

高等工程數學 (13)



廈門大學
XIAMEN UNIVERSITY



信息學院
(国家示范性软件学院)
School of Informatics

黃 烽
博士·副教授
Dr. Wei Huang

线性统计推断

数理统计 (13)



厦门大学
XIAMEN UNIVERSITY



信息学院
(国家示范性软件学院)
School of Informatics
博士·副教授
Dr. Wei Huang





4 线性统计推断

在科学的研究和处理实际问题时,分析数据常常遇到这样一类问题,即要考察若干个因素对我们所关心的某个(或某些)指标的影响.例如,要研究钢中含碳量与精炼时间之间的联系,人的身高与体重之间的联系,等等.这些问题的变量之间有一定的联系,但又不能用确定的函数关系式来表达,因为这些变量实际上是随机变量,或至少其中一个是随机变量,它们之间的这种关系称为相关关系,回归分析就是研究这种相关关系的统计方法.再则,实际问题中影响某个(或某些)指标的因素很多,而这些因素由于试验条件的限制,往往只可以取有限种状态或只能定性地描述,处理这类问题的一种有效的统计方法是方差分析.举例来说,在化工生产中,原料成分、原料剂量、催化剂、反应温度与压力、反应时间、设备装置以及操作人员素质等因素都会对产品的质量与数量产生影响.其中有的因素影响大些,有的影响小些,为了能保证优质、高产、低能耗,就需要找出对产品质量与产量有显著影响的那些因素,并研究其最优工艺条件.为此需要做科学试验,取得一系列试验数据.由于人力、物力、财力等限制,做试验时对这些因素只能取有限种状态(也称为水平),并且希望试验次数尽可能少,而又要能达到试验的目的.方差分析讨论如何充分利用试验数据进行分析、推断某个因素的影响是否显著,以及在最优工艺条件下应如何选用显著性因素.至于如何安排试验,既大大降低试验次数又基本上达到试验的目的,则是正交试验设计研究的问题.

本章只讨论因素对所关心的指标的影响为线性时的情形,因此称为线性统计推断.

4.1 线性统计模型

如前所述,在实际问题中经常遇到多个变量处于同一个过程之中,它们互相联系、互相制约,并且所要考察的那些变量之间虽有联系但无确定的函数关系.有时即使理论上存在某种函数关系,但由于具体观测时不可避免地带有误差,因此它们之间的关系仍呈现出不确定性.在数理统计学中,把变量之间的这种不确定的函数关系称为相关关系,变量之间呈现相关关系的原因是变量本身具有随机性.

在相关关系的各个变量中,有的变量是普通的实变量,而有的变量是随机变量.因此研究相关关系时,一般可以分为随机变量与随机变量之间的相关关系,以及随机变量与普通变量之间的相关关系.这两种情况的假设不同,推导过程也不相同,但某些方法和结论却有类似之处.我们只讨论后一种情况.

从一个例子出发.小麦的亩产量记为 Y ,它与水(x_1)、肥料(x_2)、土质(x_3)、麦种(x_4)、栽培技术(x_5)及管理措施(x_6)等因素有关.这就是说,小麦亩产量 Y 与 x_1, x_2, \dots, x_6 等有一定的联系.但是由于观测或试验中总存在随机因素的影响,即使 x_1, x_2, \dots, x_6 相对固定,小麦的亩产量也不完全相同,因此将 Y 与 x_1, x_2, \dots, x_6 的相关关系分为两部分来研究,即有

$$Y = f(x_1, x_2, \dots, x_6) + \varepsilon, \quad (4.1-1)$$

4.1 线性统计模型

其中 $f(x_1, x_2, \dots, x_6)$ 表示 6 个可控因素 x_1, x_2, \dots, x_6 与亩产量 Y 的确定关系, f 是确定性函数, 它表示非随机部分, 而 ϵ 表示随机因素对亩产量 Y 的影响, 一般把 ϵ 看成数学期望 $E(\epsilon)=0$ 的随机变量. 于是, Y 是一个随机变量, 它是可观测的, 其数学期望 $E(Y)=f(x_1, x_2, \dots, x_6)$. 一般情况下, $f(x_1, x_2, \dots, x_6)$ 不一定是 x_1, x_2, \dots, x_6 的线性函数, 但为数学处理方便起见, 可以近似地把它当作线性函数, 从而设为

$$f(x_1, x_2, \dots, x_6) = \beta_0 + \beta_1 x_1 + \dots + \beta_6 x_6. \quad (4.1-2)$$

于是,

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_6 x_6 + \epsilon. \quad (4.1-3)$$

这样, (4.1-3)式是 Y 关于因子 x_1, x_2, \dots, x_6 的线性函数, 其中 $\beta_0, \beta_1, \dots, \beta_6$ 是未知参数. 在数理统计中, 它们是需要推断的对象. 对 $(x_1, x_2, \dots, x_6; Y)$ 作 n 次观测或试验, 便得到数据

$$(x_{i1}, x_{i2}, \dots, x_{i6}; y_i), \quad i = 1, 2, \dots, n.$$

我们要由这些数据来推断未知参数 $\beta_0, \beta_1, \dots, \beta_6$.

值得指出的是, 本例中有些因子是非数量的, 例如小麦种类是非数量的因子. 当它的品种有甲、乙、丙三种时, 可以规定 x_4 (甲)=1, x_4 (乙)=2, x_4 (丙)=3, 从而数字化.

一般的线性统计模型如下.

设因变量 Y 与自变量 x_1, x_2, \dots, x_k 之间有下述线性关系:

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \epsilon, \quad (4.1-4)$$



4.1 线性统计模型

对(4.1-4)式作 n 次观测, 得到数据 $(x_{i1}, x_{i2}, \dots, x_{ik}; y_i), i=1, 2, \dots, n$, 其中 y_1, y_2, \dots, y_n 分别是 Y 的 n 次观测值. 若记 (Y_1, Y_2, \dots, Y_n) 是取自总体 Y 的一个容量为 n 的样本, 则 (y_1, y_2, \dots, y_n) 是样本 (Y_1, Y_2, \dots, Y_n) 的观测值, 且由(4.1-4)式得

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (4.1-5)$$

(4.1-4)式关于未知参数 $\beta_0, \beta_1, \dots, \beta_k$ 是线性的, 这是线性统计模型的本质特征.

应用向量和矩阵形式, 记

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix},$$

那么, (4.1-5)式可表示为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4.1-6)$$

其中 \mathbf{X} 是已知的 $n \times (k+1)$ 常数矩阵, $\boldsymbol{\beta}$ 是 $k+1$ 维的未知参数向量, $\boldsymbol{\varepsilon}$ 是数学期望为零的 n 维随机向量.

对 $\boldsymbol{\varepsilon}$ 作如下的假定:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n, \quad (4.1-7)$$

其中 σ^2 是未知参数, \mathbf{I}_n 是 n 阶单位矩阵, 即

$$E(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2; \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 (i \neq j; i, j = 1, 2, \dots, n). \quad (4.1-8)$$



4.1 线性统计模型

这就是说,对随机误差 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 作无偏性、等方差性和不相关性的假定.

通常所说的线性统计模型是指由(4.1-6)式与(4.1-7)式所构成的模型,简记为(\mathbf{Y} 、 $\mathbf{X}\beta$ 、 $\sigma^2 \mathbf{I}_n$).

如果再进一步假定 ε 服从 n 维正态分布,则称这个模型为正态线性模型.

由(4.1-6)式和(4.1-7)式容易得到

$$E(\mathbf{Y}) = \mathbf{X}\beta, \quad \text{Cov}(\mathbf{Y}, \mathbf{Y}) = \sigma^2 \mathbf{I}_n, \quad (4.1-9)$$

即 $E(Y_i) = \sum_{l=0}^k x_{il}\beta_l, \text{cov}(Y_i, Y_j) = \sigma^2 \delta_{ij}, \quad i, j = 1, 2, \dots, n,$ (4.1-10)

其中 $x_{i0}=1, \quad i=1, 2, \dots, n, \quad \delta_{ij} = \begin{cases} 1, & i=j, \\ 0, & i \neq j. \end{cases}$

对于由(4.1-6)式和(4.1-7)式构成的线性模型($\mathbf{Y}, \mathbf{X}\beta, \sigma^2 \mathbf{I}_n$),所要讨论的统计推断问题是:

- 1) 对未知参数向量 β 和未知参数 σ^2 进行估计;
- 2) 对 \mathbf{Y} 服从线性模型的假设和有关 β 的某些假设进行检验;
- 3) 对 \mathbf{Y} 进行预测.

在今后的讨论中,我们总是假定 $n > k$,且 $k+1$ 阶方阵 $\mathbf{L} = \mathbf{X}^T \mathbf{X}$ 是可逆矩阵.



4.2 最小二乘估计及其性质

本节主要讨论线性统计模型 $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ 的未知参数向量 $\boldsymbol{\beta}$ 和未知参数 σ^2 的点估计问题。对于线性统计模型来说,由于 \mathbf{Y} 的分布并未给出,所以要用最小二乘估计法求点估计;但若 \mathbf{Y} 是 n 维正态随机向量(即 $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ 为正态线性模型),那么也可以用极大似然估计法求出点估计。

这里着重研究最小二乘估计的性质。

$$\text{令 } Q(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n \left[y_i - \sum_{j=0}^k \beta_j x_{ij} \right]^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (4.2-1)$$

其中 $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$ 为 \mathbf{Y} 的观测值。如果把 ϵ_i 理解为第 i 次观测中的“误差”,那么 Q 是 n 次观测中误差平方的和,故也称为误差平方和。

定义 4.2-1 如果 $\hat{\beta}_j = \hat{\beta}_j(y_1, y_2, \dots, y_n)$ ($j = 0, 1, \dots, k$) 满足

$$Q(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) = \min_{\beta_0, \beta_1, \dots, \beta_k} Q(\beta_0, \beta_1, \dots, \beta_k),$$

那么称 $\hat{\beta}_j$ 为 β_j 的最小二乘估计值,称相应的 $\hat{\beta}_j(Y_1, Y_2, \dots, Y_n)$ 为 β_j 的最小二乘估计量, $j = 0, 1, \dots, k$, 仍简记为 $\hat{\beta}_j$ 。

记 $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k]^\top$, 则 $\hat{\boldsymbol{\beta}}$ 满足正规方程组

$$(\mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}, \quad (4.2-2)$$

即

$$\mathbf{L} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}. \quad (4.2-2')$$

由于 \mathbf{L} 是可逆的,所以(4.2-2')式有唯一解,从而得到 $\boldsymbol{\beta}$ 的最小二乘估计量为

$$\hat{\boldsymbol{\beta}} = \mathbf{L}^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (4.2-3)$$



4.2 最小二乘估计及其性质

求得 β 的最小二乘估计值 $\hat{\beta}$ 后, 令

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k, \quad (4.2-4)$$

那么, 当 x_1, \dots, x_k 分别取值 x_{11}, \dots, x_{1k} 时, 则由(4.2-4)式得到 $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \cdots + \hat{\beta}_k x_{1k}$, 它可以作为随机变量 $Y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \epsilon_1$ 的一个估计值. 再则, 若 $\hat{\beta}_j$ ($0 \leq j \leq k$) 是 β_j 的最小二乘估计量, 则(4.2-4)式的 \hat{y} 是统计量

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \quad (4.2-4')$$

的观测值.

由(4.2-4)式所得到的函数

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

称为经验(线性)回归函数.

下面在对 ϵ 作(4.1-7)式假定下, 讨论由(4.2-3)式确定的最小二乘估计量 $\hat{\beta}$ 的一些基本性质.

性质 1 $\hat{\beta}$ 是 β 的线性无偏估计量, 且

$$\text{Cov}(\hat{\beta}, \hat{\beta}) = \sigma^2 L^{-1}. \quad (4.2-5)$$

证 由于 $\hat{\beta} = L^{-1} X^T Y$, 所以 $\hat{\beta}_j$ ($j = 0, 1, \dots, k$) 是样本 (Y_1, Y_2, \dots, Y_n) 的线性函数, 这种估计称为线性估计, 故 $\hat{\beta}_j$ 是线性估计, 从而 $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k]^T$ 是 β 的线性估计量. 又因

$$E(\hat{\beta}) = E(L^{-1} X^T Y) = L^{-1} X^T E(Y) = L^{-1} X^T X \beta = L^{-1} L \beta = \beta,$$

故 $\hat{\beta}$ 是 β 的线性无偏估计量.

4.2 最小二乘估计及其性质

记 $\mathbf{G} = \mathbf{L}^{-1}\mathbf{X}^T$, 则 $\hat{\boldsymbol{\beta}} = \mathbf{G}\mathbf{Y}$. 于是

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) &= \text{Cov}(\mathbf{G}\mathbf{Y}, \mathbf{G}\mathbf{Y}) = \mathbf{G} \cdot \text{Cov}(\mathbf{Y}, \mathbf{Y}) \cdot \mathbf{G}^T = \sigma^2 \mathbf{G} \mathbf{I}_n \mathbf{G}^T = \sigma^2 \mathbf{G} \mathbf{G}^T \\ &= \sigma^2 \mathbf{L}^{-1} \mathbf{X}^T (\mathbf{L}^{-1} \mathbf{X}^T)^T = \sigma^2 \mathbf{L}^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{L}^{-1})^T = \sigma^2 \mathbf{L}^{-1}.\end{aligned}$$

■

定义 4.2-2 记 $\hat{\mathbf{Y}} = [\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n]^T$, 则由(4.2-4')式知, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, 称 n 维随机向量

$$\mathbf{e} = [e_1, e_2, \dots, e_n]^T = \mathbf{Y} - \hat{\mathbf{Y}} \quad (4.2-6)$$

为残差向量.

性质 2 对于残差向量 \mathbf{e} , 有

$$\left. \begin{array}{l} (1) E(\mathbf{e}) = \mathbf{0}; \\ (2) \text{Cov}(\mathbf{e}, \mathbf{e}) = \sigma^2 (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T); \\ (3) \text{Cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) = \mathbf{0}. \end{array} \right\} \quad (4.2-7)$$

证 (1) $E(\mathbf{e}) = E(\mathbf{Y} - \hat{\mathbf{Y}}) = E(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = E(\mathbf{Y}) - \mathbf{X}E(\hat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$.

$$\begin{aligned}(2) \text{Cov}(\mathbf{e}, \mathbf{e}) &= \text{Cov}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \text{Cov}((\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T)\mathbf{Y}, (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T)\mathbf{Y}) \\ &= (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T) \text{Cov}(\mathbf{Y}, \mathbf{Y}) (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T)^T \\ &= \sigma^2 (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T) \mathbf{I}_n (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T)^T \\ &= \sigma^2 (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T) (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T) = \sigma^2 (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T).\end{aligned}$$

$$\begin{aligned}(3) \text{Cov}(\hat{\boldsymbol{\beta}}, \mathbf{e}) &= \text{Cov}(\mathbf{L}^{-1}\mathbf{X}^T\mathbf{Y}, (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T)\mathbf{Y}) = \mathbf{L}^{-1}\mathbf{X}^T \text{Cov}(\mathbf{Y}, \mathbf{Y}) (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T)^T \\ &= \sigma^2 \mathbf{L}^{-1}\mathbf{X}^T (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T) = \sigma^2 (\mathbf{L}^{-1}\mathbf{X}^T - \mathbf{L}^{-1}\mathbf{X}^T) = \mathbf{0}.\end{aligned}$$

■

4.2 最小二乘估计及其性质

性质 3 记

$$Q_e = e^T e, \quad (4.2-8)$$

则有

$$E(Q_e) = (n - k - 1)\sigma^2, \quad (4.2-9)$$

从而

$$E\left(\frac{Q_e}{n - k - 1}\right) = \sigma^2, \quad (4.2-9')$$

因此 $\frac{Q_e}{n - k - 1}$ 是 σ^2 的无偏估计量.

证 由于 $E(e) = \mathbf{0}$, 所以

$$Q_e = e^T e = (e - E(e))^T (e - E(e)) = \text{tr}[(e - E(e))(e - E(e))^T].$$

$$\begin{aligned} \text{于是, } E(Q_e) &= E\{\text{tr}[(e - E(e))(e - E(e))^T]\} = \text{tr}[E([e - E(e)][e - E(e)]^T)] \\ &= \text{tr}[\text{Cov}(e, e)] = \sigma^2 \text{tr}[I_n - \mathbf{X}L^{-1}\mathbf{X}^T] = \sigma^2 [\text{tr}I_n - \text{tr}(\mathbf{X}L^{-1}\mathbf{X}^T)] \\ &= \sigma^2 [\text{tr}I_n - \text{tr}(\mathbf{L}^{-1}\mathbf{X}^T\mathbf{X})] = \sigma^2 (\text{tr}I_n - \text{tr}I_{k+1}) = (n - k - 1)\sigma^2. \end{aligned}$$

记

$$\hat{\sigma}_e^2 = \frac{Q_e}{n - k - 1}, \quad (4.2-10)$$

则由(4.2-9')式知, $\hat{\sigma}_e^2$ 是未知参数 σ^2 的无偏估计量. ■

性质 4 设 $c^T \beta$ 是待估函数, 其中 $c = [c_0, c_1, \dots, c_k]^T$ 是任一已知的常数向量, 则 $c^T \hat{\beta}$ 是 $c^T \beta$ 的最小方差线性无偏估计量. 在这个意义上, 称 $\hat{\beta}$ 是 β 的最小方差线性无偏估计.

