3.2 正交回归设计

王正明　易泰河

系统工程学院　军事建模与仿真系

2019 年 12 月 18 日

引言

例

利用一架精度为 σ 的天平称 4 个不同的物体, 能否给出一种精度高于 σ 且只需称 4 次的称重方案?

引言

- 称重方案:

序号	左侧	右侧	差异
1	1 2 3 4	空	y_1
2	1 2	3 4	y_2
3	1 3	2 4	y_3
4	1 4	2 3	y_4

- 则

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

- 矩阵形式: $y = X\beta + \varepsilon$, 广义设计矩阵 X 为正交矩阵!

引言

- 称重方案:

序号	左侧	右侧	差异
1	1 2 3 4	空	y_1
2	1 2	3 4	y_2
3	1 3	2 4	y_3
4	1 4	2 3	y_4

- 则

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

- 矩阵形式: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 广义设计矩阵 $\boldsymbol{X}$ 为正交矩阵!

引言

- 参数的最小二乘估计:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \frac{1}{4}\begin{bmatrix} y_1 + y_2 + y_3 + y_4 \\ y_1 + y_2 - y_3 - y_4 \\ y_1 - y_2 + y_3 - y_4 \\ y_1 - y_2 - y_3 + y_4 \end{bmatrix}$$

- $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \mathrm{diag}\left\{\frac{\sigma^2}{4}, \frac{\sigma^2}{4}, \frac{\sigma^2}{4}, \frac{\sigma^2}{4}\right\}$;
- 每一个物体的测量精度都提高了 4 倍!

知识回顾

- 线性模型 $\boldsymbol{y} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$ 的参数估计为

$$\begin{cases} \widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \\ \widehat{\sigma^2} = \dfrac{\mathrm{RSS}}{n-m} = \dfrac{\boldsymbol{y}^{\mathrm{T}}\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\right]\boldsymbol{y}}{n-m}. \end{cases}$$

- $\hat{\boldsymbol{\beta}}$ 服从正态分布, 且与 $\hat{\sigma^2}$ 独立;
- 线性模型的假设检验: t 检验和 F 检验.

引言

- 回归分析中, 不论是参数估计还是显著性检验, 都需要求矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的逆.
- 当 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 为对角矩阵时, 求逆十分简单.
- 回归的正交设计恰好可达到这一要求.

引言

- 各因子的量纲可能各不相同，变化范围可能极其悬殊，给试验设计带来一定的困难.
- 编码变换使诸因子的取值区域转化为中心在原点的“立方体”.
- 设第 i 个变量 $z_i \in [z_{1i}, z_{2i}]$, $i = 1, 2, \cdots, p$. 作变换:

$$x_i = \frac{(z_i - z_{1i}) - (z_{2i} - z_i)}{z_{2i} - z_{1i}},$$

则 $x_i \in [-1, 1]$. 试验区间 $\mathcal{X} = [-1, 1]^p$.

3.2 正交回归设计

3.2.1 正交回归设计的定义

(1) 定义与案例

(2) 利用二水平正交表构造正交回归设计

3.2.2 正交回归设计的统计分析

3.2.3 添加中心点的重复试验

定义

若线性模型 $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ 中, 设计 ξ_n 使得矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 为对角矩阵, 则称设计 ξ_n 为正交回归设计.

- 为简单起见, 只考虑模型 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$.
- 如果 $\xi_n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ 是上述回归模型的正交回归设计, 其设计矩阵为

$$\boldsymbol{D} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

则

$$\begin{cases} \sum_{j=1}^{n} x_{ji} = 0, & i = 1, 2, \cdots, p, \\ \sum_{j=1}^{n} x_{ji_1} x_{ji_2} = 0, & i_1 \neq i_2. \end{cases}$$

例

某种橡胶制品由橡胶、树脂和改良剂复合而成. 试验的目的是改良橡胶制品的性能, 即提高撕裂强度. 考察的变量有三个:

- 橡胶中成分 A 的百分比 z_1, 变化范围是 $[0, 20]$;
- 树脂中成分 B 的百分比 z_2, 变化范围是 $[10, 30]$;
- 改良剂的百分比 z_3, 范围为 $[0.1, 0.3]$.

