第十章　线性回归分析和方差分析

本章中, 我们将讨论线性统计推断的两个基本问题: 线性回归分析和方差分析.

§10.1　线性回归分析

在实际工作中, 遇到的两个变量通常有相互制约、相互关联的关系. 这些关系又可分为两大类: 一类是确定性关系, 如理想状态下匀速直线运动中时间 t 与距离 s 的关系; 另一类是相关关系, 如中老年健康调查中年龄 x 与血压 Y 的关系, 水稻的单位面积施肥量 x 与单位面积产量 Y 的关系等.

当 x 与 Y 存在相关关系时, 最常见、最普通的这种相关关系是线性相关关系, 即 x 与 Y 的相关关系有线性趋势. 围绕线性相关关系的线性趋势展开的统计分析叫做线性回归分析.

"回归" 这个概念是英国统计学家高尔顿在研究人类遗传规律时提出来的. 如大多数身高 175 cm 的父亲, 其儿子的身高虽各有不同, 但一般总是与父亲及其家族的平均身高相近, 即儿子的身高在总体上有一种 "回归" 到父亲及其家族平均身高的趋势. 回归概念产生以后, 被广泛应用于各个领域, 并成为研究变量之间相关关系的一种有效方法.

线性回归分析中, 通常要求 x 是普通变量, 是可以控制的变量; Y 是随机变量, 是可观测但不可控制的变量. 如前面所述 x 是施肥量, 是可控制的, 而产量 Y 是可观测但不可控制的.

§10.1.1　线性回归模型

设 x 是普通变量, Y 是随机变量, 且

$$Y = \alpha + \beta x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2), \tag{10.1.1}$$

其中 α, β, σ^2 是不依赖于 x 的未知参数. 称此模型为**一元线性回归模型**.

易见, 此时

$$Y \sim N(\alpha + \beta x, \sigma^2). \tag{10.1.2}$$

称 Y 的期望

$$\tilde{Y} = E(Y) = \alpha + \beta x \tag{10.1.3}$$

为 Y 关于 x 的**线性回归函数**. α, β 为**回归系数**, x 为**回归变量**.

在模型 (10.1.1) 下, x 与 Y 就存在线性相关关系, 也叫做**线性回归关系**.

取得样本观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 后, 若能得 α, β 的点估计 $\hat{\alpha}$ 和 $\hat{\beta}$, 称

$$\hat{Y} = \hat{\alpha} + \hat{\beta} x \tag{10.1.4}$$

为**线性回归方程**, 它所代表的直线叫做**线性回归直线**. 线性回归直线描述了 x 与 Y 线性相关关系中的线性趋势. 因此, 求线性回归方程是线性回归分析的最基本任务之一.

注意, (10.1.1) 式中, 若 $Y = f(x) + \varepsilon$, 则是一般的回归模型, 其中 $f(x) = E(Y)$ 是一般的回归函数.

§10.1.2 α, β 和 σ^2 的极大似然估计及性质

取得样本观测值 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 后, 由 (10.1.2), 得到似然函数

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{[y_i - (\alpha + \beta x_i)]^2}{2\sigma^2}} \\ &= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum\limits_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2}, \end{aligned}$$

从而有

$$\ln L(\alpha, \beta, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2.$$

故可得对数似然方程组

$$\begin{cases} \dfrac{\partial(\ln L(\alpha, \beta, \sigma^2))}{\partial \alpha} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)] = 0, \\ \dfrac{\partial(\ln L(\alpha, \beta, \sigma^2))}{\partial \beta} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} x_i [y_i - (\alpha + \beta x_i)] = 0, \\ \dfrac{\partial(\ln L(\alpha, \beta, \sigma^2))}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4} \displaystyle\sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 = 0. \end{cases}$$

整理为关于 α,β,σ^2 的方程组

$$\begin{cases} n\alpha+\beta\sum\limits_{i=1}^{n}x_i=\sum\limits_{i=1}^{n}y_i, \\ \alpha\sum\limits_{i=1}^{n}x_i+\beta\sum\limits_{i=1}^{n}x_i^2=\sum\limits_{i=1}^{n}x_iy_i, \\ \sigma^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}[y_i-(\alpha+\beta x_i)]^2. \end{cases}$$

解得 α,β,σ^2 的极大似然估计值为

$$\begin{cases} \hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}, \\ \hat{\beta}=\dfrac{s_{xy}}{s_{xx}}, \\ \hat{\sigma}^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}[y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2, \end{cases} \tag{10.1.5}$$

