

回归的正交设计

蒋 炳 鑫

摘要 本文借助于线性变换及正交表的特性,提供一个解线性回归方程的简便方法。

这个方法适用于任意位级和混合位级的等距多因素正交试验。

关键词 正交试验; 线性; 回归方程

0 引 言

正交试验设计在工农业生产及科学研究中已有了广泛的应用,它以较少次数的试验提供了含有最大信息的数据,给出了寻求最优工艺和配方的途径。大量实践证明它是处理多因素问题的行之有效的办法。

随着研究的深入往往还要寻求数学模型——指标和因素之间的回归方程。然而多元线性回归方程的求得需要大量的计算。而本文将提供一个对于任意位级以至混合位级的等距多因素正交试验都适用的较为简便的方法。

1 线性代换

1.1 问题

a 设考察的指标为 y ,考察的因素为 m 个等距 q 位级因素。要求通过正交表 $L_n(q^m)$ 排出试验方案,并求出线性回归方程。

b 设考察的指标为 y ,考虑的因素为 m_1 个等距 q_1 位级因素和 m_2 个等距 q_2 位级因素。要求通过正交表 $L_n(q_1^{m_1} \times q_2^{m_2})$ 排出试验方案,并求出线性回归方程。

1.2 线性代换

设第 r 个因素 $z^{(r)}$ 的取值为

$$z_i^{(r)} = z_1^{(r)} + (i-1)h^{(r)} \quad (i=1,2,\dots,q^{(r)})$$

其中

$h^{(r)} > 0$ 为因素 $z^{(r)}$ 的级距

$q^{(r)}$ 为因素 $z^{(r)}$ 的位级数

令

$$x^{(r)} = \frac{2z^{(r)} - (z_1^{(r)} + z_{q^{(r)}}^{(r)})}{z_{q^{(r)}}^{(r)} - z_1^{(r)}}$$

则

$$x_i^{(r)} = \frac{2z_i^{(r)} - (z_1^{(r)} + z_q^{(r)})}{z_1^{(r)} - z_q^{(r)}} = \frac{2i - q^{(r)} - 1}{q^{(r)} - 1} \quad i = 1, 2, \dots, q^{(r)} \quad (1)$$

通过线性代换后, $x_i^{(r)}$ 的取值仅与 i 和 $q^{(r)}$ 有关, 而与 $z_1^{(r)}$ 和 $h^{(r)}$ 无关。下面对正交设计中常用的位级列出代换后的取值。

线性代换后 $x^{(r)}$ 的取值

$q^{(r)}$	$z_1^{(r)}$												
	$z_1^{(r)}$	$z_2^{(r)}$	$z_3^{(r)}$	$z_4^{(r)}$	$z_5^{(r)}$	$z_6^{(r)}$	$z_7^{(r)}$	$z_8^{(r)}$	$z_9^{(r)}$	$z_{10}^{(r)}$	$z_{11}^{(r)}$	$z_{12}^{(r)}$	$z_{13}^{(r)}$
2	-1	1											
3	-1	0	1										
4	-1	-1/3	1/3	1									
5	-1	-1/2	0	1/2	1								
6	-1	-3/5	-1/5	1/5	3/5	1							
7	-1	-2/3	-1/3	0	1/3	2/3	1						
8	-1	-5/7	-3/7	-1/7	1/7	3/7	5/7	1					
9	-1	-3/4	-2/4	-1/4	0	1/4	2/4	3/4	1				
11	-1	-4/5	-3/5	-2/5	-1/5	0	1/5	2/5	3/5	4/5	1		
13	-1	-5/6	-4/6	-3/6	-2/6	-1/6	0	1/6	2/6	3/6	4/6	5/6	1