- 首先利用编码变换

$$x_1 = \frac{z_1 - 10}{10}, \qquad x_2 = \frac{z_2 - 20}{10}, \qquad x_3 = \frac{z_3 - 0.2}{0.1}$$

把畸形的长方体试验区域变换成中心在原点的立方体.

- 考虑试验方案如下

原始变量			编码变量		
z_1	z_2	z_3	x_1	x_2	x_3
0	10	0.1	-1	-1	-1
0	10	0.3	-1	-1	$+1$
0	30	0.1	-1	$+1$	-1
0	30	0.3	-1	$+1$	$+1$
20	10	0.1	$+1$	-1	-1
20	10	0.3	$+1$	-1	$+1$
20	30	0.1	$+1$	$+1$	-1
20	30	0.3	$+1$	$+1$	$+1$

- 首先利用编码变换

$$x_1 = \frac{z_1 - 10}{10}, \qquad x_2 = \frac{z_2 - 20}{10}, \qquad x_3 = \frac{z_3 - 0.2}{0.1}$$

把畸形的长方体试验区域变换成中心在原点的立方体.

- 考虑试验方案如下

原始变量			编码变量		
z_1	z_2	z_3	x_1	x_2	x_3
0	10	0.1	-1	-1	-1
0	10	0.3	-1	-1	$+1$
0	30	0.1	-1	$+1$	-1
0	30	0.3	-1	$+1$	$+1$
20	10	0.1	$+1$	-1	-1
20	10	0.3	$+1$	-1	$+1$
20	30	0.1	$+1$	$+1$	-1
20	30	0.3	$+1$	$+1$	$+1$

- 对于一次回归模型

$$y = \beta_1 + \beta_1 x_2 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

该试验方案是一个正交回归设计.

- 对于包含乘积项的一次回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon,$$

该方案还是一个正交回归设计吗?

- 对于一次回归模型

$$y = \beta_1 + \beta_1 x_2 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

该试验方案是一个正交回归设计.

- 对于包含乘积项的一次回归模型

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon, \end{aligned}$$

该方案还是一个正交回归设计吗?

3.2 正交回归设计

3.2.1 正交回归设计的定义

(1) 定义与案例

(2) 利用二水平正交表构造正交回归设计

3.2.2 正交回归设计的统计分析

3.2.3 添加中心点的重复试验

- 由于只考虑因子的线性规律, 可采用二水平正交表来构造正交回归设计, 如

试验号	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
1	1	1	1	1	1	1	1
2	1	1	−1	1	−1	−1	−1
3	1	−1	1	−1	1	−1	−1
4	1	−1	−1	−1	−1	1	1
5	−1	1	1	−1	−1	1	−1
6	−1	1	−1	−1	1	−1	1
7	−1	−1	1	1	−1	−1	1
8	−1	−1	−1	1	1	1	−1

- 安排试验时, 只需要前三列即可.

3.2 正交回归设计

3.2.1 正交回归设计的定义

3.2.2 正交回归设计的统计分析

(1) 参数估计

(2) 假设检验

3.2.3 添加中心点的重复试验

- 设 m 个自变量的线性回归模型为

$$
\begin{cases}
y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1, \\
y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_1, \\
\qquad\qquad \cdots, \\
y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n,
\end{cases}
$$

- $\boldsymbol{D} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$ 为正交回归设计.