证 由于 $\hat{\beta}$ 是 β 的线性无偏估计, 所以 $c^T \hat{\beta}$ 是 $c^T \beta$ 的线性无偏估计量, 并且有

$$\begin{aligned} D(c^T \hat{\beta}) &= E([c^T(\hat{\beta} - E(\hat{\beta}))]^2) = E([c^T(\hat{\beta} - E(\hat{\beta}))][c^T(\hat{\beta} - E(\hat{\beta}))]^T) \\ &= c^T E((\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))^T)c = c^T \text{Cov}(\hat{\beta}, \hat{\beta})c. \end{aligned} \quad (4.2-11)$$

4.2 最小二乘估计及其性质

下面证明 $\mathbf{c}^T \hat{\boldsymbol{\beta}}$ 的方差最小性. 设 $\mathbf{a}^T \mathbf{Y}$ 是 $\mathbf{c}^T \boldsymbol{\beta}$ 的任一线性无偏估计量, 则由无偏性的要求知, 等式

$$E(\mathbf{a}^T \mathbf{Y}) = \mathbf{a}^T E(\mathbf{Y}) = \mathbf{a}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{c}^T \boldsymbol{\beta}$$

对一切 $\boldsymbol{\beta}$ 都应成立, 故有

$$\mathbf{a}^T \mathbf{X} = \mathbf{c}^T. \quad (4.2-12)$$

类似于(4.2-11)式, 可得

$$D(\mathbf{a}^T \mathbf{Y}) = \mathbf{a}^T \text{Cov}(\mathbf{Y}, \mathbf{Y}) \mathbf{a} = \sigma^2 \mathbf{a}^T \mathbf{a}. \quad (4.2-13)$$

于是, 由(4.2-13)式、(4.2-5)式和(4.2-12)式得

$$D(\mathbf{a}^T \mathbf{Y}) - D(\mathbf{c}^T \hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{a}^T \mathbf{a} - \mathbf{c}^T \mathbf{L}^{-1} \mathbf{c}) = \sigma^2 (\mathbf{a}^T \mathbf{a} - \mathbf{a}^T \mathbf{X} \mathbf{L}^{-1} \mathbf{X}^T \mathbf{a}) = \sigma^2 \mathbf{a}^T (\mathbf{I}_n - \mathbf{X} \mathbf{L}^{-1} \mathbf{X}^T) \mathbf{a}.$$

从而由(4.2-7)式的第二式可知,

$$D(\mathbf{a}^T \mathbf{Y}) - D(\mathbf{c}^T \hat{\boldsymbol{\beta}}) = \mathbf{a}^T \text{Cov}(\mathbf{e}, \mathbf{e}) \mathbf{a} \geq 0.$$

这就证明了 $\mathbf{c}^T \hat{\boldsymbol{\beta}}$ 是 $\mathbf{c}^T \boldsymbol{\beta}$ 的最小方差线性无偏估计量. ■

特别地, 若取 $\mathbf{c} = [0, \dots, 0, \overset{j}{1}, 0, \dots, 0]^T$, 则得 $\hat{\beta}_j (j=0, 1, \dots, k)$ 是 β_j 的最小方差线性无偏估计量.

现在假定在(4.1-7)式的基础上, 再进一步假定 $\epsilon_i (i=1, 2, \dots, n)$ 服从正态分布 $N(0; \sigma^2)$, 即所讨论的模型(4.1-6)式是正态线性模型. 我们研究最小二乘估计量 $\hat{\boldsymbol{\beta}}$ 的分布和统计量 Q_e 的分布问题.



4.2 最小二乘估计及其性质

性质 5 若(4.1-6)式是正态线性模型,则

- 1) $\hat{\beta}$ 与 e 相互独立,从而 $\hat{\beta}$ 与 Q_e 相互独立;
- 2) $\hat{\beta}$ 服从 $k+1$ 维正态分布,它的均值向量为 β ,协方差矩阵是 $\sigma^2 L^{-1}$,即

$$\hat{\beta} \sim N(\beta; \sigma^2 L^{-1});$$

- 3) e 服从 n 维正态分布,它的均值向量为 0 ,协方差矩阵是 $\sigma^2(I_n - XL^{-1}X^T)$,即

$$e \sim N(0; \sigma^2(I_n - XL^{-1}X^T));$$

- 4) $\frac{Q_e}{\sigma^2}$ 服从自由度为 $n-k-1$ 的 χ^2 分布,即

$$\frac{Q_e}{\sigma^2} \sim \chi^2(n-k-1).$$

证 由于 Y 是 n 维正态随机向量,且 $\hat{\beta} = L^{-1}X^TY$, $e = Y - \hat{Y} = (I_n - XL^{-1}X^T)Y$,所以由正态随机变量经线性变换后仍为正态随机变量的定理知, $\hat{\beta}$ 和 e 分别服从 $k+1$ 维和 n 维正态分布. 又由上述的性质 1 和性质 2 便得(2)和(3). 再由性质 2 的(3)知 $\text{Cov}(\hat{\beta}, e) = 0$,因而 $\hat{\beta}$ 与 e 相互独立,这就证明了(1).

最后证(4). 令 $A = I_n - XL^{-1}X^T$, 则有

$$A^T = A, \quad A^2 = A, \quad \text{tr}A = n - k - 1.$$

从而 A 是对称等幂矩阵,故由矩阵论知, A 的秩为 $n-k-1$,且存在 n 阶正交矩阵 U 使

$$A = U^T \begin{bmatrix} I_{n-k-1} & O \\ O & O \end{bmatrix} U.$$



4.2 最小二乘估计及其性质

现在, $e = Y - \hat{Y} = AY$, 因此

$$\frac{Q_e}{\sigma^2} = \frac{(AY)^T(AY)}{\sigma^2} = \frac{Y^T A^T A Y}{\sigma^2} = \frac{Y^T A Y}{\sigma^2}.$$

另一方面, 由于 $X^T A = X^T (I_n - XL^{-1}X^T) = O, AX = O$, 故有

$$e^T A e = (Y - X\beta)^T A (Y - X\beta) = Y^T A Y - \beta^T X^T A Y - Y^T A X \beta + \beta^T X^T A X \beta = Y^T A Y.$$

于是,

$$\frac{Q_e}{\sigma^2} = \frac{e^T A e}{\sigma^2} = \left(U \frac{e}{\sigma} \right)^T \begin{bmatrix} I_{n-k-1} & O \\ O & O \end{bmatrix} \left(U \frac{e}{\sigma} \right).$$

令

$$Z = [z_1, z_2, \dots, z_{n-k-1}]^T = U \frac{e}{\sigma},$$

那么, 由于 $\frac{e}{\sigma}$ 的分量是相互独立的随机变量, 且每个分量都服从标准正态分布 $N(0; 1)$, U

又是正交矩阵, 所以 Z 的分量是相互独立的正态随机变量, 且都服从 $N(0; 1)$. 因此, 由 $\frac{Q_e}{\sigma^2}$

$= \sum_{i=1}^{n-k-1} z_i^2$ 知, $\frac{Q_e}{\sigma^2}$ 服从自由度为 $n - k - 1$ 的 χ^2 分布. ■

性质 6 若(4.1-6)式是正态线性模型, 则 $\hat{\beta}$ 和 $\hat{\sigma}^2 = \frac{Q_e}{n}$ 分别是 β 和 σ^2 的极大似然估计量.

4.2 最小二乘估计及其性质

证 由于 Y_1, Y_2, \dots, Y_n 是相互独立的随机变量, 且 $Y_i \sim N(\sum_{j=0}^k x_{ij}\beta_j; \sigma^2)$, $i = 1, 2, \dots, n$, 所以关于 β 和 σ^2 的似然函数为

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)},$$

从而 $\ln L(\beta, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$.

于是, β 和 σ^2 的极大似然估计值可以由方程组

$$\begin{cases} \nabla_{\beta} \ln L(\beta, \sigma^2) = \frac{1}{\sigma^2} X^T(y - X\beta) = \mathbf{0}, \\ \frac{\partial}{\partial \sigma^2} \ln L(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)^T(y - X\beta) = 0 \end{cases}$$

解得 $\hat{\beta} = (X^T X)^{-1} X^T y$, $\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$.

从而 $\hat{\beta} = (X^T X)^{-1} X^T Y = L^{-1} X^T Y$ 和 $\hat{\sigma}^2 = \frac{Q_e}{n}$ 分别是 β 和 σ^2 的极大似然估计量. ■

由于 $\hat{\sigma}^2 = \frac{Q_e}{n} = \frac{n-k-1}{n} \hat{\sigma}_e^2$, (4.2-14)

所以 $\hat{\sigma}^2$ 不是 σ^2 的无偏估计量, 但由于 $\lim_{n \rightarrow +\infty} \hat{\sigma}^2 = \hat{\sigma}_e^2$, 所以 $\hat{\sigma}^2$ 是 σ^2 的渐近无偏估计量.



4.2 最小二乘估计及其性质

例 1 在硝酸钠的溶解度试验中, 测得在不同温度 x (单位: $^{\circ}\text{C}$)下, 硝酸钠溶解于水中的溶解度 $y\%$ 的数据如下:

温 度	0	4	10	15	21	29	36	51	68
溶解度(%)	66.7	71.0	76.3	80.6	85.7	92.9	99.4	113.6	125.1

求 y 与 x 之间的经验回归函数.

解 在坐标平面上点出这些数据点, 可知其大致位于一条直线附近, 从而可用线性模型. 这时,

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 4 & 10 & 15 & 21 & 29 & 36 & 51 & 68 \end{bmatrix}^T,$$

$$\mathbf{y} = [66.7, 71.0, 76.3, 80.6, 85.7, 92.9, 99.4, 113.6, 125.1]^T,$$

将此代入(4.2-2)式并解之, 得

$$\hat{\beta}_0 = 67.508, \quad \hat{\beta}_1 = 0.871.$$

于是, 所求的经验回归函数为

$$y = 67.508 + 0.871x.$$

例 2 为了考察汽油的两种添加剂(记为 x_1, x_2)对汽车消耗一公升汽油所行驶的平均公里数 y 的影响, 今就 x_1, x_2 各取 0, 1 和 2 个单位所有可能组合情况进行测试, 得到的数据如表 4.2-1:

4.2 最小二乘估计及其性质

表 4.2-1

y	x_2	0	1	2
x_1	0	20.4	21.0	22.0
	1	19.0	22.3	24.5
	2	18.9	18.7	22.4

求 y 与 x_1, x_2 之间的经验回归函数.

解 由于

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \end{bmatrix}^T,$$

$$\mathbf{y} = [20.4 \ 19.0 \ 18.9 \ 21.0 \ 22.3 \ 18.7 \ 22.0 \ 24.5 \ 22.4]^T,$$

所以由(4.2-3)式解得 $\hat{\beta}_0 = 19.82, \hat{\beta}_1 = -0.57, \hat{\beta}_2 = 1.77$.

于是, 经验回归函数为 $y = 19.82 - 0.57x_1 + 1.77x_2$.

从这个回归函数看出, 就统计规律来说, 第一种添加剂可能是不利的, 而第二种添加剂可能是有利的.

在具体计算最小二乘估计值 $\hat{\beta}$ 和经验回归函数时, 常常采用中心化方法. 记



4.2 最小二乘估计及其性质

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i; \quad \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad j = 1, 2, \dots, k;$$

$$l_{ji} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m) = \sum_{i=1}^n x_{ij}x_{im} - n\bar{x}_j\bar{x}_m, \quad j, m = 1, 2, \dots, k;$$

$$l_{iy} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y}) = \sum_{i=1}^n x_{ij}y_i - n\bar{x}_j\bar{y}, \quad j = 1, 2, \dots, k;$$

$$l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2;$$

$$\mathbf{L}_{k \times k} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1k} \\ l_{21} & l_{22} & \cdots & l_{2k} \\ \vdots & \vdots & & \vdots \\ l_{k1} & l_{k2} & \cdots & l_{kk} \end{bmatrix}, \mathbf{L}_{k \times k}^{-1} = \begin{bmatrix} l^{(11)} & l^{(12)} & \cdots & l^{(1k)} \\ l^{(21)} & l^{(22)} & \cdots & l^{(2k)} \\ \vdots & \vdots & & \vdots \\ l^{(k1)} & l^{(k2)} & \cdots & l^{(kk)} \end{bmatrix}.$$

则有

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^k \hat{\beta}_j \bar{x}_j, \tag{4.2-15}$$

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} l^{(11)} & l^{(12)} & \cdots & l^{(1k)} \\ l^{(21)} & l^{(22)} & \cdots & l^{(2k)} \\ \vdots & \vdots & & \vdots \\ l^{(k1)} & l^{(k2)} & \cdots & l^{(kk)} \end{bmatrix} \begin{bmatrix} l_{1y} \\ l_{2y} \\ \vdots \\ l_{ky} \end{bmatrix}, \tag{4.2-16}$$



4.2 最小二乘估计及其性质

经验回归函数为

$$y = \bar{y} + \hat{\beta}_1(x_1 - \bar{x}_1) + \hat{\beta}_2(x_2 - \bar{x}_2) + \cdots + \hat{\beta}_k(x_k - \bar{x}_k); \quad (4.2-17)$$

且统计量 Q_e 的观测值(仍记为 Q_e)为

$$Q_e = l_{yy} - \sum_{j=1}^k \hat{\beta}_j l_{jy}. \quad (4.2-18)$$

事实上,由于 $n \times (k+1)$ 矩阵

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},$$

所以正规方程组(4.2-2)为

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \cdots & \sum_{i=1}^n x_{i1} x_{ik} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik} x_{i1} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \\ \vdots \\ \sum_{i=1}^n x_{ik} y_i \end{bmatrix}. \quad (4.2-19)$$



4.2 最小二乘估计及其性质

应用矩阵的行初等变换,不难验证方程组(4.2-19)等价于下述方程组

$$\begin{bmatrix} 1 & \bar{x}_1 & \cdots & \bar{x}_k \\ 0 & l_{11} & \cdots & l_{1k} \\ \vdots & \vdots & & \vdots \\ 0 & l_{k1} & \cdots & l_{kk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \bar{y} \\ l_{1y} \\ \vdots \\ l_{ky} \end{bmatrix}. \quad (4.2-20)$$

容易看出,方程组(4.2-20)的解是(4.2-15)式和(4.2-16)式.

将经验回归函数

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

中的 $\hat{\beta}_0$ 用(4.2-15)式代入,即得(4.2-17)式.

再则, Q_e 的观测值为

$$\begin{aligned} & \sum_{i=1}^n \left[y_i - \bar{y} - \sum_{j=1}^k \hat{\beta}_j (x_{ij} - \bar{x}_j) \right]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2 \sum_{i=1}^n \left[(y_i - \bar{y}) \sum_{j=1}^k \hat{\beta}_j (x_{ij} - \bar{x}_j) \right] + \sum_{i=1}^n \left[\sum_{j=1}^k \hat{\beta}_j (x_{ij} - \bar{x}_j) \right]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2 \sum_{j=1}^k \hat{\beta}_j \left[\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y}) \right] + \sum_{j=1}^k \hat{\beta}_j \left(\sum_{m=1}^k \hat{\beta}_m l_{jm} \right) \\ &= l_{yy} - 2 \sum_{j=1}^k \hat{\beta}_j l_{jy} + \sum_{j=1}^k \hat{\beta}_j l_{jy} = l_{yy} - \sum_{j=1}^k \hat{\beta}_j l_{jy}, \end{aligned}$$

这就证明了(4.2-18)式.

4.2 最小二乘估计及其性质

例 3 在平炉炼钢过程中,由于矿石及炉气的氧化作用,铁水的总含碳量在不断降低. 一炉钢在冶炼初期(熔化期)总的去碳量(单位:吨)Y 与所加的两种矿石(天然矿石与烧结矿石)的量(单位:吨) x_1 、 x_2 及熔化时间(单位:5 分钟) x_3 有关. 现对某号平炉作了 49 次实测,得到的数据如表 4.2-2 所示.

表 4.2-2

i	y_i	x_{i1}	x_{i2}	x_{i3}	i	y_i	x_{i1}	x_{i2}	x_{i3}
1	4.3302	2	18	50	26	2.7066	9	6	39
2	3.6458	7	9	40	27	5.6314	12	5	51
3	4.4830	5	14	46	28	5.8152	6	13	41
4	5.5468	12	3	43	29	5.1302	12	7	47
5	5.4970	1	20	64	30	5.3910	0	24	61
6	3.1125	3	12	40	31	4.4583	5	12	37
7	5.1182	3	17	64	32	4.6569	4	15	49

4.2 最小二乘估计及其性质

续表

i	y_i	x_{i1}	x_{i2}	x_{i3}	i	y_i	x_{i1}	x_{i2}	x_{i3}
8	3.8759	6	5	39	33	4.5212	0	20	45
9	4.6700	7	8	37	34	4.8650	6	16	42
10	4.9536	0	23	55	35	5.3566	4	17	48
11	5.0060	3	16	60	36	4.6098	10	4	48
12	5.2701	0	18	49	37	2.3815	4	14	36
13	5.3772	8	4	50	38	3.8746	5	13	36
14	5.4849	6	14	51	39	4.5919	9	18	51
15	4.5960	0	21	51	40	5.1588	6	13	54
16	5.6645	3	14	51	41	5.4373	5	18	100
17	6.0795	7	12	56	42	3.9960	5	11	44
18	3.2194	16	0	48	43	4.3970	8	6	63
19	5.8076	6	16	45	44	4.0622	2	13	55
20	4.7306	0	15	52	45	2.2905	7	8	50
21	4.6805	9	0	40	46	4.7115	4	10	45
22	3.1272	4	6	32	47	4.5310	10	5	40
23	2.6104	0	17	47	48	5.3637	3	17	64
24	3.7174	9	0	44	49	6.0771	4	15	72
25	3.8946	2	16	39					

由经验知道, Y 与 x_1, x_2, x_3 之间的相关关系服从正态线性模型:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

其中 $\epsilon \sim N(0; \sigma^2)$, 求 $\beta_0, \beta_1, \beta_2, \beta_3$ 的最小二乘估计及 σ^2 的估计.