显见

$$a. \quad \sum_{i=1}^{q^{(r)}} x_i^{(r)} = 0$$

b. 任意两个同位级的因素 $z^{(s)}$ 、 $z^{(t)}$ 不论它们的取值是否相同, 通过上述线性代换都有

$$x_i^{(s)} = x_i^{(t)} \quad i = 1, 2, \dots, q \quad (q = q^{(s)} = q^{(t)})$$

为了证明方便, 设

$$x_i^{(s)} = x_i^{(t)} = x_i \quad i = 1, 2, \dots, q$$

此时有

$$\sum_{i=1}^q x_i = 0$$

2 定 理

在通过上述线性代换后, 求关于指标 y 对因素 $z^{(1)}$, $z^{(2)}$, $\dots$, $z^{(m)}$ 的线性回归方程就转化为求 y 对 $x^{(1)}$, $x^{(2)}$, $\dots$, $x^{(m)}$ 的线性回归方程, 而后的计算将是方便的。

在应用上述线性代换后正交表中各因素的取值即为 $x_1, x_2, \dots, x_q$; 此时由于正交表的

特点有下列定理。

定理1 对于正交表 $L_n(q^m)$, 任意两列所对应的因素处于同一横行的取值之积的总和为零, 即任两列的内积为零。

证明: 由正交表 $L_n(q^m)$ 的构造可知试验次数 $n = Nq^2$, 其中 N 依赖于 m 和 q 的某一正整数。

当 $N=1$ 时 $n=q^2$, 由于正交表的匀称性任意两列(例如: i 列和 j 列)所对应的因素处于同一横行的取值数对为:

$$(x_l, x_1), (x_l, x_2), \dots, (x_l, x_q)$$

$$l=1, 2, \dots, q$$

共有 q^2 对。

当 $N \neq 1$ 时任意两列所对应的因素处于同一横行的取值数对仍是上面所列的那些, 不过每一数对都重复出现了 N 次。

所以若设任意两列(i 列和 j 列)所对应的因素处于同一横行的取值数对之积的总和为 S_{ij} , 则

$$\begin{aligned} S_{ij} &= N \sum_{K=1}^q x_{iK} x_{jK} + N \sum_{K=1}^q x_{jK} x_{iK} + \dots + N \sum_{K=1}^q x_{iK} x_{jK} \\ &= N (x_{i1} \sum_{K=1}^q x_{jK} + x_{i2} \sum_{K=1}^q x_{jK} + \dots + x_{iq} \sum_{K=1}^q x_{jK}) \\ &= N \sum_{K=1}^q x_{iK} \cdot \sum_{l=1}^q x_{jl} \\ &= 0 \end{aligned}$$

定理2 对正交表 $L_n(q_1^{m_1} \times q_2^{m_2})$ 任意两列所对应的因素处于同一横行的取值之积的总和为零, 即任两列的内积为零。

证明: 设第 i 列和第 j 列所对应的因素处于同一横行的取值之积的总和为 S_{ij} 。

由正交表的构造可知, 试验次数 n 为 q_1^a 和 q_2^b 的公倍数。

其中

$$a = \begin{cases} 1 & m_1 = 1 \\ 2 & m_1 > 1 \end{cases} \quad b = \begin{cases} 1 & m_2 = 1 \\ 2 & m_2 > 1 \end{cases}$$

例: 对 $L_{16}(4^2 \times 2^6)$ 有 $n=16$ 为 4^2 和 2^2 的公倍数,

对 $L_{18}(6^1 \times 3^6)$ 有 $n=18$ 为 6 和 3^2 的公倍数。

下面分两种情况证明,

a. 设第 i 列和第 j 列所对应的因素为同位级因素, 不失一般性设它们都是 q_1 位级的因素。此时由于正交表的匀称性, 这两列所对应的因素处于同一横行的取值为:

$$(x_l, x_1), (x_l, x_2), \dots, (x_l, x_{q_1})$$

$$l=1, 2, \dots, q_1$$

而每一数对都重复出现 n/q_1^2 次。

因此仿照定理1的证明即得:

$$S_{ij} = n/q_1^2 \sum_{l=1}^{q_1} \sum_{k=1}^{q_1} x_l x_k = n/q^2 \sum_{k=1}^{q_1} x_k \sum_{l=1}^{q_1} x_l = 0$$