- 由设计的正交性, 信息矩阵为

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = (\mathbf{1}_n, \boldsymbol{D})^{\mathrm{T}}(\mathbf{1}_n, \boldsymbol{D}) = \operatorname{diag}\left(n, \sum_{j=1}^{n} x_{j1}^2, \cdots, \sum_{j=1}^{n} x_{jp}^2\right).$$

- 常数项矩阵为

$$\boldsymbol{b} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \begin{bmatrix} \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n} x_{j1} y_j \\ \vdots \\ \sum_{j=1}^{n} x_{jp} y_j \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}.$$

- 由设计的正交性, 信息矩阵为

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = (\mathbf{1}_n, \boldsymbol{D})^{\mathrm{T}}(\mathbf{1}_n, \boldsymbol{D}) = \mathrm{diag}\left(n, \sum_{j=1}^{n} x_{j1}^2, \cdots, \sum_{j=1}^{n} x_{jp}^2\right).$$

- 常数项矩阵为

$$\boldsymbol{b} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \begin{bmatrix} \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n} x_{j1} y_j \\ \vdots \\ \sum_{j=1}^{n} x_{jp} y_j \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}.$$

- 回归系数的最小二乘估计为

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\
&= \left[\frac{1}{n}\sum_{j=1}^{n} y_j, \frac{\sum_{j=1}^{n} x_{j1}y_j}{\sum_{j=1}^{n} x_{j1}^2}, \cdots, \frac{\sum_{j=1}^{n} x_{jp}y_j}{\sum_{j=1}^{n} x_{jp}^2}\right]^{\mathrm{T}}.
\end{aligned}
$$

- $\hat{\beta}$ 的分布是什么? 它的各分量独立吗?
- $\hat{\beta}$ 的第 j 个分量完全由 X 的第 j 列决定, 且 $\hat{\beta}$ 的各分量互相独立.

- 回归系数的最小二乘估计为

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

$$= \left[\frac{1}{n}\sum_{j=1}^{n} y_j, \frac{\sum_{j=1}^{n} x_{j1}y_j}{\sum_{j=1}^{n} x_{j1}^2}, \cdots, \frac{\sum_{j=1}^{n} x_{jp}y_j}{\sum_{j=1}^{n} x_{jp}^2}\right]^{\mathrm{T}}.$$

- $\hat{\beta}$ 的分布是什么? 它的各分量独立吗?
- $\hat{\beta}$ 的第 j 个分量完全由 $\boldsymbol{X}$ 的第 j 列决定, 且 $\hat{\beta}$ 的各分量互相独立.

- 回归系数的最小二乘估计为

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\
&= \left[\frac{1}{n}\sum_{j=1}^{n} y_j, \frac{\sum_{j=1}^{n} x_{j1}y_j}{\sum_{j=1}^{n} x_{j1}^2}, \cdots, \frac{\sum_{j=1}^{n} x_{jp}y_j}{\sum_{j=1}^{n} x_{jp}^2}\right]^{\mathrm{T}}.
\end{aligned}
$$

- $\hat{\boldsymbol{\beta}}$ 的分布是什么? 它的各分量独立吗?
- $\hat{\boldsymbol{\beta}}$ 的第 j 个分量完全由 $\boldsymbol{X}$ 的第 j 列决定, 且 $\hat{\boldsymbol{\beta}}$ 的各分量互相独立.

3.2 正交回归设计

3.2.1 正交回归设计的定义

3.2.2 正交回归设计的统计分析

(1) 参数估计

(2) 假设检验

3.2.3 添加中心点的重复试验

- 对 $j = 1, 2, \cdots, m$, 考虑检验问题

$$H_{0j} : \beta_j = 0, \quad \text{v.s.} \quad H_{ij} : \beta_j \neq 0.$$

- 回归系数的 t 检验?
- 回归系数的 F 检验?