4.2 最小二乘估计及其性质

解 本例中, $n=49$, $k=3$. 由所给数据算得:

$$\bar{y} = \frac{1}{49} \sum_{i=1}^{49} y_i = 4.582, \quad \bar{x}_1 = \frac{1}{49} \sum_{i=1}^{49} x_{i1} = 5.286,$$

$$\bar{x}_2 = \frac{1}{49} \sum_{i=1}^{49} x_{i2} = 11.796, \quad \bar{x}_3 = \frac{1}{49} \sum_{i=1}^{49} x_{i3} = 49.204,$$

$$l_{11} = \sum_{i=1}^{49} x_{i1}^2 - 49\bar{x}_1^2 = 662.000,$$

$$l_{22} = \sum_{i=1}^{49} x_{i2}^2 - 49\bar{x}_2^2 = 1753.959,$$

$$l_{33} = \sum_{i=1}^{49} x_{i3}^2 - 49\bar{x}_3^2 = 6247.959,$$

$$l_{12} = l_{21} = \sum_{i=1}^{49} x_{i1}x_{i2} - 49\bar{x}_1\bar{x}_2 = -918.143,$$

$$l_{13} = l_{31} = \sum_{i=1}^{49} x_{i1}x_{i3} - 49\bar{x}_1\bar{x}_3 = -388.857,$$

$$l_{23} = l_{32} = \sum_{i=1}^{49} x_{i2}x_{i3} - 49\bar{x}_2\bar{x}_3 = 776.041,$$

$$l_{1y} = \sum_{i=1}^{49} x_{i1}y_i - 49\bar{x}_1\bar{y} = -6.433,$$



4.2 最小二乘估计及其性质

$$l_{2y} = \sum_{i=1}^{49} x_{i2} y_i - 49 \bar{x}_2 \bar{y} = 69.130,$$

$$l_{3y} = \sum_{i=1}^{49} x_{i3} y_i - 49 \bar{x}_3 \bar{y} = 245.571,$$

$$l_{yy} = \sum_{i=1}^{49} y_i^2 - 49 \bar{y}^2 = 44.905.$$

从而由(4.2-16)式得

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = L_{3 \times 3}^{-1} \begin{bmatrix} -6.433 \\ 69.130 \\ 245.571 \end{bmatrix} = \begin{bmatrix} 0.1606 \\ 0.1076 \\ 0.0359 \end{bmatrix},$$

再由(4.2-15)式得 $\hat{\beta}_0 = 0.6974$. 于是, 经验回归函数为

$$y = 0.6974 + 0.1606x_1 + 0.1076x_2 + 0.0359x_3.$$

又由(4.2-18)式算得

$$Q_e = 29.684.$$

因此, σ^2 的极大似然估计值

$$\hat{\sigma}^2 = \frac{29.684}{49} = 0.606.$$



4.3 线性模型的假设检验和统计推断

在线性模型中,要考虑下列的假设检验与统计推断问题:

- (1) 因变量 Y 与自变量 x_1, x_2, \dots, x_k 之间有线性相关关系(4.1-4)只是一种假设,它们是否具有这种线性关系,需要进行检验. 如果它们之间没有线性关系,则意味着所有的 $\beta_j (j=1, 2, \dots, k)$ 都应为零,这相当于检验假设 $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ 是否成立;
- (2) 如果 Y 与 x_1, x_2, \dots, x_k 之间确有线性关系,但并不意味着每个自变量对因变量 Y 都有显著影响. 若 x_j 对 Y 的影响不显著,就应该有 $\beta_j \approx 0$. 因此,要检验 x_j 对 Y 是否有显著影响,相当于检验假设 $H_0: \beta_j = 0$ 是否成立;
- (3) 在上述(1)与(2)的基础上,对未知参数 $\beta_0, \beta_1, \dots, \beta_k$ 作进一步统计推断,例如它们的点估计与区间估计,又如何根据经验回归函数来预测,当自变量 $x_j (j=1, 2, \dots, k)$ 取给定值时,因变量 Y 的值或取值范围,以及如果需要把 Y 的取值限制在某个范围内,那么 x_1, x_2, \dots, x_k 的取值应如何控制. 这些就是线性模型的预测和控制问题.



4.3 线性模型的假设检验和统计推断

4.3.1 线性模型的假设检验

首先对数据 y_1, y_2, \dots, y_n 的离差平方和 $l_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$ 进行分解, 其中 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. 记统计量

$$\begin{cases} SS = \sum_{i=1}^n (Y_i - \bar{Y})^2, \\ SS_r = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \\ SS_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2, \end{cases} \quad (4.3-1)$$

其中 $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$, \hat{Y}_i 的观测值 \hat{y}_i 是经验回归函数 $y = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_j$ 在 $x_j = x_{ij}$ ($j = 1, 2, \dots, k$) 处的值, 即

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij} = \bar{y} + \sum_{j=1}^k \hat{\beta}_j (x_{ij} - \bar{x}_j), \quad i = 1, 2, \dots, n. \quad (4.3-2)$$

称 SS 或其观测值(仍记为 SS , 下同)为总离差平方和, SS_r 为回归平方和, SS_e 为残差平方和. SS 反映了数据



4.3 线性模型的假设检验和统计推断

$$(x_{i1}, x_{i2}, \dots, x_{ik}; y_i), \quad i = 1, 2, \dots, n$$

中 y_1, y_2, \dots, y_n 的分散程度; SS_r 反映了 $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ 的分散程度, 而这 n 个数 \hat{y}_i 全在回归函数所表示的超平面上, 它们的分散性来源于 $x_{i1}, x_{i2}, \dots, x_{ik}$ 的分散性, 所以称为回归平方和; $SS_e = Q_e$ 反映了 y_i 与 \hat{y}_i 的离差, 而这是由于随机误差存在而引起的, 故称它为残差平方和.

引理(平方和分解公式) 在线性模型(4.1-5)中,

$$SS = SS_r + SS_e. \quad (4.3-3)$$

证 由(4.2-3)式得

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0},$$

从而有

$$(\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}^T \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0,$$

即

$$\hat{\mathbf{Y}}^T(\mathbf{Y} - \hat{\mathbf{Y}}) = 0,$$

亦即

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)\hat{Y}_i = 0.$$

又由(4.2-19)式中的第 1 个方程得

$$n\hat{\beta}_0 + (\sum_{i=1}^n x_{i1})\hat{\beta}_1 + \dots + (\sum_{i=1}^n x_{ik})\hat{\beta}_k = \sum_{i=1}^n Y_i,$$

即

$$\sum_{i=1}^n (Y_i - \hat{Y}_i) = 0.$$

于是, $SS = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$

4.3 线性模型的假设检验和统计推断

$$\begin{aligned}
 &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\
 &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = SS_e + SS_r.
 \end{aligned}$$



在假定线性模型(4.1-5)是正态线性模型的条件下,要检验假设

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0. \quad (4.3-4)$$

由于 $Y_i \sim N(\beta_0 + \sum_{j=1}^k \beta_j x_{ij}, \sigma^2)$, $i = 1, 2, \dots, n$, 并且它们相互独立, 所以当 H_0 成立,

即 $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ 时, (Y_1, Y_2, \dots, Y_n) 可以看成是取自正态总体 $Y \sim N(\beta_0; \sigma^2)$ 的一个容量为 n 的样本, 而 \bar{Y} 是样本均值, 从而由(1.4-2)式知

$$\frac{1}{\sigma^2} SS \sim \chi^2(n-1). \quad (4.3-5)$$

又由最小二乘估计的性质 5 知

$$\frac{1}{\sigma^2} SS_e \sim \chi^2(n-k-1). \quad (4.3-6)$$

可以证明, SS_r 与 SS_e 相互独立(请读者自证), 因此由(4.3-3)式及 χ^2 分布的可加性推得, 在 H_0 成立时,

$$\frac{1}{\sigma^2} SS_r \sim \chi^2(k). \quad (4.3-7)$$



4.3 线性模型的假设检验和统计推断

根据上述的讨论可知,对于假设检验问题(4.3-4),取检验统计量

$$F = \frac{SS_r/k}{SS_e/(n-k-1)},$$

当 H_0 成立时, $F \sim F(k, n-k-1)$, 而在 H_0 不成立时, 由于

$$SS_r = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n \left[\sum_{j=1}^k \hat{\beta}_j (x_{ij} - \bar{x}_j) \right]^2,$$

所以 F 值有变大的趋势,因此由

$$F = \frac{SS_r/k}{SS_e/(n-k-1)} > F_{1-\alpha}(k, n-k-1)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

具体计算时,常用表 4.3-1 所示的方差分析表.

表 4.3-1

方差来源	平方和	自由度	均方和	F 值
回归系数	$SS_r = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	k	$MS_r = \frac{SS_r}{k}$	$F = \frac{MS_r}{MS_e}$
残差	$SS_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$		$MS_e = \frac{SS_e}{n-k-1}$	
总和	$SS = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n-1$		

4.3 线性模型的假设检验和统计推断

4.3.2 回归系数的假设检验

如果经上述检验后拒绝了 H_0 , 那么可以认为回归系数在一定程度反映了整体上自变量 x_1, x_2, \dots, x_k 与因变量 Y 之间的相关关系, 然而不能排除个别自变量实际上对 Y 并无显著影响. 另一方面, 如果经检验后不能拒绝 H_0 , 那么也不能排除某些自变量对 Y 有显著影响, 只不过有可能由于这些自变量之间相互影响而使其回归效果不显著. 因此, 除了检验形如(4.3-4)式那样的假设外, 还需要对每个回归系数 β_j 分别进行检验假设

$$H_{0j}: \beta_j = 0. \quad (4.3-8)$$

如果经检验后不能拒绝 H_{0j} , 则认为 x_j 对 Y 的作用不显著, 否则便可以认为 x_j 对 Y 有显著影响.

由最小二乘估计的性质 5 可知, 当 H_0 成立即 $\beta_j = 0$ 时, $\hat{\beta}_j \sim N(0; l^{(jj)}\sigma^2)$, 其 $l^{(jj)}$ 是 k 阶方阵 $L_{k \times k}$ 的逆矩阵 $L_{k \times k}^{-1}$ 中的第 j 行第 j 列处的元素. 记

$$SS_{rj} = \frac{\hat{\beta}_j^2}{l^{(jj)}} \quad (j = 1, 2, \dots, k), \quad (4.3-9)$$

称它为第 j 个偏回归平方和. 在 H_0 成立时, 由于 $\frac{\hat{\beta}_j^2}{l^{(jj)}\sigma^2} \sim \chi^2(1)$, 所以

$$F_j = \frac{SS_{rj}}{SS_e/(n-k-1)} \sim F(1, n-k-1);$$

而在 H_0 不成立时, 由于

$$E(SS_{rj}) = E\left(\frac{\hat{\beta}_j^2}{l^{(jj)}}\right) = \frac{1}{l^{(jj)}}E(\hat{\beta}_j^2) = \frac{\beta_j^2}{l^{(jj)}} + \sigma^2,$$

4.3 线性模型的假设检验和统计推断

所以 SS_{ij} 有偏大的趋势,从而 F_j 也随之变大.因此,取检验统计量 F_j ,则由

$$F_j = \frac{SS_{ij}}{SS_e/(n-k-1)} > F_{1-\alpha}(1, n-k-1) \quad (4.3-10)$$

所确定的拒绝域给出了这个假设检验问题在显著性水平 α 下的一个检验,称其为偏 F 检验.

当剔除自变量 x_j 后,由剩下的 $k-1$ 个自变量所构成的 $k-1$ 元回归分析问题,其中诸量不必一一重新计算,它们可以借助于原 k 元回归分析问题中已经算得的数据.为了区别起见,我们把 $k-1$ 元回归分析问题中的诸量在右上角加“*”号.由(4.2-20)式知

$$\begin{bmatrix} l_{11} & \cdots & l_{1k} \\ \vdots & \ddots & \vdots \\ l_{j1} & \cdots & l_{jk} \\ \vdots & \ddots & \vdots \\ l_{k1} & \cdots & l_{kk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_j \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} l_{1y} \\ \vdots \\ l_{jy} \\ \vdots \\ l_{ky} \end{bmatrix}, \quad \begin{bmatrix} l_{11} & \cdots & l_{1k} \\ \vdots & \ddots & \vdots \\ l_{j1} & \cdots & l_{jk} \\ \vdots & \ddots & \vdots \\ l_{k1} & \cdots & l_{kk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1^* \\ \vdots \\ 0 \\ \vdots \\ \hat{\beta}_k^* \end{bmatrix} = \begin{bmatrix} l_{1y} \\ \vdots \\ \sum_{\substack{k=1 \\ i \neq j}}^j l_{ji} \hat{\beta}_i^* \\ \vdots \\ l_{ky} \end{bmatrix}.$$

将上述两式相减,得到

$$\begin{bmatrix} l_{11} & \cdots & l_{1k} \\ \vdots & \ddots & \vdots \\ l_{j1} & \cdots & l_{jk} \\ \vdots & \ddots & \vdots \\ l_{k1} & \cdots & l_{kk} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1^* - \hat{\beta}_1 \\ \vdots \\ -\hat{\beta}_j \\ \vdots \\ \hat{\beta}_k^* - \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ \sum_{\substack{i=1 \\ i \neq j}}^k l_{ji} \hat{\beta}_i^* - l_{jy} \\ \vdots \\ 0 \end{bmatrix},$$



4.3 线性模型的假设检验和统计推断

解此方程组得

$$-\hat{\beta}_j = l^{(jj)} \left(\sum_{\substack{i=1 \\ i \neq j}}^k l_{ji} \hat{\beta}_i^* - l_{jy} \right),$$

$$\hat{\beta}_m^* - \hat{\beta}_m = l^{(mj)} \left(\sum_{\substack{i=1 \\ i \neq j}}^k l_{ji} \hat{\beta}_i^* - l_{jy} \right), m = 1, \dots, j-1, j+1, \dots, k.$$

由此可得剔除 x_j 后的计算公式为

$$\hat{\beta}_m^* = \hat{\beta}_m - \frac{l^{(mj)}}{l^{(jj)}} \hat{\beta}_j, \quad m = 1, \dots, j-1, j+1, \dots, k, \quad (4.3-11)$$

$$\hat{\beta}_0^* = \bar{y} - (\hat{\beta}_1^* \bar{x}_1 + \dots + \hat{\beta}_{j-1}^* \bar{x}_{j-1} + \hat{\beta}_{j+1}^* \bar{x}_{j+1} + \dots + \hat{\beta}_k^* \bar{x}_k), \quad (4.3-12)$$

$$\begin{cases} SS^* = SS, \\ SS_r^* = SS_r - SS_{rj}, \\ SS_e^* = SS_e + SS_{rj}. \end{cases} \quad (4.3-13)$$

由于涉及 k 阶方阵 $L_{k \times k}$ 的逆矩阵 $L_{k \times k}^{-1}$ 的元素, 所以当 $L_{k \times k}$ 为对角矩阵时, 计算特别简单. 这时, $L_{k \times k}^{-1}$ 也为对角矩阵, 故有 $l^{(mj)} = 0, m \neq j$. 从而由(4.3-11)式看出, $\hat{\beta}_m^* = \hat{\beta}_m, m = 1, \dots, j-1, j+1, \dots, k$. 又因 $l_{jj} = \frac{1}{l^{(jj)}}$, 故有

$$SS_r = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n \left[\sum_{j=1}^k \hat{\beta}_j (x_{ij} - \bar{x}_j) \right]^2 = \sum_{j=1}^k l_{jj} \hat{\beta}_j^2 = \sum_{j=1}^k SS_{rj}.$$

这表明回归平方和恰是 k 个偏回归平方和之和.

4.3 线性模型的假设检验和统计推断

由 L_{pr} 的定义看出, k 阶方阵 $L_{k \times k}$ 为对角矩阵的充分必要条件是, $n \times k$ 矩阵

$$\begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}$$

中 k 个列向量两两正交. 因此, 如何安排试验使 $L_{k \times k}$ 为对角矩阵是一个有意义的研究课题.

例 1 设有下列数据:

x_1	1000	600	1200	500	300	400	1300	1100	1300	300
x_2	5	7	6	6	8	7	5	4	3	9
y	100	75	80	70	50	65	90	100	110	60

若用回归分析方法来处理这些数据, 则要建立 y 与 x_1, x_2 之间的经验回归函数, 并对其进行检验, 显著性水平 α 取为 0.05.

这时, $n=10, k=2$. 由所给数据算得

$$\bar{x}_1 = 800, \quad \bar{x}_2 = 6, \quad \bar{y} = 80;$$

$$l_{11} = 1580000, \quad l_{22} = 30, \quad l_{12} = l_{21} = -5900;$$

$$l_{1y} = 65000, \quad l_{2y} = -300, \quad l_{yy} = 3450.$$

从而, (4.2-16)式和(4.2-15)式给出

4.3 线性模型的假设检验和统计推断

$$\hat{\beta}_1 = 0.0143, \hat{\beta}_2 = -7.1882, \hat{\beta}_0 = 111.6892.$$

于是,经验回归函数为 $y = 111.6892 + 0.0143x_1 - 7.1882x_2$.

