关变量的相关系数非常之高，表明自变量X2对劳动力动力因素起关键作用.

在第二对典型变量中，Y1与劳动力动力因素关系也非常密切。因为在第二对典型变量中，Y1在v2中的载荷14.7443，与Y2差距并不明显。由此可以分析的处，用Y1作为产出水平的代表，X1、X2、X3作为影响变量建立因果拟合模型效果是最好的。