定理 (回归系数的分组检验)

当假设 $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$ 成立时, RSS 与 $\text{RSS}_{H_0} - \text{RSS}$ 相互独立, 且

$$\frac{\text{RSS}_{H_0} - \text{RSS}}{\sigma^2} \sim \chi^2(r), \quad \frac{(\text{RSS}_{H_0} - \text{RSS})/r}{\text{RSS}/(n-m)} \sim F(r, n-m).$$

推论 当假设 $H_{0j} : \beta_j = 0$ 成立时, RSS 与 $\text{RSS}_{H_{0j}} - \text{RSS}$ 相互独立, 且

$$\frac{\text{RSS}_{H_{0j}} - \text{RSS}}{\text{RSS}/(n-p-1)} \sim F(1, n-p-1).$$

- 问题转化为计算 $\text{RSS}_{H_{0j}} - \text{RSS}$.

定理 (回归系数的分组检验)

当假设 $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$ 成立时, RSS 与 $\mathrm{RSS}_{H_0} - \mathrm{RSS}$ 相互独立, 且

$$\frac{\mathrm{RSS}_{H_0} - \mathrm{RSS}}{\sigma^2} \sim \chi^2(r), \qquad \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/r}{\mathrm{RSS}/(n-m)} \sim F(r, n-m).$$

推论 当假设 $H_{0j} : \beta_j = 0$ 成立时, RSS 与 $\mathrm{RSS}_{H_{0j}} - \mathrm{RSS}$ 相互独立, 且

$$\frac{\mathrm{RSS}_{H_{0j}} - \mathrm{RSS}}{\mathrm{RSS}/(n-p-1)} \sim F(1, n-p-1).$$

- 问题转化为计算 $\mathrm{RSS}_{H_{0j}} - \mathrm{RSS}$.

- 残差平方和 $\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}\right)^2$.
- 由于 $\hat{\boldsymbol{\beta}}$ 的各分量仅与 $\boldsymbol{X}$ 的对应列有关, 剔除第 j 列后得到的残差平方和为

$$\mathrm{RSS}_{H_{0j}} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip} + \hat{\beta}_j x_{ij}\right)^2 .$$

- 利用矩阵的正交性, 得到

$$\mathrm{RSS}_{H_{0j}} - \mathrm{RSS} = \hat{\beta}_j b_j$$

仅与第 j 个分量有关!

- 定义 $Q_j := \hat{\beta}_j b_j$ 为第 j 个变量的平方和.
- 回归平方和 SS_R 是在假设 $H_0 : (\beta_1, \beta_2, \cdots, \beta_p) = \mathbf{0}$ 成立下的 RSS_{H_0} 与 RSS 之间的差. 根据正交性,

$$SS_R = \mathrm{RSS}_{H_0} - \mathrm{RSS} = \sum_{j=1}^{m} Q_j.$$

- 正交回归设计中回归平方和可分解为各变量的平方和的和.
- 参数 $\boldsymbol{\beta}$ 的维数为 $p+1$, 残差平方和 RSS 的自由度为 $n-p-1$.

来源	平方和	自由度	均方和	F 值
x_1	$Q_1 = \frac{\left(\sum_{i=1}^n x_{i1} y_i\right)^2}{\sum_{i=1}^n x_{i1}^2}$	1	Q_1	$\frac{Q_1}{SS_E/(n-p-1)}$
x_2	$Q_2 = \frac{\left(\sum_{i=1}^n x_{i2} y_i\right)^2}{\sum_{i=1}^n x_{i2}^2}$	1	Q_2	$\frac{Q_2}{SS_E/(n-p-1)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$Q_p = \frac{\left(\sum_{i=1}^n x_{ip} y_i\right)^2}{\sum_{i=1}^n x_{ip}^2}$	1	Q_p	$\frac{Q_p}{SS_E/(n-p-1)}$
回归	$SS_R = \sum_{j=1}^p Q_j$	p	SS_R/p	$\frac{SS_R/p}{SS_E/(n-p-1)}$
剩余	$SS_E = SS_T - SS_R$	$n-p-1$	$SS_E/(n-p-1)$	
总	$SS_T = \sum_{i=1}^n y_i^2 - \frac{1}{n}\left(\sum_{i=1}^n y_i\right)^2$	$n-1$		

- 当显著性检验的结果出现某些回归系数不显著时，可从回归方程中直接剔除相应的项，而无需重新计算回归方程.