下面对这个模型进行显著性检验.

列出方差分析如表 4.3-2 所示.

表 4.3-2

方差来源	平方和	自由度	均方和	F 值
回归系数	3086	2	1543	29.673
残 差	364	7	52	
总 和	3450	9		

对于上述线性模型,要检验假设

$$H_0: \beta_1 = \beta_2 = 0.$$

由 $\alpha=0.05$ 得到临界值 $F_{0.95}(2,7)=4.74$, 显然有 $29.673 > 4.74$, 因此拒绝 H_0 , 即可以认为回归系数在一定程度上反映了 x_1, x_2 与 Y 之间的相关关系.

考虑到 $|\hat{\beta}_1| < |\hat{\beta}_2|$, 故再检验

$$H_{01}: \beta_1 = 0.$$

由于

$$t^{(11)} = \frac{l_{22}}{l_{11}l_{22} - l_{12}l_{21}} = 2.383 \times 10^{-6},$$



4.3 线性模型的假设检验和统计推断

因此偏回归平方和的观测值为

$$SS_{r1} = \frac{(0.0143)^2}{2.383 \times 10^{-6}} = 85.812.$$

从而偏 F 检验统计量 F_1 的观测值为

$$F_1 = \frac{85.812}{52} = 1.650.$$

由 $\alpha=0.05$ 得到临界值 $F_{0.95}(1, 7)=5.59$, 显然有 $1.650 < 5.59$, 因此不能拒绝 H_{01} , 即可以认为 x_1 对 y 的影响不大.

剔除了变量 x_1 之后, 需要建立 y 与 x_2 之间的经验回归函数. 由已算得的 t_{22} 和 t_{2y} 得

$$\hat{\beta}_2^* = \frac{t_{2y}}{t_{22}} = -10,$$

再由(4.3-12)式得 $\hat{\beta}_0^* = \bar{y} - \hat{\beta}_2^* \bar{x}_2 = 140$. 于是, y 与 x_2 之间的经验回归函数为

$$y = 140 - 10x_2. \quad (4.3-14)$$

现对(4.3-14)式所对应的线性正态模型进行检验. 为此, 列出其方差分析表 4.3-3.

表 4.3-3

方差来源	平方和	自由度	均方和	F 值
回归系数	3000	1	3000	53.33
残 差	450	8	56.25	
总 和	3450	9		



4.3 线性模型的假设检验和统计推断

对于检验

$$H_0: \beta_2^* = 0,$$

由 $\alpha=0.05$ 得到临界值 $F_{0.95}(1, 8)=5.32$. 显然 $53.33 > 5.32$, 故拒绝 H_0 , 亦即认为经验回归函数(4.3-14)是可以接受的. ■

4.3.3 统计推断

当回归分析指出某些自变量对因变量的影响有显著作用时, 则可以进一步讨论统计推断的问题.

(1) 关于回归系数的区间估计.

设 $\theta = \sum_{j=0}^k c_j \beta_j$ 是待估函数, 其中 $c_j (j = 0, 1, \dots, k)$ 是给定的常数, 则由最小二乘估计的性质 4 知, $\hat{\theta} = \sum_{j=0}^k c_j \hat{\beta}_j$ 是 θ 的一个较优的点估计量, 其均值 $E(\hat{\theta}) = E\left(\sum_{j=0}^k c_j \hat{\beta}_j\right) = \sum_{j=0}^k c_j E(\hat{\beta}_j) = \sum_{j=0}^k c_j \beta_j = \theta$, 且(4.2-11)式给出 $\hat{\theta}$ 的方差为

$$D(\hat{\theta}) = D\left(\sum_{j=0}^k c_j \hat{\beta}_j\right) = \sigma^2 \mathbf{c}^\top \mathbf{L}^{-1} \mathbf{c}.$$



4.3 线性模型的假设检验和统计推断

又因为 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 都是正态随机变量, 所以 $\hat{\theta}$ 也是正态随机变量, 从而 $\hat{\theta} \sim N(\theta; \sigma^2 \mathbf{c}^\top \mathbf{L}^{-1} \mathbf{c})$. 再则, 由最小二乘估计的性质 5 知, $\hat{\theta}$ 与 \mathbf{Q}_e 相互独立, 故有

$$\frac{\hat{\theta} - \theta}{\sigma \sqrt{\mathbf{c}^\top \mathbf{L}^{-1} \mathbf{c}}} \Bigg/ \frac{1}{\sigma} \sqrt{\frac{\mathbf{Q}_e}{n-k-1}} = \frac{\hat{\theta} - \theta}{\hat{\sigma}_e \sqrt{\mathbf{c}^\top \mathbf{L}^{-1} \mathbf{c}}} \sim t(n-k-1). \quad (4.3-15)$$

因此, 对于给定的置信水平 $1-\alpha$, θ 的双侧 $1-\alpha$ 置信区间的上、下限是

$$\hat{\theta} \pm t_{1-\alpha/2}(n-k-1) \hat{\sigma}_e \sqrt{\mathbf{c}^\top \mathbf{L}^{-1} \mathbf{c}}.$$

特别地, 若取 $\mathbf{c} = [0, \dots, 0, \overset{j}{1}, 0, \dots, 0]^\top$, 则得 $\theta = \beta_j$, 从而 β_j 的双侧 $1-\alpha$ 置信区间的上、下限为

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-k-1) \hat{\sigma}_e \sqrt{(\mathbf{L}^{-1})_{jj}} \quad (j = 0, 1, \dots, k), \quad (4.3-16)$$

其中 $(\mathbf{L}^{-1})_{jj}$ 表示 $k+1$ 阶方阵 \mathbf{L}^{-1} 中的第 j 行第 j 列处的元素.

(2) 关于 Y_t 的预测区域.

对于正态线性模型(4.1-6), 求得 $\boldsymbol{\beta}$ 的最小二乘估计值 $\hat{\boldsymbol{\beta}}$ 后, 可以用(4.2-4)式对 Y 进行预测. 事实上, 若 x_1, x_2, \dots, x_k 在时刻 t 的值依次为 $x_{t1}, x_{t2}, \dots, x_{tk}$, 则(4.2-4)式给出

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \dots + \hat{\beta}_k x_{tk},$$

它可以作为随机变量 Y 在时刻 t 的预测值. 记

$$\mathbf{X}_t = [1, x_{t1}, \dots, x_{tk}]^\top, \quad (4.3-17)$$

则统计量 \hat{Y}_t 可以表示为

$$\hat{Y}_t = \mathbf{X}_t^\top \hat{\boldsymbol{\beta}}.$$

4.3 线性模型的假设检验和统计推断

由最小二乘估计的性质 5 知, \hat{Y}_t 服从正态分布 $N(\mathbf{X}_t^T \boldsymbol{\beta}; \sigma^2 \mathbf{X}_t^T \mathbf{L}^{-1} \mathbf{X}_t)$; 而 $Y_t \sim N(\mathbf{X}_t^T \boldsymbol{\beta}; \sigma^2)$, 所以有

$$Y_t - \hat{Y}_t \sim N(0; \sigma^2(1 + \mathbf{X}_t^T \mathbf{L}^{-1} \mathbf{X}_t)).$$

由于 Y_t 与 \mathbf{Y} 相互独立, 所以 Y_t 与 $e = (\mathbf{I}_n - \mathbf{X}\mathbf{L}^{-1}\mathbf{X}^T)\mathbf{Y}$ 相互独立, 故 Y_t 与 Q_e 相互独立. 又由最小二乘估计的性质 5 知, $\hat{\boldsymbol{\beta}}$ 与 Q_e 相互独立, 故 $\hat{Y}_t = \mathbf{X}_t^T \hat{\boldsymbol{\beta}}$ 与 Q_e 相互独立. 于是, $Y_t - \hat{Y}_t$ 与 Q_e 相互独立. 注意到 $Q_e / \sigma^2 \sim \chi^2(n-k-1)$, 从而有

$$\frac{Y_t - \hat{Y}_t}{\sigma \sqrt{1 + \mathbf{X}_t^T \mathbf{L}^{-1} \mathbf{X}_t}} / \sqrt{\frac{Q_e}{n-k-1}} = \frac{Y_t - \hat{Y}_t}{\hat{\sigma}_e \sqrt{1 + \mathbf{X}_t^T \mathbf{L}^{-1} \mathbf{X}_t}} \sim t(n-k-1). \quad (4.3-18)$$

于是, 对于给定的置信水平 $1-\alpha$, Y_t 的双侧 $1-\alpha$ 预测区间的上、下限为

$$\hat{Y}_t \pm \hat{\sigma}_e \sqrt{1 + \mathbf{X}_t^T \mathbf{L}^{-1} \mathbf{X}_t} t_{1-\alpha/2}(n-k-1). \quad (4.3-19)$$



4.4 方差分析

在方差分析问题中,称影响我们所关心的某个指标(因变量)的那些因素(自变量)为因子,常用 $A, B \dots$ 来表示. 因子所取的不同状态(相当于自变量的取值)称为水平,因子 A 的 a 个不同水平用 A_1, A_2, \dots, A_a 表示. 例如,三名实验员各用四种不同型号的仪器对同一物理常数进行测定,我们希望了解不同实验员及不同型号的仪器对测定值有何影响. 这时,实验员和测量仪器是因子,分别记为 A 和 B ,那么因子 A 有 3 个水平,记为 A_1, A_2, A_3 ;而因子 B 有 4 个水平,记为 B_1, B_2, B_3, B_4 .

由于因子的水平不同对指标的影响一般是不同的,所以根据样本对此进行统计分析,就构成了方差分析研究的对象. 当只考虑一个因子的不同水平的影响,则称单因子方差分析;若考虑两个因子的不同水平的影响,则称双因子方差分析;若考虑两个以上因子,则称多因子方差分析. 由于多因子方差分析的方法与双因子的方差分析方法并无本质上的差别,所以对它不再讨论.

在方差分析中,总是假定样本取自正态总体,并且各个正态总体的方差都相等.

方差分析问题的统计模型本质上是一个多元正态线性模型,但由于自变量只取有限个状态(即水平),所以通常不用形如(4.1-6)式与(4.1-7)式所构成的正态线性模型.

在单因子方差分析中,设因子 A 有 a 个不同水平 A_1, A_2, \dots, A_a ,把每个水平 A_i 下我们所关心的因变量看作一个总体 $X^{(i)}$,且 $X^{(i)} \sim N(\mu_i; \sigma^2)$. 从第 i 个总体 $X^{(i)}$ 抽取的一个容量为 m 的样本 $(X_{i1}, X_{i2}, \dots, X_{im})$ 称为第 i 个样本,并假定所取得的 a 个容量为 m 的样本是相互独立的,即 $X_{11}, X_{12}, \dots, X_{1m}; \dots; X_{a1}, X_{a2}, \dots, X_{am}$ 相互独立. 这时,我们可以用下述正态线性模型来表征单因子方差分析问题:



4.4 方差分析

$$X_{ik} = \mu_i + \varepsilon_{ik}, \quad k = 1, 2, \dots, m, \quad i = 1, 2, \dots, a, \quad (4.4-1)$$

其中 $\varepsilon_{11}, \varepsilon_{12}, \dots, \varepsilon_{1m}; \dots; \varepsilon_{a1}, \varepsilon_{a2}, \dots, \varepsilon_{an}$ 是独立同分布的随机变量,且每一个 $\varepsilon_{ik} \sim N(0; \sigma^2)$. 这里的 $\mu_1, \mu_2, \dots, \mu_a$ 及 σ^2 是 $a+1$ 个未知参数. 因子 A 在不同水平下的作用是通过均值 $\mu_1, \mu_2, \dots, \mu_a$ 来反映的.

记

$$\mu = \frac{1}{a} \sum_{i=1}^a \mu_i, \quad (4.4-2)$$

称 μ 为总平均; 又记 $\delta_i = \mu_i - \mu, \quad i = 1, 2, \dots, a,$ (4.4-3)

称 δ_i 为因子 A 在第 i 个水平 A_i 下的效应. 显然, 这 a 个效应 $\delta_1, \delta_2, \dots, \delta_a$ 满足关系式

$$\sum_{i=1}^a \delta_i = 0. \quad (4.4-4)$$

引进效应的概念后, 因子 A 在不同水平下的作用差异就可通过效应来表示, 且可以把(4.4-1)式等价地表示为

$$X_{ik} = \mu + \delta_i + \varepsilon_{ik}, \quad k = 1, 2, \dots, m, \quad i = 1, 2, \dots, a. \quad (4.4-5)$$

这里 $\mu, \delta_1, \delta_2, \dots, \delta_a$ 及 σ^2 均是未知参数, 但由于约束条件(4.4-4)式, 故(4.4-5)式实际上只包含 $a+1$ 个未知参数.

在双因子方差分析中, 设因子 A 取 a 个不同的水平 A_1, A_2, \dots, A_a , 因子 B 取 b 个不同的水平 B_1, B_2, \dots, B_b . 把因子 A 取水平 A_i 且因子 B 取水平 B_j 时, 我们所关心的因变量看作一个总体 $X^{(ij)}$, 并设 $X^{(ij)} \sim N(\mu_{ij}; \sigma^2)$. 从每个总体 $X^{(ij)}$ 抽取一个容量为 m 的样本



4.4 方差分析

$(X_{ij1}, X_{ij2}, \dots, X_{ijm}), i=1, 2, \dots, a, j=1, 2, \dots, b$, 并且假定所有 X_{ijk} 相互独立. 于是可以建立下述的正态线性模型来讨论双因子方差分析问题:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad k = 1, 2, \dots, m; \quad i = 1, 2, \dots, a, \quad j = 1, 2, \dots, b, \quad (4.4-6)$$

其中 $\varepsilon_{111}, \varepsilon_{112}, \dots, \varepsilon_{11m}; \dots; \varepsilon_{ab1}, \varepsilon_{ab2}, \dots, \varepsilon_{abm}$ 是独立同分布的随机变量, 且每一个 $\varepsilon_{ijk} \sim N(0; \sigma^2)$. 这里,

$$\mu = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b \mu_{ij}, \quad (4.4-7)$$

$$\mu_{ii} = \frac{1}{b} \sum_{j=1}^b \mu_{ij}, \quad \alpha_i = \mu_{ii} - \mu, \quad i = 1, 2, \dots, a, \quad (4.4-8)$$

$$\mu_{..j} = \frac{1}{a} \sum_{i=1}^a \mu_{ij}, \quad \beta_j = \mu_{..j} - \mu, \quad j = 1, 2, \dots, b. \quad (4.4-9)$$

并且称 μ 为总平均, 称 α_i 为因子 A 在第 i 个水平 A_i 下的效应, 称 β_j 为因子 B 在第 j 个水平 B_j 下的效应. 显然

$$\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0. \quad (4.4-10)$$

再则, 记

$$\begin{aligned} \gamma_{ij} &= \mu_{ij} - (\mu + \alpha_i + \beta_j) = \mu_{ij} - \mu_{ii} - \mu_{..j} + \mu \\ &\quad i = 1, 2, \dots, a, \quad j = 1, 2, \dots, b, \end{aligned} \quad (4.4-11)$$

4.4 方差分析

称 γ_{ij} 为因子 A 的第 i 个水平 A_i 与因子 B 的第 j 个水平 B_j 之间的交互效应. 如果 $\gamma_{ij} = 0$, 即 $\mu_{ij} = \mu + \alpha_i + \beta_j$ ($i = 1, 2, \dots, a, j = 1, 2, \dots, b$), 则称这类方差分析为无交互作用的双因子方差分析问题. 这时, 两个因子的联合作用可以由每个因子效应的简单迭加来表示.

由(4.4-11)式得,

$$\begin{cases} \sum_{j=1}^b \gamma_{ij} = 0, & i = 1, 2, \dots, a, \\ \sum_{i=1}^a \gamma_{ij} = 0, & j = 1, 2, \dots, b, \end{cases} \quad (4.4-12)$$

因而对于有交互作用的双因子方差分析问题(4.4-6)式来说, $\mu, \alpha_1, \dots, \alpha_a, \beta_1, \dots, \beta_b, \gamma_{11}, \dots, \gamma_{ab}$ 及 σ^2 都是未知参数, 但由于(4.4-10)式和(4.4-12)式, 故实际上只有

$$1 + a + b + ab + 1 - [1 + 1 + (a + b - 1)] = ab + 1$$

个未知参数.

对于无交互作用的双因子方差分析问题来说, $\mu, \alpha_1, \dots, \alpha_a, \beta_1, \dots, \beta_b$ 及 σ^2 是未知参数, 但由于(4.4-10)式, 故有

$$1 + a + b + 1 - (1 + 1) = a + b$$

个未知参数.

对于多因子方差分析问题, 也可以按(4.4-6)式类似地建立相应的正态线性模型, 只是交互效应的形式更为复杂. 这时, 除了任意两个因子在各对水平搭配下有交互效应外, 还需考虑任意三个因子在各组水平搭配下的交互效应, 等等.

4.4 方差分析

4.4.1 单因子方差分析

在单因子方差分析问题(4.4-1)中,因子A在不同水平下对我们所关心的指标的影响是通过均值 $\mu_1, \mu_2, \dots, \mu_a$ 来反映的.因此,如果水平的改变并不影响指标,那么在各个水平 A_1, A_2, \dots, A_a 下的总体均值 $\mu_1, \mu_2, \dots, \mu_a$ 应该相等,从而 $\delta_1 = \delta_2 = \dots = \delta_a = 0$.反之,如果总有一些 δ_i 不等于零,那么说明因子A的影响是显著的,因而要进一步找出因子A取哪个水平时效果最佳.这就是方差分析所要解决的问题.