同理, 若都是 q_2 位级的因素, 则有

$$S_{ij} = n/q_2^2 \sum_{k=1}^{q_2} x_k \sum_{l=1}^{q_2} x_l = 0$$

b. 设第 i 列和第 j 列所对应的因素分别是 q_1 位级和 q_2 位级的因素。此时由正交表的匀称性有:

q_1 位级因素的取值为 $x_1^{(1)}, x_2^{(1)}, \dots, x_{q_1}^{(1)}$, 且各重复 n/q_1 次。

q_2 位级因素的取值为 $x_1^{(2)}, x_2^{(2)}, \dots, x_{q_2}^{(2)}$, 且各重复 n/q_2 次。

这两因素处于同一横行的取值为:

$$(x_1^{(1)}, x_K^{(2)}), (x_2^{(1)}, x_K^{(2)}), \dots, (x_{q_1}^{(1)}, x_K^{(2)}) \\ K = 1, 2, \dots, q_2$$

而每一数对都重复出现 $n/q_1 q_2$ 次。

故

$$\begin{aligned} S_{ij} &= n/q_1 q_2 \left(\sum_{K=1}^{q_2} x_1^{(1)} x_K^{(2)} + \sum_{K=1}^{q_2} x_2^{(1)} x_K^{(2)} + \dots + \sum_{K=1}^{q_2} x_{q_1}^{(1)} x_K^{(2)} \right) \\ &= n/q_1 q_2 \left(x_1^{(1)} \sum_{K=1}^{q_2} x_K^{(2)} + x_2^{(1)} \sum_{K=1}^{q_2} x_K^{(2)} + \dots + x_{q_1}^{(1)} \sum_{K=1}^{q_2} x_K^{(2)} \right) \\ &= n/q_1 q_2 \sum_{K=1}^{q_2} x_K^{(2)} \cdot \sum_{l=1}^{q_1} x_l^{(1)} \\ &= 0 \end{aligned}$$

定理3, 对于正交表 $L_n(q^m)$, $L_n(q_1^{m_1} \times q_2^{m_2})$ 任意一列所对应的因素取值之和为零。

证明: 设第 i 列所对应的因素取值之和为 S_i 。

a. 对正交表 $L_n(q^m)$ 第 i 列所对应的因素的取值为 $x_1, x_2, \dots, x_q$ 其中 $\sum_{i=1}^q x_i = 0$

且每一取值都重复 n/q 次, 故 $S_i = n/q \sum_{i=1}^q x_i = 0$

b. 对正交表 $L_n(q_1^{m_1} \times q_2^{m_2})$,

若

第 i 列所对应的因素为 q_1 位级, 则数值 $x_1^{(1)}, x_2^{(1)}, \dots, x_{q_1}^{(1)}$ 各重复出现 n/q_1 次且 $\sum_{K=1}^{q_1} x_K^{(1)} = 0$,

故

若

$$S_i = n/q_1 \sum_{K=1}^{q_1} x_K^{(1)} = 0$$

(2)

第*i*列所对应的因素为 q_2 位级, 则数值 $x_1^{(2)}, x_2^{(2)}, \dots, x_{q_2}^{(2)}$ 各重复出现 n/q_2 次且 $\sum_{K=1}^{q_2} x_K^{(2)} = 0$,

故

$$S_i = n/q_2 \sum_{K=1}^{q_2} x_K^{(2)} = 0$$

由上述定理可见, 在对各因素进行线性代换(1)以后再利用正交表进行试验设计, 则在试验计划中有:

任一列的和 $S_i = 0$

任两列的内积 $S_{ij} = 0$

故称这种试验设计为正交设计, 所得试验计划为正交计划。

3 线性回归方程的求法

3.1 正交计划中的线性回归方程

定理4, 没有指标 y , 考察的因素为 $x_1, x_2, \dots, x_m$ 。根据正交表做 n 次试验, 试验数据为 $(y_K, x_{1K}, x_{2K}, \dots, x_{mK}) K=1, 2, \dots, n$ 。