例

某化工产品的产量 y 受反应时间 z_1 和反应温度 z_2 的影响，现有条件 $z_1 = 35\texttt{min}$，$z_2 = 155$℃. 为寻找最优生产条件，在现有试验范围：$z_1 \in [30, 40]$，$z_2 \in [150, 160]$ 内设计一个试验，拟合一次回归模型

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon.$$

- 先通过编码变换把原始变量 z_1 和 z_2 变换为

$$x_1 = \frac{z_1 - 35}{5} \in [-1, 1], \qquad x_2 = \frac{z_2 - 155}{5} \in [-1, 1].$$

- 采用表 $L_4(2^3)$ 设计试验方案, 得到试验数据如下:

试验号	x_0	x_1	x_2	y
1	1	1	1	39.4
2	1	1	-1	40.0
3	1	-1	1	40.9
4	1	-1	-1	41.5

- 计算得到

$$\boldsymbol{b} = \begin{bmatrix} \sum_{j=1}^{4} y_j \\ \sum_{j=1}^{4} x_{j1} y_j \\ \sum_{j=1}^{4} x_{j2} y_j \end{bmatrix} = \begin{bmatrix} 161.7 \\ 3.1 \\ 1.3 \end{bmatrix},$$

- 回归系数的最小二乘估计为

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left[\frac{1}{4} \sum_{j=1}^{4} y_j, \frac{\sum_{j=1}^{4} x_{j1} y_j}{\sum_{j=1}^{4} x_{j1}^2}, \frac{\sum_{j=1}^{4} x_{j2} y_j}{\sum_{j=1}^{4} x_{j2}^2} \right]^{\mathrm{T}} \\ &= [40.425, 0.775, 0.325]^{\mathrm{T}}, \end{aligned}$$

- 拟合的回归方程（关于编码后的变量）为

$$\hat{y} = 40.425 + 0.775 x_1 + 0.325 x_2.$$

- 计算诸平方和,

$$SS_T = \sum_{j=1}^{4} y_j^2 - \frac{1}{4}\left(\sum_{j=1}^{4} y_j\right)^2 = 2.8275,$$

$$Q_1 = \hat{\beta}_1 b_1 = 2.4025, \quad Q_2 = \hat{\beta}_2 b_2 = 0.4225,$$

$$SS_R = Q_1 + Q_2 = 2.825,$$

$$RSS = SS_T - Q_1 - Q_2 = 0.0025.$$

- 方差分析表:

来源	平方和	自由度	均方和	F 值
x_1	2.4025	1	2.4025	961
x_2	0.4225	1	0.4225	169
回归	2.8250	2	1.4125	565
剩余	0.0025	1	0.0025	
总	2.8275	3		

- 因为 $F_{0.005}(1,1) = 161$, $F_{0.005}(2,1) = 200$, 当显著性水平 $\alpha = 0.05$ 时, 回归方程与回归方程中的两个变量都是显著的.

3.2 正交回归设计

3.2.1 正交回归设计的定义

3.2.2 正交回归设计的统计分析

3.2.3 添加中心点的重复试验

(1) 参数估计

(2) 假设检验

- 正交回归设计的试验点都在试验区域的边界上.
- 回归方程显著只意味着在试验区域的边界上一次模型与实际情况相符.
- 为查看试验区域内部回归模型的拟合情况, 可在原点补充几次试验来检验.

设在原点重复 k 个试验，试验结果为 $y_{01}, \cdots, y_{0k}$. 结合先前做的 n 次试验，得到线性统计模型

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \\ y_{01} \\ \vdots \\ y_{0k} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \\ \varepsilon_{n+1} \\ \vdots \\ \varepsilon_{n+k} \end{bmatrix}$$

$$= \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

- 补充中心点试验后的设计依然是正交回归设计!
- $x_1, \cdots, x_p$ 的回归系数的估计值 $\hat{\beta}_1, \cdots, \hat{\beta}_p$ 和相应的回归平方和 $Q_1, \cdots, Q_p$ 都与未补充试验前的结果一致;
- 但总平方和 SS_T 和残差平方和 RSS 的值改变了.