检验假设 $H_0: \mu_1 = \mu_2 = \dots = \mu_a$ (4.4-13)

等价于检验假设 $H_0: \delta_1 = \delta_2 = \dots = \delta_a = 0$. (4.4-13')

容易看出,这与4.3节所述的线性模型假设检验问题类似.因此,设第*i*个样本均值为

$$\bar{X}_{i\cdot} = \frac{1}{m} \sum_{j=1}^m X_{ij}, i = 1, 2, \dots, a,$$

*a*个样本的总均值为 $\bar{X} = \frac{1}{am} \sum_{i=1}^a \sum_{j=1}^m X_{ij} = \frac{1}{a} \sum_{i=1}^a \bar{X}_{i\cdot}$ (4.4-14)

记统计量
$$\begin{cases} SS = \sum_{i=1}^a \sum_{j=1}^m (X_{ij} - \bar{X})^2, \\ SS_A = m \sum_{i=1}^a (\bar{X}_{i\cdot} - \bar{X})^2, \\ SS_e = \sum_{i=1}^a \sum_{j=1}^m (X_{ij} - \bar{X}_{i\cdot})^2. \end{cases}$$
 (4.4-15)

4.4 方差分析

这里, SS 反映了全体数据的波动; SS_A 反映由于因子 A 在各个水平下的不同作用而在数据中引起的波动, 其中常数因子 m 表示对每个总体 $X^{(i)}$ 重复观测了 m 次; SS_e 反映由于随机误差的作用而在数据中引起的波动. SS 称为总离差平方和, SS_A 称为因子 A 的离差平方和(也称为组间离差平方和), SS_e 称为误差平方和(也称为组内离差平方和).

引理 1(平方和分解公式) 在单因子方差分析问题(4.4-5)中,

$$SS = SS_A + SS_e. \quad (4.4-16)$$

证 由于 $\sum_{j=1}^m (X_{ij} - \bar{X}_{i\cdot}) = \sum_{j=1}^m X_{ij} - m \bar{X}_{i\cdot} = 0$,

所以 $SS = \sum_{i=1}^a \sum_{j=1}^m (X_{ij} - \bar{X})^2 = \sum_{i=1}^a \sum_{j=1}^m [(X_{ij} - \bar{X}_{i\cdot}) + (\bar{X}_{i\cdot} - \bar{X})]^2$

$$= \sum_{i=1}^a \sum_{j=1}^m (X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^a \sum_{j=1}^m (\bar{X}_{i\cdot} - \bar{X})^2 + 2 \sum_{i=1}^a \sum_{j=1}^m (X_{ij} - \bar{X}_{i\cdot})(\bar{X}_{i\cdot} - \bar{X})$$

$$= SS_e + SS_A + 2 \sum_{i=1}^a [(\bar{X}_{i\cdot} - \bar{X}) \sum_{j=1}^m (X_{ij} - \bar{X}_{i\cdot})]$$

$$= SS_e + SS_A.$$

定理 4.4-1 在单因子方差分析问题(4.4-5)中, SS_A 与 SS_e 相互独立, 且 SS_e/σ^2 服从 $\chi^2(n-a)$, 其 $n=am$. 当 H_0 成立, 即 $\delta_1=\delta_2=\dots=\delta_a=0$ 时, SS_A/σ^2 服从 $\chi^2(a-1)$, 从而

$$F = \frac{SS_A/(a-1)}{SS_e/(n-a)} \sim F(a-1, n-a). \quad (4.4-17)$$



4.4 方差分析

证 对每一个 $i (i=1, 2, \dots, a)$, 总体 $X^{(i)}$ 的样本均值 $\bar{X}_{i..}$ 与样本方差 $\frac{1}{m-1} \sum_{j=1}^m (X_{ij} - \bar{X}_{i..})^2$ 相互独立; 又由全体样本相互独立知

$$(\bar{X}_{1..}, \sum_{j=1}^m (X_{ij} - \bar{X}_{1..})^2), \dots, (\bar{X}_{a..}, \sum_{j=1}^m (X_{aj} - \bar{X}_{a..})^2)$$

相互独立; 因此 $2a$ 个随机变量 $\bar{X}_{1..}, \dots, \bar{X}_{a..}, \sum_{j=1}^m (X_{1j} - \bar{X}_{1..})^2, \dots, \sum_{j=1}^m (X_{aj} - \bar{X}_{a..})^2$

相互独立. 从而 $(\bar{X}_{1..}, \bar{X}_{2..}, \dots, \bar{X}_{a..})$ 与 SS_e 相互独立, 由此推知 SS_A 与 SS_e 相互独立.

对于第 i 个样本, 有 $\sum_{j=1}^m (X_{ij} - \bar{X}_{i..})^2 / \sigma^2 \sim \chi^2(m-1), i = 1, 2, \dots, a$, 因而得到

$$\frac{SS_e}{\sigma^2} = \sum_{i=1}^a \frac{\sum_{j=1}^m (X_{ij} - \bar{X}_{i..})^2}{\sigma^2} \sim \chi^2(n-a), \quad n = am.$$

再则, 当 H_0 成立时, $X_{11}, X_{12}, \dots, X_{am}$ 是独立同分布的随机变量, 且每一个 X_{ij} 服从正态分布 $N(\mu, \sigma^2)$, 从而 $\bar{X}_{i..} = \frac{1}{m} \sum_{j=1}^m X_{ij} \sim N\left(\mu; \frac{\sigma^2}{m}\right), i = 1, 2, \dots, a$. 因此, 可以把 $(\bar{X}_{1..}, \bar{X}_{2..}, \dots, \bar{X}_{a..})$ 看为取自正态总体 $N\left(\mu; \frac{\sigma^2}{m}\right)$ 的一个大小为 a 的样本. 于是, 由 (4.4-14) 式和 (1.4-18) 式知,

$$\frac{SS_A}{\sigma^2} = \frac{\sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2}{\sigma^2/m} \sim \chi^2(a-1).$$



4.4 方差分析

这样便由 F 分布的定义知,

$$F = \frac{SS_A/(a-1)}{SS_e/(n-a)} \sim F(a-1, n-a).$$

因此,对于假设检验问题(4.4-13)式,可以取检验统计量为

$$F = \frac{SS_A/(a-1)}{SS_e/(n-a)},$$

且当 H_0 成立时, $F \sim F(a-1, n-a)$; 而当 H_0 不成立时, 由于

$$E(SS_A) = mE\left[\sum_{i=1}^a (\bar{X}_{ii} - \bar{X})^2\right] = m \sum_{i=1}^a E[(\bar{X}_{ii} - \bar{X})^2] = (a-1)\sigma^2 + m \sum_{i=1}^a \delta_i^2,$$

所以 SS_A 有偏大的趋势. 于是, 对于假设检验(4.4-13)式, 由

$$F = \frac{SS_A/(a-1)}{SS_e/(n-a)} > F_{1-\alpha}(a-1, n-a) \quad (4.4-18)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验. 在具体计算时, 常用方差分析表 4.4-1.

表 4.4-1

方差来源	平方和	自由度	均方和	F 值
因子 A	$SS_A = m \sum_{i=1}^a (\bar{X}_{ii} - \bar{X})^2$	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F = \frac{MS_A}{MS_e}$
误差	$SS_e = \sum_{i=1}^a \sum_{j=1}^m (X_{ij} - \bar{X}_{ii})^2$	$n-a$	$MS_e = \frac{SS_e}{n-a}$	
总和	$SS = \sum_{i=1}^a \sum_{j=1}^m (X_{ij} - \bar{X})^2$	$n-1$		



4.4 方差分析

如果检验结果是拒绝 H_0 , 那么希望进一步找出因子 A 取哪个水平效果最佳. 这时可以通过检验

$$H_0: \mu_{i_1} = \mu_{i_2} \text{ (即 } \mu_{i_1} - \mu_{i_2} = 0) \quad (4.4-19)$$

来解决, 这里 A_{i_1}, A_{i_2} 是因子 A 的某两个水平. 为了方便起见, 不妨假定 $i_1=1, i_2=2$.

$\bar{X}_{1\cdot}, \bar{X}_{2\cdot}$ 是 $\mu_{1\cdot}, \mu_{2\cdot}$ 的一个较好的估计, 且 $\bar{X}_{1\cdot}, \bar{X}_{2\cdot} \sim N(\mu_{1\cdot}, \mu_{2\cdot}; \frac{2}{m}\sigma^2)$. 因此, 当 H_0 成立, 即 $\mu_{1\cdot} - \mu_{2\cdot} = 0$ 时, 检验统计量

$$T_{1,2} = \sqrt{\frac{m}{2}} \frac{\bar{X}_{1\cdot} - \bar{X}_{2\cdot}}{\sqrt{\frac{SS_e}{n-a}}} \sim t(n-a). \quad (4.4-20)$$

于是, 对于假设检验(4.4-19), 由

$$|T_{1,2}| = \frac{\sqrt{\frac{m}{2}} |\bar{X}_{1\cdot} - \bar{X}_{2\cdot}|}{\sqrt{\frac{SS_e}{n-a}}} > t_{1-\alpha/2}(n-a) \quad (4.4-21)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

例 1 为了比较四种不同肥料对小麦亩产量的影响, 取一片土壤肥沃程度和水利灌溉条件差不多的土地, 分成 16 块. 将肥料品种记为 A_1, A_2, A_3, A_4 , 每种肥料施在四块土地上, 得亩产量(单位: 公斤)如下:

4.4 方差分析

肥料品种	亩产量				$\bar{x}_{i\cdot}$
A_1	198	196	190	166	187.50
A_2	160	169	167	150	161.50
A_3	179	164	181	170	173.50
A_4	190	170	179	188	181.75

问施肥品种对小麦亩产量有无显著影响? ($\alpha=0.05$)

解 由所给数据算得 $\bar{x}=176.06$, 列出方差分析表 4.4-2:

表 4.4-2

方差来源	平方和	自由度	均方和	F 值
因子(施肥品种)	1527.19	3	509.06	4.65
误差	1313.75	12	109.48	
总和	2840.94	15		

对于假设检验(4.4-13), 取显著性水平 $\alpha=0.05$, 查表得临界值 $F_{0.95}(3,12)=3.49$. 显然, $4.65>3.49$, 因此拒绝 H_0 , 即认为施肥品种对小麦亩产量有显著影响. 进一步, 自然希望找出最佳的施肥品种. 由所给数据表中的最后一列看出, 施肥品种 A_1, A_4 较优, 因此要检验

$$H_0: \mu_{1\cdot} = \mu_{4\cdot}$$

按(4.4-20)式, 检验统计量 $T_{1,4}$ 的观测值为

$$t_{1,4} = \sqrt{\frac{4}{2}} \times \frac{187.50 - 181.75}{\sqrt{109.48}} = 0.7772.$$



4.4 方差分析

在显著性水平 $\alpha=0.05$ 下, 查表得临界值 $t_{0.975}(12)=2.1788$. 显然, $0.7772 < 2.1788$, 因此不能拒绝 H_0 , 即可以认为施肥品种 A_1 与 A_4 对小麦亩产量的作用无显著差异, 从而这两种品种都是较佳的施肥品种. ■

4.4.2 双因子方差分析

首先讨论无交互作用的双因子方差分析模型:

$$\begin{cases} X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, i = 1, 2, \dots, a, j = 1, 2, \dots, b; \\ \sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0; \\ \epsilon_{11}, \epsilon_{12}, \dots, \epsilon_{ab} \text{ 是独立同分布的随机变量, 且每一个 } \epsilon_{ij} \sim N(0; \sigma^2). \end{cases} \quad (4.4-22)$$

这就是说, 现有两个因子 A 和 B , 因子 A 取 a 个不同水平, 因子 B 取 b 个不同水平, 且对这两个因子的每一对水平搭配 (A_i, B_j) , 有容量为 1 的样本. 通常称其为双因子每对水平搭配只观测一次(或非重复)试验的方差分析.

记

$$\bar{X}_{i\cdot} = \frac{1}{b} \sum_{j=1}^b X_{ij}, \quad i = 1, 2, \dots, a,$$

$$\bar{X}_{\cdot j} = \frac{1}{a} \sum_{i=1}^a X_{ij}, \quad j = 1, 2, \dots, b,$$

$$\bar{X} = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b X_{ij} = \frac{1}{a} \sum_{i=1}^a \bar{X}_{i\cdot} = \frac{1}{b} \sum_{j=1}^b \bar{X}_{\cdot j},$$

4.4 方差分析

并记 $n=ab$. 现要分别检验假设

$$H_{0A}: \alpha_1 = \alpha_2 = \dots = \alpha_a = 0 \quad (4.4-23)$$

及

$$H_{0B}: \beta_1 = \beta_2 = \dots = \beta_b = 0. \quad (4.4-24)$$

采用的方法仍是把总离差平方和分解为由因素 A 引起的离差平方和、由因素 B 引起的离差平方和以及误差平方和三项之和.

记统计量 $SS = \sum_{i=1}^a \sum_{j=1}^b (X_{ij} - \bar{X})^2$, $SS_A = b \sum_{i=1}^a (\bar{X}_{i\cdot} - \bar{X})^2$,

$$SS_B = a \sum_{j=1}^b (\bar{X}_{\cdot j} - \bar{X})^2, \quad SS_e = \sum_{i=1}^a \sum_{j=1}^b (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2.$$

直观上, SS 反映全体数据中的波动; SS_A 反映由于因子 A 在各个水平下的不同作用而引起的波动, 其中常数 b 表示每个水平 A_i 在各对水平搭配中出现了 b 次; SS_B 的意义与 SS_A 类同; SS_e 反映由于随机误差的作用而在数据中引起的波动. 分别称 SS 为总离差平方和, SS_A 为因子 A 的离差平方和, SS_B 为因子 B 的离差平方和, SS_e 为误差平方和.

引理 2(平方和分解公式) 在无交互作用的双因子方差分析模型(4.4-22)中,

$$SS = SS_A + SS_B + SS_e. \quad (4.4-25)$$

证明的思路与单因子方差分析的类同, 请读者自证, 或参阅文献[2]. ■

定理 4.4-2 在无交互作用的双因子方差分析模型(4.4-22)中, SS_A 、 SS_B 、 SS_e 相互独立, 且 $SS_e/\sigma^2 \sim \chi^2(n-a-b+1)$. 当 H_{0A} 成立, 即 $\alpha_1 = \alpha_2 = \dots = \alpha_a = 0$ 时, $SS_A/\sigma^2 \sim \chi^2(a-1)$; 而当 H_{0B} 成立, 即 $\beta_1 = \beta_2 = \dots = \beta_b = 0$ 时, $SS_B/\sigma^2 \sim \chi^2(b-1)$.

4.4 方差分析

证明请参阅文献[2].

对于假设检验问题(4.4-23), 取检验统计量为

$$F_A = \frac{SS_A/(a-1)}{SS_e/(n-a-b-1)},$$

则当 H_{0A} 成立时, $F_A \sim F(a-1, n-a-b+1)$, 而当 H_{0A} 不成立时, 由于

$$E(SS_A) = bE\left[\sum_{i=1}^a (\bar{X}_{i\cdot} - \bar{X})^2\right] = b\sum_{i=1}^a E[(\bar{X}_{i\cdot} - \bar{X})^2] = (a-1)\sigma^2 + b\sum_{i=1}^a \alpha_i^2,$$

所以 SS_A 有偏大的趋势, 从而由 $F_A > F_{1-\alpha}(a-1, n-a-b+1)$ (4.4-26)
 所确定的拒绝域给出了显著性水平 α 下的一个检验.

对于假设检验问题(4.4-24), 取检验统计量为 $F_B = \frac{SS_B/(b-1)}{SS_e/(n-a-b+1)}$,

则当 H_{0B} 成立, 即 $\beta_1 = \beta_2 = \dots = \beta_b = 0$ 时, $F_B \sim F(b-1, n-a-b+1)$; 而当 H_{0B} 不成立时,
 由于

$$E(SS_B) = aE\left[\sum_{j=1}^b (\bar{X}_{\cdot j} - \bar{X})^2\right] = a\sum_{j=1}^b E[(\bar{X}_{\cdot j} - \bar{X})^2] = (b-1)\sigma^2 + a\sum_{j=1}^b \beta_j^2,$$

所以 SS_B 有偏大的趋势, 从而由

$$F_B > F_{1-\alpha}(b-1, n-a-b+1) \quad (4.4-27)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

4.4 方差分析

在具体计算时,常用方差分析表 4.4-3.

表 4.4-3

方差来源	平方和	自由度	均方和	F 值
因子 A	$SS_A = b \sum_{i=1}^a (\bar{X}_{i\cdot} - \bar{X})^2$	$a - 1$	$MS_A = \frac{SS_A}{a - 1}$	$F_A = \frac{MS_A}{MS_e}$
因子 B	$SS_B = a \sum_{j=1}^b (\bar{X}_{\cdot j} - \bar{X})^2$	$b - 1$	$MS_B = \frac{SS_B}{b - 1}$	$F_B = \frac{MS_B}{MS_e}$
误差	$SS_e = \sum_{i=1}^a \sum_{j=1}^b (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2$	$n - a - b + 1$	$MS_e = \frac{SS_e}{n - a - b + 1}$	
总和	$SS = \sum_{i=1}^a \sum_{j=1}^b (X_{ij} - \bar{X})^2$	$n - 1$		

例 2 某型号火箭使用了四种燃料、三种推进器做射程试验. 每种燃料与每种推进器搭配做一次试验, 测得的火箭射程(单位: km)如下:

射程 推进器 B 燃料 A	B_1	B_2	B_3	$\bar{x}_{\cdot j}$
A_1	158.2	156.2	165.3	159.90
A_2	149.1	154.1	151.6	151.60
A_3	160.1	170.9	139.2	156.73
A_4	175.8	158.2	148.7	160.90
$\bar{x}_{\cdot j}$	160.80	159.88	151.20	$\bar{x} = 157.28$



4.4 方差分析

问在显著性水平 $\alpha=0.05$ 下, 燃料与推进器对射程是否有显著影响?