若设 y 对 $x_1, x_2, \dots, x_m$ 的线性回归方程为 $y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m$

则

$$b_i = l_{iy}/l_{ii} \quad b_0 = \bar{y} = 1/n \sum_{K=1}^n y_K$$

其中

$$l_{iy} = \sum_{K=1}^n x_{iK} y_K$$

$$l_{ii} = \sum_{K=1}^n x_{iK}^2 \quad i=1, 2, \dots, m$$

(l_{ii} 可根据正交表表头查附表得到)

证明: 已知在一般情况下 $b_0, b_1, \dots, b_m$ 应满足如下正规方程。

$$\begin{cases} l_{11}b_1 + l_{12}b_2 + \dots + l_{1m}b_m = l_{1y} \\ l_{21}b_1 + l_{22}b_2 + \dots + l_{2m}b_m = l_{2y} \\ \dots\dots\dots \\ l_{m1}b_1 + l_{m2}b_2 + \dots + l_{mm}b_m = l_{my} \end{cases} \quad (2)$$

$$b_0 = \bar{y} - \sum_{i=1}^m b_i \bar{x}_i$$

其中

$$\bar{y} = 1/n \sum_{K=1}^n y_K$$

$$\bar{x}_i = 1/n \sum_{K=1}^n x_{iK} \quad i = 1, 2, \dots, m$$

$$l_{ij} = l_{ji} = \sum_{K=1}^n x_{iK} x_{jK} - n \bar{x}_i \bar{x}_j \quad i, j = 1, 2, \dots, m$$

$$l_{iy} = \sum_{K=1}^n x_{iK} y_K - n \bar{x}_i \bar{y} \quad i = 1, 2, \dots, m$$

而根据定理三、四可知

$$\bar{x}_i = 1/n \sum_{K=1}^n x_{iK} = S_i/n = 0 \quad i = 1, 2, \dots, m$$

$$l_{ij} = \sum_{K=1}^n x_{iK} x_{jK} = S_{ij} = 0 \quad i \neq j, i, j = 1, 2, \dots, m$$

$$l_{ii} = \sum_{K=1}^n x_{iK}^2 \quad i = 1, 2, \dots, m$$

(对 l_{iy} , 为了应用的方便, 已对常用正交表进行计算。其结果见附表)

$$l_{iy} = \sum_{K=1}^n x_{iK} y_K \quad i = 1, 2, \dots, m$$

故正规方程(2)化为

$$\begin{cases} l_{11}b_1 & = l_{1y} \\ l_{22}b_2 & = l_{2y} \\ & \vdots \\ l_{mm}b_m & = l_{my} \end{cases}$$

即

$$b_0 = \bar{y}$$

$$b_i = l_{iy}/l_{ii} \quad i = 1, 2, \dots, m$$

3.2 正交计划中回归方程的显著性检验

3.2.1 算统计量 $F = \frac{(n-m-1)U}{m(l_{yy}-U)}$

其中

$$l_{yy} = \sum_{K=1}^n y_K^2 - n \bar{y}^2$$

$$U = \sum_{K=1}^n (\hat{y}_K - \bar{y})^2 = \sum_{i=1}^m b_i l_{iy}$$

3.2.2 查表对已给的显著水平 α , 查 $N_1 = m$, $N_2 = n - m - 1$ 的 F 分布表得单侧临界值 F_α .