- 称 $SS_{E_1} = \sum_{j=1}^{k}(y_{0j} - \bar{y}_0)^2$ 为误差平方和, 其中 $\bar{y}_0 = \frac{1}{k}\sum_{j=1}^{k} y_{0j}$.
- 称 $SS_{\mathrm{Lf}} = RSS - SS_{E_1}$ 为失拟平方和.
- 显然 $SS_T = SS_R + SS_{\mathrm{Lf}} + SS_{E_1}$.
- SS_T 的自由度 $f_T = n + k - 1$, SS_R 的自由度 $f_R = p$, SS_{E_1} 的自由度 $f_{E_1} = k - 1$, SS_{Lf} 的自由度 $f_{\mathrm{Lf}} = n + k - 1 - p - (k - 1) = n - p$.

- 根据模型的假定和正态随机变量的二次型理论, 可以证明在假设
$$H_0 : \mathbb{E}(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
成立的条件下,
$$\frac{SS_{\text{Lf}}}{\sigma^2} \sim \chi^2(n-p), \qquad \frac{SS_{E_1}}{\sigma^2} \sim \chi^2(k-1),$$
且它们相互独立.
- 以统计量
$$F = \frac{SS_{\text{Lf}}/f_{\text{Lf}}}{SS_{E_1}/f_{E_1}} \sim F(f_{\text{Lf}}, f_{E_1})$$
来检验线性模型是否恰当. 当 $F < F_\alpha(f_{\text{Lf}}, f_{E_1})$ 时, 认为在显著水平为 α 时, 拟合线性回归模型是恰当的, 否则不恰当.

例

在前一例中试验的基础上, 在原点补充 5 次试验, 试验方案和结果如下:

试验号	x_0	x_1	x_2	x_1^2	x_2^2	y
1	1	1	1	1	1	39.4
2	1	1	−1	1	1	40.0
3	1	−1	1	1	1	40.9
4	1	−1	−1	1	1	41.5
5	1	0	0	0	0	40.3
6	1	0	0	0	0	40.5
7	1	0	0	0	0	40.7
8	1	0	0	0	0	40.2
9	1	0	0	0	0	40.6

拟合一次回归方程, 并检验回归方程的好坏.

- $\beta_0 = \bar{y}_{\cdot} = 40.444$, 拟合的回归方程为

$$\hat{y} = 40.444 + 0.775x_1 + 0.325x_2.$$

- Q_1、Q_2 和 SS_R 在前面已经计算过了,

$$SS_T = \sum_{j=1}^{9} y_j^2 - \frac{1}{4}\left(\sum_{j=1}^{9} y_j\right)^2 = 3.0022,$$

$$RSS = SS_T - Q_1 - Q_2 = 0.1772,$$

$$SS_{E_1} = \sum_{j=1}^{k} (y_{0j} - \bar{y}_0)^2 = 0.1720,$$

$$SS_{\mathrm{Lf}} = RSS - SS_{E_1} = 0.0052.$$

来源	平方和	自由度	均方和	F 值
回归	2.8250	2	1.4125	47.88
剩余	0.1772	6	0.0295	
失拟	0.0052	2	0.0026	$0.06 < 1$
误差	0.1720	4	0.0430	
总	3.0022	8		

- 当显著性水平 $\alpha = 0.01$ 时, 回归方程是显著的.

总结

- 正交回归设计使得矩阵 $X^{\mathrm{T}}X$ 为对角矩阵;
- 正交回归设计使得回归系数的估计计算简单, 且各分量互相独立;
- 正交回归设计使得回归平方和可分解为各变量的平方和的和.
- 如果把多因子试验的固定效应模型写成回归模型, 那么正交设计也是正交回归设计!