解 由所给数据列出方差分析表 4.4-4 ($a=4, b=3, n=12$) 如下:

表 4.4-4

方差来源	平方和	自由度	均方和	F 值
因子 A	157.59	3	52.53	0.43
因子 B	223.85	2	111.93	0.92
误差	731.98	6	121.99	
总和	1113.42	11		

对于假设检验 H_{0A} , 查表得临界值 $F_{0.95}(3, 6) = 4.76$. 显然, $0.43 < 4.76$, 因此不能拒绝 H_{0A} , 即各种燃料的差异对火箭射程的影响不显著. 对于假设检验 H_{0B} , 查表得临界值 $F_{0.95}(2, 6) = 5.14$. 显然, $0.92 < 5.14$, 因此也不能拒绝 H_{0B} , 即各种推进器的差异对火箭射程的影响并不显著.

本例中误差的均方和 $MS_e = \frac{SS_e}{n-a-b+1} = \frac{731.98}{6} = 121.99$ 比因子 A、B 的均方和

大, 因而 F_A, F_B 较小, 从而不能拒绝 H_{0A}, H_{0B} . 但是, 误差均方和 MS_e 是 σ^2 的无偏估计, 通常不能太大, 而在本例竟出现较大的值, 这可能是没有考虑因子 A 与 B 的不同水平搭配所引起的交互作用的缘故. 从所测得的数据看出, A_4 与 B_1 的搭配或 A_3 与 B_2 的搭配, 其效果很可能最佳, 而 A_3 与 B_3 的搭配效果可能最差. 但由于本例对每一对水平搭配都只做一次试验, 所以不能分辨出交互作用.



4.4 方差分析

现在讨论有交互作用的双因子方差分析模型(4.4-6). 设有两个因子A和B, 因子A取 a 个不同水平, 因子B取 b 个不同水平, 对这两个因子的每一对水平搭配(A_i, B_j), 有容量为 $m (> 1)$ 的样本 $(X_{ij1}, X_{ij2}, \dots, X_{ijm})$, $i = 1, 2, \dots, a, j = 1, 2, \dots, b$. 记

$$\bar{X}_{ij\cdot} = \frac{1}{m} \sum_{k=1}^m X_{ijk}, \quad i = 1, 2, \dots, a, j = 1, 2, \dots, b;$$

$$\bar{X}_{i\cdot\cdot} = \frac{1}{bm} \sum_{j=1}^b \sum_{k=1}^m X_{ijk} = \frac{1}{b} \sum_{j=1}^b \bar{X}_{ij\cdot}, \quad i = 1, 2, \dots, a;$$

$$\bar{X}_{\cdot j\cdot} = \frac{1}{am} \sum_{i=1}^a \sum_{k=1}^m X_{ijk} = \frac{1}{a} \sum_{i=1}^a \bar{X}_{ij\cdot}, \quad j = 1, 2, \dots, b;$$

$$\bar{X} = \frac{1}{abm} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^m X_{ijk} = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b \bar{X}_{ij\cdot} = \frac{1}{a} \sum_{i=1}^a \bar{X}_{i\cdot\cdot} = \frac{1}{b} \sum_{j=1}^b \bar{X}_{\cdot j\cdot},$$

并记 $n = abm$. 要分别检验假设

$$H_{0A}: \alpha_1 = \alpha_2 = \dots = \alpha_a = 0; \tag{4.4-28}$$

$$H_{0B}: \beta_1 = \beta_2 = \dots = \beta_b = 0; \tag{4.4-29}$$

$$H_{0A \times B}: \gamma_{11} = \dots = \gamma_{1b} = \dots = \gamma_{a1} = \dots = \gamma_{ab} = 0. \tag{4.4-30}$$

记统计量 $SS = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^m (X_{ijk} - \bar{X})^2$, $SS_A = bm \sum_{i=1}^a (\bar{X}_{i\cdot\cdot} - \bar{X})^2$,

$$SS_B = am \sum_{j=1}^b (\bar{X}_{\cdot j\cdot} - \bar{X})^2, \quad SS_{A \times B} = m \sum_{i=1}^a \sum_{j=1}^b (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2,$$

4.4 方差分析

$$SS_e = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^m (X_{ijk} - \bar{X}_{ij.})^2,$$

其中 SS 反映全体数据中的波动, 称为总离差平方和; SS_A 、 SS_B 分别反映因子 A 、 B 在各个水平下的不同作用而在数据中引起的波动, 分别称为因子 A 、 B 的离差平方和; SS_e 反映由于随机误差的作用而在数据中引起的波动, 称为误差平方和; $SS_{A \times B}$ 反映由于交互效应的存在而在数据中引起的波动, 称为交互效应的离差平方和.

与无交互作用的双因子方差分析类似, 可以证明下述的平方和分解公式及有关的定理.

引理 3(平方和分解公式) 在有交互作用的双因子方差分析模型(4.1-6)式中,

$$SS = SS_A + SS_B + SS_{A \times B} + SS_e.$$

定理 4.4-3 在有交互作用的双因子方差分析模型(4.4-6)中, SS_A 、 SS_B 、 $SS_{A \times B}$ 、 SS_e 相互独立, 且 $SS_e/\sigma^2 \sim \chi^2(n-ab)$. 当 H_{0A} 成立, 即 $\alpha_1 = \alpha_2 = \dots = \alpha_a = 0$ 时, $SS_A/\sigma^2 \sim \chi^2(a-1)$; 当 H_{0B} 成立, 即 $\beta_1 = \beta_2 = \dots = \beta_b = 0$ 时, $SS_B/\sigma^2 \sim \chi^2(b-1)$; 当 $H_{0A \times B}$ 成立, 即 $\gamma_{11} = \dots = \gamma_{1b} = \dots = \gamma_{a1} = \dots = \gamma_{ab} = 0$ 时, $SS_{A \times B}/\sigma^2 \sim \chi^2((a-1)(b-1))$.

证明请参阅文献[2].

对于假设检验问题(4.4-28), 取检验量为

$$F_A = \frac{SS_A/(a-1)}{SS_e/(n-ab)},$$

则当 H_{0A} 成立时, $F_A \sim F(a-1, n-ab)$; 而在 H_{0A} 不成立时, 由于

4.4 方差分析

$$SS_e = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^m (X_{ijk} - \bar{X}_{ij.})^2,$$

其中 SS 反映全体数据中的波动, 称为总离差平方和; SS_A, SS_B 分别反映因子 A, B 在各个水平下的不同作用而在数据中引起的波动, 分别称为因子 A, B 的离差平方和; SS_e 反映由于随机误差的作用而在数据中引起的波动, 称为误差平方和; $SS_{A \times B}$ 反映由于交互效应的存在而在数据中引起的波动, 称为交互效应的离差平方和.

与无交互作用的双因子方差分析类似, 可以证明下述的平方和分解公式及有关的定理.

引理 3(平方和分解公式) 在有交互作用的双因子方差分析模型(4.1-6)式中,

$$SS = SS_A + SS_B + SS_{A \times B} + SS_e.$$

定理 4.4-3 在有交互作用的双因子方差分析模型(4.4-6)中, $SS_A, SS_B, SS_{A \times B}, SS_e$ 相互独立, 且 $SS_e/\sigma^2 \sim \chi^2(n-ab)$. 当 H_{0A} 成立, 即 $\alpha_1 = \alpha_2 = \dots = \alpha_a = 0$ 时, $SS_A/\sigma^2 \sim \chi^2(a-1)$; 当 H_{0B} 成立, 即 $\beta_1 = \beta_2 = \dots = \beta_b = 0$ 时, $SS_B/\sigma^2 \sim \chi^2(b-1)$; 当 $H_{0A \times B}$ 成立, 即 $\gamma_{11} = \dots = \gamma_{1b} = \dots = \gamma_{a1} = \dots = \gamma_{ab} = 0$ 时, $SS_{A \times B}/\sigma^2 \sim \chi^2((a-1)(b-1))$.

证明请参阅文献[2].

对于假设检验问题(4.4-28), 取检验量为

$$F_A = \frac{SS_A/(a-1)}{SS_e/(n-ab)},$$

则当 H_{0A} 成立时, $F_A \sim F(a-1, n-ab)$; 而在 H_{0A} 不成立时, 由于

$$E(SS_A) = bmE\left[\sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2\right] = bm \sum_{i=1}^a E[(\bar{X}_{i..} - \bar{X})^2] = (a-1)\sigma^2 + bm \sum_{i=1}^a \alpha_i^2,$$



4.4 方差分析

所以 SS_A 有偏大的趋势. 因此, 由

$$F_A > F_{1-\alpha}(a-1, n-ab) \quad (4.4-31)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

对于假设检验问题(4.4-29), 取检验量为 $F_B = \frac{SS_B/(b-1)}{SS_e/(n-ab)}$,

则当 H_{0B} 成立时, $F_B \sim F(b-1, n-ab)$; 而在 H_{0B} 不成立时, 由于

$$E(SS_B) = amE\left[\sum_{j=1}^b (\bar{X}_{\cdot j} - \bar{X})^2\right] = am \sum_{j=1}^b E[(\bar{X}_{\cdot j} - \bar{X})^2] = (b-1)\sigma^2 + am \sum_{j=1}^b \beta_j^2,$$

所以 SS_B 有偏大的趋势. 因此, 由

$$F_B > F_{1-\alpha}(b-1, n-ab)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

对于假设检验问题(4.4-30), 取检验统计量为

$$F_{A \times B} = \frac{SS_{A \times B}/(a-1)(b-1)}{SS_e/(n-ab)},$$

则当 $H_{0A \times B}$ 成立时, $F_{A \times B} \sim F((a-1)(b-1), n-ab)$; 而在 $H_{0A \times B}$ 不成立时, 由于

$$E(SS_{A \times B}) = (a-1)(b-1)\sigma^2 + m \sum_{i=1}^a \sum_{j=1}^b \gamma_{ij}^2,$$

所以 $SS_{A \times B}$ 有偏大的趋势. 因此, 由

4.4 方差分析

$$F_{A \times B} > F_{1-\alpha}((a-1)(b-1), n-ab)$$

所确定的拒绝域给出了显著性水平 α 下的一个检验.

在具体计算时, 常用有交互作用的双因子方差分析表 4.4-5.

表 4.4-5 有交互作用的双因子方差分析表

方差来源	平方和	自由度	均方和	F 值
因子 A	$SS_A = bm \sum_{i=1}^a (\bar{X}_{i..} - \bar{X})^2$	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_A = \frac{MS_A}{MS_e}$
因子 B	$SS_B = am \sum_{j=1}^b (\bar{X}_{.j.} - \bar{X})^2$	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$F_B = \frac{MS_B}{MS_e}$
交互效应	$SS_{A \times B} = m \sum_{i=1}^a \sum_{j=1}^b (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2$	$(a-1)(b-1)$	$MS_{A \times B} = \frac{SS_{A \times B}}{(a-1)(b-1)}$	$F_{A \times B} = \frac{MS_{A \times B}}{MS_e}$
A \times B 误差	$SS_e = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^m (X_{ijk} - \bar{X}_{ij.})^2$	$n-ab$	$MS_e = \frac{SS_e}{n-ab}$	
总和	$SS = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^m (X_{ijk} - \bar{X})^2$	$n-1$		

例 3 在例 2 中, 对于燃料与推进器的每一对水平搭配, 各发射火箭两次, 测得射程(单位: km)如下:

4.4 方差分析

推进器 燃料	B_1	B_2	B_3	$\bar{x}_{i..}$
A_1	158.2	152.6	165.3	155.72
	152.6	141.2	160.8	
A_2	149.1	154.1	151.6	149.42
	142.8	150.5	148.4	
A_3	160.1	170.9	139.2	157.07
	158.3	173.2	140.7	
A_4	175.8	158.2	148.7	157.77
	171.5	151.0	141.4	
$\bar{x}_{..j..}$	158.55	156.91	149.51	$\bar{x} = 154.99$

在显著性水平 $\alpha=0.05$ 下, 分别检验各种燃料、各种推进器、交互效应是否对火箭射程有显著性影响.

解 按所给数据列出方差分析表 4.4-6 ($a=4, b=3, m=2, n=abm=24$):

表 4.4-6 方差分析表

方差来源	平方和	自由度	均方和	F 值
因子 A	261.68	3	87.23	4.42
因子 B	370.98	2	185.49	9.39
交互效应 $A \times B$	1768.69	6	294.78	14.93
误差	236.95	12	19.75	
总和	2638.30	23		



4.4 方差分析

对于假设检验(4.4-28),查表得临界值 $F_{0.95}(3, 12) = 3.49$. 显然, $4.42 > 3.49$, 故拒绝 H_{0A} , 即可以认为不同燃料对火箭射程有显著差异.

对于假设检验(4.4-29),查表得临界值 $F_{0.95}(2, 12) = 3.89$. 显然, $9.39 > 3.89$, 故拒绝 H_{0B} , 即可以认为不同推进器对火箭射程有显著差异.

对于假设检验(4.4-30),查表得临界值 $F_{0.95}(6, 12) = 3.00$. 显然, $14.93 > 3.00$, 故拒绝 $H_{0A \times B}$, 即可以认为交互效应显著.

由于交互效应 $A \times B$ 的 F 值为 14.93, 与因子 A, B 的 F 值相比要大得多, 所以本例的交互作用特别显著, 也就是说要注意燃料与推进器的搭配. ■

值得指出的是, 实际应用中一个重要问题是进一步求出两个因子的最佳水平搭配, 我们将在 4.5 节中讨论这个问题.

4.5 正交试验设计及其应用

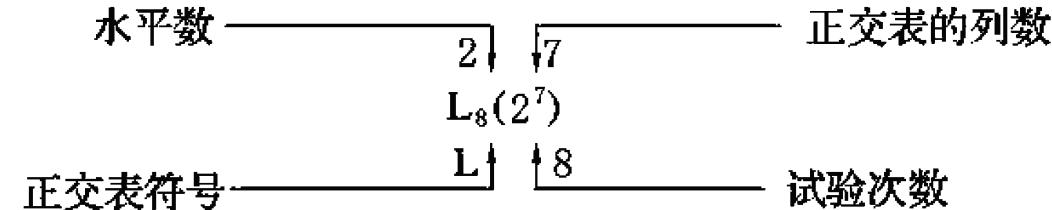
试验设计是数理统计学的一个分支,它主要研究如何收集数据以供统计推断应用. 正交试验设计是最常用的一类试验设计方法,其利用一套现成的规格化的表——正交表,科学地挑选试验条件和合理地安排试验. 它的主要优点是,能在很多的试验条件下选出代表性强的少数几个试验条件,并能通过较少次数的试验,分析推断出最好的情况.

本节通过几个例子介绍正交表及正交试验设计的应用.

1. 正交表

正交表(见附表五)是正交试验设计中安排试验,并对试验结果数据进行统计分析的重要工具.

正交表符号的含义如下:



以正交表符号 $L_8(2^7)$ 为例说明正交表的基本结构. 其中:①“L”是正交表符号;②“8”表示这张正交表有 8 行, 使用这张正交表需要安排 8 次试验;③“2”表示这张正交表适用于每个因子安排 2 个水平, 表中的数字 1 与 2 分别代表两个不同的水平;④“7”表示这张正交表有 7 列, 每列可以安排一个因子或一种交互作用.

仔细观察正交表 $L_8(2^7)$ (见附表五(2)), 发现它有以下两个特点:

4.5 正交试验设计及其应用

1) 表中任何一列,其所含各个水平的个数都相同,亦即不同数字出现的次数相等. 在 $L_8(2^7)$ 的每列中,数字 1 与 2 出现的次数都是 4;

2) 表中任何两列,所有各种可能的数对出现的次数都相同. 在 $L_8(2^7)$ 的任意两列中, 1 与 2 两个数字的各种可能数对 $(1,1), (1,2), (2,1), (2,2)$ 各出现两次.

附表五收录了一些最常用的正交表:二水平正交表 $L_4(2^3)$ 、 $L_8(2^7)$ 、 $L_{16}(2^{15})$ 与三水平正交表 $L_9(3^4)$ 、 $L_{27}(3^{13})$.

每张正交表都附有一张两列间交互作用表,它是用来说明如何安放因子之间交互作用的. 以 $L_8(2^7)$ 的两列间交互作用表(见附表五(2))为例加以说明. 假定要求考察三个因子 A, B, C . 如果因子 A 放在 $L_8(2^7)$ 的第 1 列, 因子 B 放在 $L_8(2^7)$ 的第 2 列, 那么交互作用 $A \times B$ 必须放在 $L_8(2^7)$ 的第 3 列. 这个 3 是由 $L_8(2^7)$ 的两列间交互作用表中两个因子列号所对应的数字给出的. 类似地,若因子 A 放在 $L_8(2^7)$ 的第 7 列,因子 B 放在 $L_8(2^7)$ 的第 3 列,则由 $L_8(2^7)$ 的两列间交互作用表给出数字 4,故交互作用 $A \times B$ 必须放在 $L_8(2^7)$ 的第 4 列.