3.2.3 判断, 若 $F > F_\alpha$ 则回归效果显著, 否则不显著。

3.3 正交计划中的预测区间

算统计量 $Q = l_{yy} - U$

得 y 的置信概率为0.95的预测区间为

$$\left(b_0 + \sum_{i=1}^m b_i x_i - 1.96 \sqrt{\frac{Q}{n-m-1}}, b_0 + \sum_{i=1}^m b_i x_i + 1.96 \sqrt{\frac{Q}{n-m-1}} \right)$$

3.4 使用步骤

设利用正交表进行 n 次试验后, 经直观分析得某指标 y 和因素 $z^{(r)}$ ($r = 1, 2, \dots, m$)之间存在线性关系。 $z^{(r)}$ 的取值为 $z_i^{(r)} = z_1^{(r)} + (i-1)h^{(r)}$ $i = 1, 2, \dots, q^{(r)}$

其中

$q^{(r)}$ 为 $z^{(r)}$ 的位级数, $h^{(r)} > 0$ 为 $z^{(r)}$ 的级距。

a. 作线性代换

设

$$x^{(r)} = \frac{2z^{(r)} - (z_1^{(r)} + z_{q^{(r)}}^{(r)})}{z_{q^{(r)}}^{(r)} - z_1^{(r)}} \quad r = 1, 2, \dots, m \quad (1)$$

由附表一可查得 $x^{(r)}$ 的取值。

b. 求 y 对 $x^{(1)}, x^{(2)}, \dots, x^{(m)}$ 的线性回归方程。

设第 K 次试验的试验数据为 $(y_K, z_{1K}, z_{2K}, \dots, z_{mK})$ $K = 1, 2, \dots, n$ 。与之对应, 有数据 $(y_K, x_{1K}, x_{2K}, \dots, x_{mK})$

其中

z_{iK} 表示因素 $z^{(i)}$ 在第 K 次试验中的取值。

x_{iK} 表示与 z_{iK} 相对应的 $x^{(i)}$ 的取值。

计算 $l_{iy} = \sum_{K=1}^n x_{iK} y_K, \quad i = 1, 2, \dots, m$

$$b_i = l_{iy} / l_{ii}, \quad i = 1, 2, \dots, m, \quad b_0 = 1/n \sum_{K=1}^n y_K$$

其中

l_{ii} 可由所用正交表的表头查附录得到, 由此得 y 对 $x^{(1)}, x^{(2)}, \dots, x^{(m)}$ 的线性回归方程

$$\hat{y} = b_0 + b_1 x^{(1)} + b_2 x^{(2)} + \dots + b_m x^{(m)} \quad (2)$$

c. 求 y 对 $z^{(1)}, z^{(2)}, \dots, z^{(m)}$ 的线性回归方程

将(1)代入(2)并化简即得 y 对 $z^{(1)}, z^{(2)}, \dots, z^{(m)}$ 的线性回归方程

$$\hat{y} = a_0 + a_1 z^{(1)} + a_2 z^{(2)} + \dots + a_m z^{(m)} \quad (3)$$

b. 通过正交计划中回归方程的显著性检验决定(3)式是否有效.

e. 若(3)式有效, 求出 y 的预测区间.

4 结束语

本文提供的回归正交设计在很大程度上减少了求等距多因素线性回归方程的难度, 由于不必算 $l_{ij}(i \neq j, i, j=1, 2, \dots, m)$, $l_{ii}(i=1, 2, \dots, m)$, 不必解 m 元线性方程组(正规方程)计算量大大减少. 因此只要具备初步的概率统计知识就能应用本文介绍的方法.