在实际问题中,三个(或更多个)因子之间的交互作用常常可以忽略不计;任意两个因子之间的交互作用也不必都去考察,而只需选取一些应该重视的交互作用加以考虑.

2. 无交互作用的正交设计及其方差分析

我们用一个例子来说明如何选用正交表安排试验,并作相应的统计分析.

例 1 为了提高某种化工产品的得率,选择有关的三个因素(即因子)做试验,以确定一个较好的方案.这三个因素及各个水平如下:

4.5 正交试验设计及其应用

因子 \ 水平	1	2	3
反应温度 A(℃)	80	85	90
反应时间 B(分)	90	120	150
用碱量 C(%)	5	6	7

这个问题的具体处理方法按下列步骤进行.

1) 选正交表. 由于每个因子取三种水平, 所以在三水平正交表中选取. 一般应尽可能选 L 右下角表示试验次数的数字较小的表. 现取 $L_9(3^4)$, 这个表最多可安排四个因子, 且表明只要安排 9 次试验. 如果不用正交表, 则由于三个因子的所有可能的水平组搭配共有 $3^3 = 27$ 种, 所以一般要安排 27 次试验. 因此使用正交表大大节省了试验次数, 并且因子越多、水平数越多, 节省的试验次数也越多.

2) 表头设计. 把各个因子安放在正交表 $L_9(3^4)$ 的各列上端. 现只有三个因子, 且无交互作用, 故可以把三个因子放在正交表 $L_9(3^4)$ 的四列中任意三列的上端. 例如把因子 A、B、C 分别安放在前三列上. 这样, 表头设计为

因子	A	B	C	空列
列号	1	2	3	4



4.5 正交试验设计及其应用

其中不安放因子(及交互作用)的列称为空列,空列在今后的统计分析中将发挥重要的作用,一般要求至少有一个空列.如果没有空列,那么应该在每一组水平搭配下做重复试验.

3) 制订试验方案.把每个因子中的三个水平与正交表 $L_9(3^4)$ 中代表水平的数字 1、2、3 建立一一对应关系.为了避免系统误差,可以采用抽签或查随机数表的办法来决定对应关系.这里假定水平 A_i, B_i, C_i 用数字 i 表示, $i=1, 2, 3$.从而建立试验方案表(暂时抛开最后一列)如表 4.5-1.

试验方案表给出了要做的 9 次试验应该采取的水平搭配.例如试验号 5 这一行上的三个数字依次为 2、2、3,这表示第 5 号试验的水平搭配为 $A_2B_2C_3$,即反应温度取 85℃、反应时间取 120 分钟、用碱量取 7%.

表 4.5-1 试验方案表

因子 列号 试验号	A(温度)	B(时间)	C(用碱量)	得率(%)
	1	2	3	试验结果 y_i
1	1(80℃)	1(90分钟)	1(5%)	31
2	1	2(120分钟)	2(6%)	54
3	1	3(150分钟)	3(7%)	38
4	2(85℃)	1	2	53
5	2	2	3	49
6	2	3	1	42
7	3(90℃)	1	3	57
8	3	2	1	62
9	3	3	2	64

4.5 正交试验设计及其应用

4) 按规定的方案做试验, 记录下试验结果 y_1, y_2, \dots, y_9 , 标在试验方案表的最后一列上.

5) 计算各个统计量的观测值. 先定义有关的统计量(或其观测值). 假定每个因子取 r 种不同的水平, 每种水平在试验方案中出现了 m 次, 那么总的试验次数(即所用正交表的行数) $n=rm$, 设试验结果为 Y_1, Y_2, \dots, Y_n . 对所用正交表的第 j 列(包括空列), 令 $K_{jl} =$ 第 j 列中相应于水平 l 的 m 个试验结果之和, $l = 1, 2, \dots, r$.

例如, 现在 $r=3, m=3, n=rm=9$; 则 $K_{11}=y_1+y_2+y_3, K_{23}=y_3+y_6+y_9$ 等. 记

$$K = \sum_{l=1}^r K_{jl}, \quad (4.5-1)$$

易见, K 是全体试验结果的和 $\sum_{i=1}^n Y_i$, 因而与 j 无关. 若令

$$P = \frac{1}{n} K^2, \quad (4.5-2)$$

$$Q_j = \frac{1}{m} \sum_{l=1}^r K_{jl}^2, \quad S_j = Q_j - P, \quad (4.5-3)$$

$$Q = \sum_{i=1}^n Y_i^2, \quad S_T = Q - P, \quad (4.5-4)$$

则可以证明

$$S_T = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_j S_j, \quad (4.5-5)$$

其中 $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} K$. 例如, 现在 $L_9(3^4)$ 中有 4 列, 因此 $S_T = S_1 + S_2 + S_3 + S_4$.

按所给数据算出上述统计量的观测值后, 列出计算表 4.5-2.

4.5 正交试验设计及其应用

6) 方差分析. 首先推断每个因子在各个不同水平下的效应是否有显著性差异, 即要分别检验

$$H_{0A}: \alpha_1 = \alpha_2 = \alpha_3 = 0,$$

$$H_{0B}: \beta_1 = \beta_2 = \beta_3 = 0,$$

$$H_{0C}: \delta_1 = \delta_2 = \delta_3 = 0,$$

表 4.5-2

因子 列号 试验号	A	B	C	空列	y_i
	1	2	3	4	
1	1	1	1	1	31
2	1	2	2	2	54
3	1	3	3	3	38
4	2	1	2	3	53
5	2	2	3	1	49
6	2	3	1	2	42
7	3	1	3	2	57
8	3	2	1	3	62
9	3	3	2	1	64
K_{j1}	123	141	135	144	$K=450$
K_{j2}	144	165	171	153	$P=22500$
K_{j3}	183	144	144	153	
Q_j	23118	22614	22734	22518	$Q=23484$
S_j	618	114	234	18	$S_T=984$



4.5 正交试验设计及其应用

其中 α_j 表示因子 A 在第 j 种水平下的效应, β_j 表示因子 B 在第 j 种水平下的效应, δ_j 表示因子 C 在第 j 种水平下的效应, $j=1, 2, 3$. 记 μ 为总平均效应. 类似于无交互作用的双因子方差分析模型(4.4-22), 可以建立如下的统计模型:

$$\left\{ \begin{array}{l} Y_1 = \mu + \alpha_1 + \beta_1 + \delta_1 + \varepsilon_1, \\ Y_2 = \mu + \alpha_1 + \beta_2 + \delta_2 + \varepsilon_2, \\ Y_3 = \mu + \alpha_1 + \beta_3 + \delta_3 + \varepsilon_3, \\ Y_4 = \mu + \alpha_2 + \beta_1 + \delta_2 + \varepsilon_4, \\ Y_5 = \mu + \alpha_2 + \beta_2 + \delta_3 + \varepsilon_5, \\ Y_6 = \mu + \alpha_2 + \beta_3 + \delta_1 + \varepsilon_6, \\ Y_7 = \mu + \alpha_3 + \beta_1 + \delta_3 + \varepsilon_7, \\ Y_8 = \mu + \alpha_3 + \beta_2 + \delta_1 + \varepsilon_8, \\ Y_9 = \mu + \alpha_3 + \beta_3 + \delta_2 + \varepsilon_9, \\ \sum_{j=1}^3 \alpha_j = \sum_{j=1}^3 \beta_j = \sum_{j=1}^3 \delta_j = 0, \end{array} \right. \quad (4.5-6)$$

其中 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_9$ 是独立同分布的随机变量, 且每一个 ε_i 服从 $N(0; \sigma^2)$.

方差分析中因子的离差平方和恰是该因子在表头设计中所占那一列的相应的 S_j , 且自由度为该因子所取的水平数减 1(即 $r-1$). 现在

$$SS_A = S_1, \quad SS_B = S_2, \quad SS_C = S_3, \quad (4.5-7)$$



4.5 正交试验设计及其应用

自由度都是 $3-1=2$. 总离差平方和

$$SS = S_T, \quad (4.5-8)$$

且自由度为总的试验次数减1(即 $n-1$). 现在自由度是 $9-1=8$. 空列所对应的 S_i 之和恰是误差平方和 SS_e , 且自由度为 S_T 的自由度减去诸因子离差平方和的自由度之和. 因而现在

$$SS_e = S_4, \quad (4.5-9)$$

且自由度为 $8-(2+2+2)=2$.

值得指出的是, 在有重复试验的情况下, 上述计算离差平方和的公式与计算自由度的法则仍然有效.

现列出方差分析表, 如表 4.5-3 所示.

表 4.5-3 无交互作用三因子方差分析表

方差来源	平方和	自由度	均方和	F 值	显著性
因子 A	$SS_A = S_1 = 618$	$a-1=2$	$MS_A = \frac{SS_A}{a-1} = 309$	$F_A = \frac{MS_A}{MS_e} = 34.33$	*
因子 B	$SS_B = S_2 = 114$	$b-1=2$	$MS_B = \frac{SS_B}{b-1} = 57$	$F_B = \frac{MS_B}{MS_e} = 6.33$	
因子 C	$SS_C = S_3 = 234$	$c-1=2$	$MS_C = \frac{SS_C}{c-1} = 117$	$F_C = \frac{MS_C}{MS_e} = 13.00$	(*)
误差	$SS_e = S_4 = 18$	$n-1-(a-1+b-1+c-1)=2$	$MS_e = \frac{SS_e}{2} = 9$		
总和	$SS = S_T = 984$	$n-1=8$			



4.5 正交试验设计及其应用

表中显著性一栏的含义是,符号“(*)”、“*”、“**”分别表示在显著性水平 $\alpha=0.10$ 、 $\alpha=0.05$ 、 $\alpha=0.01$ 下检验结果是拒绝原假设.一般,若在 $\alpha=0.10$ 下检验结果为不拒绝原假设,则可以认为该因子对所考察的指标无显著性影响.本例查表得到的临界值分别为

$$F_{0.90}(2,2) = 9.0, F_{0.95}(2,2) = 19, F_{0.99}(2,2) = 99.$$

7) 最终结论.从表 4.5-3 的最后第二列 F 值看出,因子 B 的作用不显著,即可以认为反应时间是影响该种化工产品得率的次要因素.因子 A 的作用显著,因子 C 的作用比较显著.再由计算表 4.5-2 看出,对因子 A 来说有

$$K_{13} > K_{12} > K_{11},$$

而对因子 C 来说有

$$K_{32} > K_{33} > K_{31},$$

因此较优的水平是 A_3 与 C_2 .为了节约时间,不妨采用反应时间为 90 分钟,即 B_1 ,从而本例中较好的因子水平搭配是 $A_3B_1C_2$,亦即采用反应温度为 90℃、反应时间为 90 分钟、用碱量为 6% 这一方案来组织生产.顺便指出,这组水平搭配不在我们做过的 9 个试验之内,这并不奇怪,因为正交试验仅仅对因子的所有水平搭配选做了一部分. ■

3. 有交互作用的正交设计及其方差分析

我们也通过例子来说明如何对有交互作用的方差分析问题应用正交表及相应的两列间交互作用表作正交设计并进行相应的统计分析.

例 2 为了考察水稻品种、栽培规格与施氮肥量这三个因素及它们的交互作用对水稻产量的影响,对每个因子各取两个水平进行正交试验,各个因子水平如下:

4.5 正交试验设计及其应用

因子 A, 水稻品种, A_1 : 铁大, A_2 : 双广;

因子 B, 栽培规格(单位: 厘米 \times 厘米), B_1 : 15×12 , B_2 : 15×15 ;

因子 C, 施氮肥量(单位: 公斤/亩), C_1 : 10, C_2 : 12.5.

这个问题的具体处理方法可按下列步骤依次进行(凡是无交互作用时的方法仍有效的地方不再重复):

1) 选正交表. 由于每个因子取两个水平, 所以应该在二水平正交表中选取. 一般, 三个因子之间的交互作用 $A \times B \times C$ 可以不考虑, 因此这里只考虑两个因子之间的交互作用, 又根据专业人员的经验, 水稻品种与栽培规格之间的交互作用不显著, 故仅考虑交互作用 $B \times C$ 与 $C \times A$. 这样, 所选的正交表至少应该有 5 列, 合适的正交表是 $L_8(2^7)$.

2) 表头设计. 若先把因子 C 放在第 1 列, 因子 B 放在第 2 列, 则由两列间的交互作用表查得, $B \times C$ 只能放在第 3 列. 然后把因子 A 放在第 4 列, 这时 $C \times A$ 必须放在第 5 列. 从而表头设计为

因子	C	B	$B \times C$	A	$C \times A$	空列	空列
列号	1	2	3	4	5	6	7

3) 制定试验方案. 类似于无交互作用的情形, 建立试验方案表(暂时抛开此表的最后一列), 见表 4.5-4.



4.5 正交试验设计及其应用

表 4.5-4 试验方案表

因子 列号 试验号	C(施氮肥量)	B(栽培规格)	A(水稻品种)	水稻亩产量 (公斤)
	1	2	4	试验结果 y_i
1	1(10 公斤/亩)	1(15 厘米×12 厘米)	1(铁大)	789.7
2	1	1	2(双优)	855.0
3	1	2(15 厘米×15 厘米)	1	800.9
4	1	2	2	858.0
5	2(12.5 公斤/亩)	1	1	955.8
6	2	1	2	756.0
7	2	2	1	890.7
8	2	2	2	681.0

4) 按规定的方案做试验. 记录下试验结果 y_1, y_2, \dots, y_8 , 标在试验方案表的最后一列上.

5) 计算各个统计量的观测值. 在无交互作用情形下给出的计算公式(4.5-1)~(4.5-5)依然有效,但在二水平情形($r=2$)下,我们有

$$S_j = \frac{1}{n} (K_{j1} - K_{j2})^2, \quad (4.5-10)$$

从而可以省略 Q_j 的计算. 现在, $r=2, m=4, n=rm=8$. 按所给数据列出计算表如表4.5-5 所示.

4.5 正交试验设计及其应用

6) 方差分析. 交互效应的离差平方和恰是该交互作用在表头设计中所占那一列的相应的 S_j , 且自由度为相应两个因子的自由度之积. 误差平方和 SS_e 的自由度仍用减法来计算. 现在,

$$SS_A = S_4, \quad SS_B = S_2, \quad SS_C = S_1,$$

且自由度都是 $2-1=1$; 交互效应的离差平方和为

$$SS_{B\times C} = S_3, \quad SS_{C\times A} = S_5,$$

其自由度都是 $1\times 1=1$. 总离差平方和为

$$SS = S_T,$$

表 4.5-5

因子 列号 试验号	C	B	B×C	A	C×A	空列	空列	亩产量/公斤
	1	2	3	4	5	6	7	y_i
1	1	1	1	1	1	1	1	789.7
2	1	1	1	2	2	2	2	855.0
3	1	2	2	1	1	2	2	800.9
4	1	2	2	2	2	1	1	858.0
5	2	1	2	1	2	1	2	955.8
6	2	1	2	2	1	2	1	756.0
7	2	2	1	1	2	2	1	890.7
8	2	2	1	2	1	1	2	681.0
K_{j1}	3303.6	3365.5	3216.4	3437.1	3027.6	3284.5	3294.4	$K=6587.1$
K_{j2}	3283.5	3230.6	3370.7	3150.0	3559.5	3302.6	3292.7	$P=5423735.8$
S_j	50.5	1981.35	2976.06	10303.33	5364.7	40.95	0.36	$S_T=50717.22$



4.5 正交试验设计及其应用

且自由度是 $8 - 1 = 7$. 误差平方和为

$$\begin{aligned} SS_e &= SS - (SS_A + SS_B + SS_C + SS_{B \times C} + SS_{C \times A}) \\ &= S_6 + S_7, \end{aligned}$$

即 SS_e 是空列所对应的 S_i 之和, 并且自由度是

$$7 - (1 + 1 + 1 + 1 + 1) = 2.$$

对于显著性水平 $\alpha = 0.10, 0.05$ 和 0.01 , 查表得到的临界值分别为

$$F_{0.90}(1, 2) = 8.53, F_{0.95}(1, 2) = 18.5, F_{0.99}(1, 2) = 98.5.$$

列出方差分析表 4.5-6.

7) 最终结论. 由表 4.5-6 看出, 因子 C 的作用不显著, 但交互效应 $B \times C, C \times A$ 却高度显著, 因此必须重视施氮肥量分别与水稻品种、栽培规格之间的交互作用. 于是分别研究这两种搭配的效应. 按所给的数据(见试验方案表 4.5-4)作因子 C 与因子 A 的搭配效应表如表 4.5-7.

4.5 正交试验设计及其应用

表 4.5-6 有交互作用的三因子方差分析表

方差来源	平方和	自由度	均方和	F 值	显著性
因子 A	$SS_A = S_4 = 10303.30$	$a-1=1$	$MS_A = \frac{SS_A}{a-1} = 10303.3$	$F_A = \frac{MS_A}{MS_e} = 498.7$	* *
因子 B	$SS_B = S_2 = 1981.35$	$b-1=1$	$MS_B = \frac{SS_B}{b-1} = 1981.35$	$F_B = \frac{MS_B}{MS_e} = 95.9$	*
因子 C	$SS_C = S_1 = 50.50$	$c-1=1$	$MS_C = \frac{SS_C}{c-1} = 50.5$	$F_C = \frac{MS_C}{MS_e} = 2.4$	
交互效应 $B \times C$	$SS_{B \times C} = S_3 = 2976.06$	$(b-1)(c-1)=1$	$MS_{B \times C} = \frac{SS_{B \times C}}{(b-1)(c-1)} = 2976.06$	$F_{B \times C} = \frac{MS_{B \times C}}{MS_e} = 144.0$	* *
交互效应 $C \times A$	$SS_{C \times A} = S_5 = 35364.70$	$(c-1)(a-1)=1$	$MS_{C \times A} = \frac{SS_{C \times A}}{(c-1)(a-1)} = 35364.7$	$F_{C \times A} = \frac{MS_{C \times A}}{MS_e} = 1711.7$	* *
误差	$SS_e = S_6 + S_7 = 41.31$	(用减法) 2	$MS_e = \frac{SS_e}{2} = 20.66$		
总和	$SS = S_T = 50717.22$	$n-1=7$			

表 4.5-7 A、C 因子搭配效应表

因子 C	因子 A	
	A_1 (铁大)	A_2 (双J)
C_1 (10 公斤/亩)	$\frac{1}{2}(y_1+y_3) = 795.3$	$\frac{1}{2}(y_2+y_4) = 856.5$
C_2 (12.5 公斤/亩)	$\frac{1}{2}(y_5+y_7) = 923.25$	$\frac{1}{2}(y_6+y_8) = 718.5$



4.5 正交试验设计及其应用

因此可以看出,搭配 C_2A_1 较好. 类似地,作因子 C 与因子 B 的搭配效应表如表 4.5-8.

表 4.5-8 B 与 C 因子搭配效应表

因子 C	因子 B	
	$B_1(15 \times 12)$	$B_2(15 \times 15)$
$C_1(10 \text{ 公斤/亩})$	$\frac{1}{2}(y_1 + y_2) = 822.35$	$\frac{1}{2}(y_3 + y_4) = 829.45$
$C_2(12.5 \text{ 公斤/亩})$	$\frac{1}{2}(y_5 + y_6) = 855.9$	$\frac{1}{2}(y_7 + y_8) = 785.85$

由此看出,搭配 B_1C_2 较好.

综合起来,本例中较好的因子水平搭配是 $A_1B_1C_2$,即采用铁大稻种、15 厘米 \times 12 厘米栽培规格和每亩施氮肥 12.5 公斤这一方案. ■

值得指出的是,本例中因子 A 的作用也是高度显著的. 由于 $K_{41} > K_{42}$, 所以较优的水平是 A_1 , 这与上述因子水平搭配的结论不矛盾. 但是,有时两者会发生矛盾,这时一般应该优先照顾按交互效应所选出的水平.

4. 多元回归的正交设计及其回归分析

在 4.3 节中已经说过,如果方阵 $L = X^T X$ 是对角矩阵,那么多元回归分析的计算特别简便. 因此,在 k 元回归分析问题

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \dots, n$$

中,如果矩阵

4.5 正交试验设计及其应用

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

中 $k+1$ 个 n 维列向量两两正交, 且诸 x_{ij} 仅取 1 或 -1 , 那么 $\mathbf{X}^T \mathbf{X} = n\mathbf{I}_{k+1}$. 从而(4.2-3)式给出 $\beta_0, \beta_1, \dots, \beta_k$ 的最小二乘估计为

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n Y_i = \bar{Y}, \quad \hat{\beta}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} Y_i, \quad j = 1, 2, \dots, k.$$

由于矩阵 \mathbf{X} 的第 1 列与其余各列正交, 因此有 $\sum_{i=1}^n x_{ij} = 0$, 即 $\bar{x}_j = 0, j = 1, 2, \dots, k$. 由此推得, 当 $j \neq m$ 时,

$$l_{jm} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m) = \sum_{i=1}^n x_{ij} x_{im} = 0, \\ j, m = 1, 2, \dots, k.$$

同时,
$$l_{jj} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = \sum_{i=1}^n x_{ij}^2 = n, \quad j = 1, 2, \dots, k.$$

于是, 由(4.3-2)式得到回归平方和

$$SS_r = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n \left[\sum_{j=1}^k (x_{ij} - \bar{x}_j) \hat{\beta}_j \right]^2 = \sum_{j=1}^k \sum_{i=1}^n \hat{\beta}_j \hat{\beta}_j l_{jj} = n \sum_{j=1}^k \hat{\beta}_j^2.$$

4.5 正交试验设计及其应用

又因为 $l^{(jj)} = \frac{1}{l_{jj}} = \frac{1}{n}$, $j=1, 2, \dots, k$, 所以由(4.3-9)式知, 偏回归平方和为

$$SS_{\beta_j} = \frac{\hat{\beta}_j^2}{l^{(jj)}} = n\hat{\beta}_j^2$$

因此,

$$SS_r = SS_{r1} + \dots + SS_{rk},$$

即回归平方和恰是 k 个偏回归平方和的和.

与方差分析问题相同, 称自变量为因子, 称因子的取值为水平. 对多元线性回归分析问题作正交设计时仅仅用到二水平正交表, 且用数字 1 与 -1 分别代表两个不同的水平.

下面举一个例子来说明, 如何对多元线性回归分析问题应用正交表作正交设计, 并作相应的统计分析.

例 3 为了提高硝基蒽醌中的某种含量 Y , 今通过试验来研究有关的三个因素(即因子)的影响, 并建立经验回归函数. 这三个因素的上、下水平(即因子变化范围的上、下限)如表 4.5-9:

表 4.5-9

因子	下水平	上水平
亚硫酸钠 z_1	5.0 克	9.0 克
大苏打 z_2	2.5 克	4.5 克
反应时间 z_3	1 小时	3 小时

为了得到因子变化范围是区间 $[-1, 1]$, 作变换

$$x_1 = \frac{1}{2}(z_1 - 7), \quad x_2 = z_2 - 3.5, \quad x_3 = z_3 - 2.$$



4.5 正交试验设计及其应用

经过变换得到的因子 x_j ($j=1, 2, 3$) 的下水平恒为 -1 , 上水平恒为 1 .

这个问题的具体处理方法按下列步骤依次进行:

1) 选正交表. 根据有关专业人员的意见, 必须重视交互作用, 但三个因子之间的交互作用可以忽略不计. 这样, 所选的正交表至少应该能安放因子 x_1, x_2, x_3 及交互作用 x_1x_2, x_1x_3, x_2x_3 . 因而选用 $L_8(2^7)$.

2) 表头设计. 我们将用经过变换后得到的因子 x_j 进行工作, 并按正交表的要求设计表头:

因子	x_1	x_2	x_1x_2	x_3	x_1x_3	x_2x_3	空列
列号	1	2	3	4	5	6	7

3) 制订试验方案. 建立试验方案表(暂时抛开最后两列)如表 4.5-10.

4) 按规定的方案做试验. 为了提高精度, 现在对每一号试验各重复做了两次, 得到 16 个数据, 并相应地标在试验方案表的最后两列上(为计算方便, 这些数据已分别减去了 85).

5) 计算各个统计量的观测值. 先定义有关的统计量(或其观测值). 假定我们共做了 r 号试验, 每号试验重复做了 m 次, 得到的试验结果为 $Y_{i1}, Y_{i2}, \dots, Y_{im}$, $i=1, 2, \dots, r$. 总的试验次数 $n=rm$. 例如, 本例中 $r=8, m=2, n=16$. 记

4.5 正交试验设计及其应用

表 4.5-10 试验方案表

试验号	因子 列号	x_1 (亚硫酸钠)	x_2 (大苏打)	x_3 (反应时间)	试验结果	
		1	2	4	y_{i1}	y_{i2}
1		1(9.0 克)	1(4.5 克)	1(3 小时)	5.98	8.73
2		1	1	-1(1 小时)	-0.46	2.67
3		1	-1(2.5 克)	1	2.70	6.46
4		1	-1	-1	0.60	3.50
5		-1(5.0 克)	1	1	0.40	1.01
6		-1	1	-1	-2.37	-1.12
7		-1	-1	1	0.50	-2.60
8		-1	-1	-1	-1.80	-1.45

$$Y_{i \cdot} = \sum_{k=1}^m Y_{ik}, \quad i = 1, 2, \dots, r, \quad (4.5-11)$$

$$K = \sum_{i=1}^r Y_{i \cdot} = \sum_{i=1}^r \sum_{k=1}^m Y_{ik}. \quad (4.5-12)$$

易见, K 是全体数据的和. 令

$$P = \frac{1}{n} K^2, \quad (4.5-13)$$



4.5 正交试验设计及其应用

$$Q_t = \frac{1}{m} \sum_{i=1}^r Y_{i\cdot}^2, \quad S_t = Q_t - P; \quad (4.5-14)$$

$$Q_T = \sum_{i=1}^r \sum_{k=1}^m Y_{ik}^2, \quad S_T = Q_T - P; \quad (4.5-15)$$

$$S_e = Q_T - Q_t, \quad (4.5-16)$$

可以证明

$$S_T = \sum_{i=1}^r \sum_{k=1}^m (Y_{ik} - \bar{Y})^2, \quad (4.5-17)$$

$$S_t = m \sum_{i=1}^r (\bar{Y}_{i\cdot} - \bar{Y})^2, \quad (4.5-18)$$

$$S_e = \sum_{i=1}^r \sum_{k=1}^m (Y_{ik} - \bar{Y}_{i\cdot})^2, \quad (4.5-19)$$

$$S_T = S_t + S_e, \quad (4.5-20)$$

其中 $\bar{Y} = \frac{1}{n} K$, $\bar{Y}_{i\cdot} = \frac{1}{m} Y_{i\cdot}$, $i=1, 2, \dots, r$. 并且, 在 $m=2$ 时, (4.5-19) 式可以简化为

$$S_e = \frac{1}{2} \sum_{i=1}^r (Y_{i1} - Y_{i2})^2. \quad (4.5-21)$$

为了方便, 在下面的计算表中添上第 0 列(也记作 x_0 列), 它的元素全是 1. 这样, 正交表便成为本段初所要求的矩阵 X . 对于第 j 列, 记

$$l_{jy} = \sum_{i=1}^r x_{ij} y_{i\cdot},$$



4.5 正交试验设计及其应用

$$\hat{\beta}_j = \frac{1}{n} l_{jy}, \quad j = 0, 1, 2, \dots,$$

$$S_j = n\hat{\beta}_j^2 = \hat{\beta}_j l_{jy}, \quad j = 1, 2, \dots,$$

可以证明

$$S_t = \sum_j S_j. \quad (4.5-22)$$

例如,本例中 $L_8(2^7)$ 有 7 列,因此 $S_t = \sum_{j=1}^7 S_j.$

按所给数据计算出有关的统计量的观测值后,列出计算表(暂时抛开最后两行)如表 4.5-11 所示,其中已将表头设计中的因子 x_3 与 x_1x_2 交换了顺序,从而列号 4 与 3 也作相应的变动.

6) 回归分析. 偏回归平方和

$$SS_{rj} = S_j, \quad j = 1, 2, \dots, 6,$$

且自由度都是 $2-1=1$; 回归平方和

$$SS_r = \sum_{j=1}^6 SS_{rj} = \sum_{j=1}^6 S_j,$$

其自由度是 6; 总离差平方和

$$SS = S_T,$$

4.5 正交试验设计及其应用

表 4.5-11

试验号	因子 列号	x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	空列	$y_{i1} + y_{i2}$ $= y_i.$	$ y_{i1} - y_{i2} $
		0	1	2	3	4	5	6	7		
1		1	1	1	1	1	1	1	1	14.71	2.75
2		1	1	1	-1	1	-1	-1	-1	2.21	3.13
3		1	1	-1	1	-1	1	-1	-1	9.16	3.76
4		1	1	-1	-1	-1	-1	1	1	4.10	2.90
5		1	-1	1	1	-1	-1	1	-1	1.41	0.61
6		1	-1	1	-1	-1	1	-1	1	-3.49	1.25
7		1	-1	-1	1	1	-1	-1	1	-2.10	3.10
8		1	-1	-1	-1	1	1	1	-1	-3.25	0.35
t_{jy}		22.75	37.61	6.93	23.61	0.39	11.51	11.19	3.69	$K=22.75$ $P=32.35$ $S_r=25.79$ $S_t=143.28$ $S_T=169.07$ $SS_e=26.64$	
$\hat{\beta}_j$		1.42	2.35	0.43	1.48	0.02	0.72	0.70	0.23		
S_j		88.38	2.98	34.94	0.01	8.29	7.83	0.85			
F_j		29.86	1.01	11.80	0.00	2.80	2.65				
显著性		**		**		(**)	(**)			** $(F_{0.99}(6,9)=5.8)$	

4.5 正交试验设计及其应用

且自由度是 $16 - 1 = 15$; 残差平方和

$$SS_e = SS - SS_r = S_7 + S_e,$$

其自由度是 $15 - 6 = 9$. 偏 F 检验统计量

$$F_j = \frac{SS_{rj}}{MS_e} = \frac{S_j}{SS_e/9}, \quad j = 1, 2, \dots, 6, \quad (4.5-23)$$

F 检验统计量为

$$F = \sum_{j=1}^6 F_j. \quad (4.5-24)$$

对于 $\alpha = 0.10, 0.05, 0.01$, 查表得到临界值分别为

$$F_{0.75}(1, 9) = 1.51, \quad F_{0.90}(1, 9) = 3.36,$$

$$F_{0.95}(1, 9) = 5.12, \quad F_{0.99}(1, 9) = 10.6.$$

在计算表 4.5-11 的显著性一行中, 除了前述的符号“(*)”, “*”, “**”外, 还用符号“(**)”表示在显著性水平 $\alpha = 0.25$ 下检验结果是拒绝原假设.

7) 最终结论. 由计算表 4.5-11 看出, 交互效应 x_1x_2 的作用不显著, 故可以把它从回归函数中剔除出去. 但交互效应 x_1x_3, x_2x_3 还是有一定的显著作用, 因此将它们保留在回归函数中. 因子 x_1 及 x_3 的作用高度显著, 而因子 x_2 本身虽然没有显著作用, 但考虑到交互效应 x_2x_3 有一定的显著性, 所以仍把 x_2 保留在回归函数之中. 于是, 我们得到的经验回归函数是



4.5 正交试验设计及其应用

$$y - 85 = 1.42 + 2.35x_1 + 0.43x_2 + 1.48x_3 + 0.72x_1x_3 \\ + 0.70x_2x_3.$$

若用原始因子 z_1, z_2, z_3 表示，则经验回归函数为

$$y = 83.68 + 0.455z_1 - 0.97z_2 - 3.49z_3 + 0.36z_1z_3 \\ + 0.70z_2z_3.$$

这个经验回归函数大致上反映了亚硫酸钠、大苏打和反应时间三个因素与硝基蒽醌中的某种含量之间的相关关系。当然，这样建立的经验回归函数只是表明在因子的上、下水平处与试验结果拟合得较好，至于区间内部拟合得如何是不清楚的。

作业 (第1部分)

习题 4

2. 今有 10 组观测数据如下

x_i	0.5	-0.8	0.9	-2.8	6.5	2.3	1.6	5.1	-1.9	-1.5
y_i	-0.3	-1.2	1.1	-3.5	4.6	1.8	0.5	2.8	-2.8	0.5

应用正态线性模型 $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, $i=1, 2, \dots, 10$, 且 $\epsilon_1, \epsilon_2, \dots, \epsilon_{10}$ 相互独立,

- (1) 求 β_0, β_1 的最小二乘估计;
- (2) 求 β_1 的置信水平为 0.95 的区间估计;
- (3) 在显著性水平 $\alpha=0.01$ 下, 检验假设 $H_0: \beta_1 = 0$;
- (4) 计算残差方差 $\hat{\sigma}_e^2$;
- (5) 求 $x=1.2$ 时 Y 的双侧 95% 的预测区间.

3. 养猪场为了估算猪的毛重(单位:公斤) Y 与其身长(单位:厘米) x_1 , 肚围(单位:厘米) x_2 之间的关系, 测量了 14 头猪, 得数据如下:

x_{i1}	41	45	51	52	59	62	69	72	78	80	90	92	98	103
x_{i2}	49	58	62	71	62	74	71	74	79	84	85	94	91	95
y_i	28	39	41	44	43	50	51	57	63	66	70	76	80	84



作业 (第2部分)

- (1) 求经验回归函数;
- (2) 在显著性水平 1% 下, 检验 $H_0: \beta_1 = \beta_2 = 0$;
- (3) 求 $x_1 = 100, x_2 = 80$ 时 Y 的预测值;
- (4) 在显著性水平 5% 下, 作偏 F 检验,

$$H_{0j}: \beta_j = 0, j = 1, 2.$$

这里, 假定猪的毛重 $Y \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$.

7. 下表给出了某种化工产品在三种不同浓度(单位: %)与四种不同温度(单位: °C)下成品的得率, 且每对水平搭配各做了两次试验的数据:

浓度 \ 温度	10	24	38	52
2	10, 14	11, 11	13, 9	10, 12
4	9, 7	10, 8	7, 11	6, 10
6	5, 11	13, 14	12, 13	14, 10

假定数据来自方差相等的正态总体.

- (1) 证明, 在显著性水平 25% 下, 浓度与温度之间的交互效应对该种化工产品的得率无显著影响.
- (2) 问在显著性水平 5% 下, 浓度与温度分别对该种化工产品的得率有无显著影响?

謝謝觀看！



廈門大學
XIAMEN UNIVERSITY



信息學院
(国家示范性软件学院)
School of Informatics

黃 烽
博士·副教授
Dr. Wei Huang