附 录

$L_4(2^3)$			
列号(i)	1	2	3
l_{ii}	4	4	4

$L_{12}(2^{11})$			
列号(i)	1	2	... 11
l_{ii}	12	12	... 12

$L_{32}(2^{31})$			
列号(i)	1	2	... 31
l_{ii}	32	32	... 32

$L_{21}(3^{13})$			
列号(i)	1	2	... 13
l_{ii}	18	18	... 18

$L_{16}(4^5)$			
列号(i)	1	2	... 5
l_{ii}	80/9	80/9	... 80/9

$L_{25}(5^6)$			
列号(i)	1	2	... 6
l_{ii}	25/2	25/2	... 25/2

$L_{49}(7^8)$			
列号(i)	1	2	... 8
l_{ii}	196/9	196/9	... 196/9

$L_{81}(9^{10})$			
列号(i)	1	2	... 10
l_{ii}	135/4	135/4	... 135/4

$L_{169}(13^{14})$			
列号(i)	1	2	... 14
l_{ii}	1183/18	1183/18	... 1183/18

$L_8(2^7)$			
列号(i)	1	2	... 7
l_{ii}	8	8	... 8

$L_{16}(2^{15})$			
列号(i)	1	2	... 15
l_{ii}	16	16	... 16

$L_9(3^4)$				
列号(i)	1	2	3	4
l_{ii}	6	6	6	6

$L_{81}(3^{40})$				
列号(i)	1	2	... 40	
l_{ii}	54	54	... 54	

$L_{64}(4^{21})$				
列号(i)	1	2	... 21	
l_{ii}	320/9	320/9	... 320/9	

$L_{125}(5^{31})$				
列号(i)	1	2	... 31	
l_{ii}	125/2	125/2	... 125/2	

$L_{64}(8^9)$				
列号(i)	1	2	... 9	
l_{ii}	192/7	192/7	... 192/7	

$L_{121}(11^{12})$				
列号(i)	1	2	... 12	
l_{ii}	242/5	242/5	... 242/5	

$$L_8(4^1 \times 2^4)$$

列号(i) 1 2 3 ... 5

l_{ii} 40/9 8 8 ... 8

$$L_{18}(3^7 \times 2^1)$$

列号(i) 1 2 ... 7 8

l_{ii} 12 12 ... 12 18

$$L_{54}(3^{25} \times 2^1)$$

列号(i) 1 2 ... 25 26

l_{ii} 36 36 ... 36 54

$$L_{16}(4^4 \times 2^3)$$

列号(i) 1 2 ... 4 5 6 7

l_{ii} 80/9 80/9 ... 80/9 16 16 16

$$L_{15}(4^3 \times 2^6)$$

列号(i) 1 2 3 4 5 ... 9

l_{ii} 80/9 80/9 80/9 16 16 ... 16

$$L_{15}(4^2 \times 2^9)$$

列号(i) 1 2 3 4 ... 11

l_{ii} 80/9 80/9 16 16 ... 16

$$L_{15}(4^1 \times 2^{12})$$

列号(i) 1 2 3 ... 13

l_{ii} 80/9 16 16 ... 16

$$L_{32}(4^9 \times 2^{14})$$

列号(i) 1 2 ... 9 10 11 ... 23

l_{ii} 160/9 160/9 ... 160/9 32 32 ... 32

$$L_{128}(4^{41} \times 2^4)$$

列号(i) 1 2 ... 41 42 43 ... 45

l_{ii} 640/9 640/9 ... 640/9 128 128 ... 128

$$L_{50}(5^{11} \times 2^1)$$

列号(i) 1 2 ... 11 12

l_{ii} 25 25 ... 25 50

$$L_{98}(7^{15} \times 2^1)$$

列号(i) 1 2 ... 15 16

l_{ii} 392/9 392/9 ... 392/9 98

$$L_{16}(8^1 \times 2^8)$$

列号(i) 1 2 3 ... 9

l_{ii} 48/7 16 16 ... 16

$$L_{18}(6^1 \times 3^6)$$

列号(i) 1 2 3 ... 7

l_{ii} 42/5 12 12 ... 12

参 考 文 献

- 1 费荣昌, 费定晖. 概率统计
- 2 中国现场统计研究会三次设计组等正交法和三次设计

Orthogonaie Design of Regression

Jiang Bingxin

(Dept of Found. Courses)

Abstract This paper deals with the simple and convenient method to solve linear regression equation with linear transformation and the characteristics of orthogonal surface. This method is suitable for orthogonal experiments of equidistance multi-factor of any levels and mixed levels. An example is also given in this paper.

Keywords orthogonal experiments, linear, regression equation