其中

$$\begin{aligned} s_{xy}&=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}(x_iy_i-\bar{x}y_i-\bar{y}x_i+\bar{x}\bar{y}) \\ &=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}, \end{aligned} \tag{10.1.6}$$

$$s_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2. \tag{10.1.7}$$

(10.1.5), (10.1.6) 和 (10.1.7) 中, 观测值 y_i 改为 Y_i, 则得到相应的极大似然估计量

$$\begin{cases} \hat{\alpha}=\bar{Y}-\hat{\beta}\bar{x}, \\ \hat{\beta}=\dfrac{s_{xY}}{s_{xx}}, \\ \hat{\sigma}^2=\dfrac{1}{n}\sum\limits_{i=1}^{n}[Y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2. \end{cases} \tag{10.1.8}$$

注意, α 和 β 的极大似然估计值是使似然函数中 e 的指数部分

$$Q(\alpha,\beta)=\sum_{i=1}^{n}[y_i-(\alpha+\beta x_i)]^2$$

达到最小的估计, 所以 α 和 β 的极大似然估计 $\hat{\alpha}$ 和 $\hat{\beta}$ 又叫做最小二乘估计.

事实上, $Q(\widehat{\alpha},\widehat{\beta})$ 是数据 (x_i,y_i) 对估计 $\widehat{y}_i=\widehat{\alpha}+\widehat{\beta}x_i$ 的偏差平方和 (又称残差平方和). 最小二乘估计使数据点到回归直线的垂直距离平方和达到最小 (图 10.1).

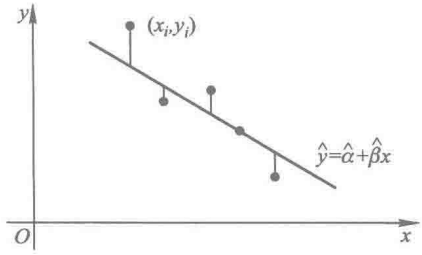

图 10.1

得到 $\hat{\alpha}$ 和 $\hat{\beta}$ 后, 可得回归方程

$$\hat{y}=\hat{\alpha}+\hat{\beta}x=\bar{y}+\hat{\beta}(x-\bar{x}), \tag{10.1.9}$$

可见回归直线过点 $(\bar{x},\bar{y})$.

例 10.1 一化学反应过程的中间步骤是在一个大气压下进行的. 反应温度为 $2\sim10$ °C, 相应的得率数据如下, 求得率 Y 关于温度 x 的线性回归方程.

温度 x/°C	2	4	6	8	10
得率 Y/%	5	10	14	17	21

解 由数据, 可得计算表 10.1.

表 10.1

x_i	2	4	6	8	10	$\sum x_i=30$
y_i	5	10	14	17	21	$\sum y_i=67$
x_i^2	4	16	36	64	100	$\sum x_i^2=220$
x_iy_i	10	40	84	136	210	$\sum x_iy_i=480$

由于 $n=5$, 故可得

$$\bar{x}=\frac{30}{5}=6,\qquad \bar{y}=\frac{67}{5}=13.4.$$

由 (10.1.5), (10.1.6) 和 (10.1.7), 得

$$\hat{\beta}=\frac{480-5\times6\times13.4}{220-5\times6^2}=1.95,$$
$$\hat{\alpha}=13.4-1.95\times6=1.7.$$

回归方程为

$$\widehat{Y} = 1.7 + 1.95x. \tag{10.1.10}$$

图 10.2 给出了回归直线与样本观测点 (x_i, y_i) $(i = 1, 2, \cdots, 5)$ 的关系, 其中回归直线过点 (6,13.4).

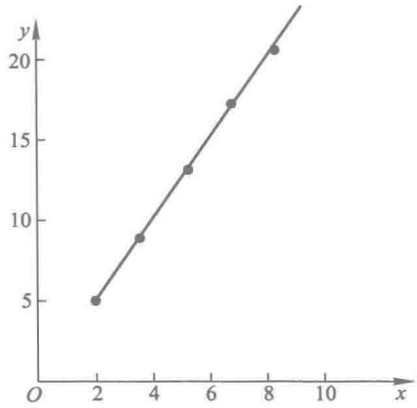

图 10.2

关于极大似然估计量 $\hat{\alpha}, \hat{\beta}$ 和 $\hat{\sigma}^2$, 我们有如下性质:

定理 10.1 (1) $\bar{Y}, \hat{\beta}$ 和 $\hat{\sigma}^2$ 相互独立;

$$(2)\ (\hat{\alpha}, \hat{\beta}) \sim N\left(\alpha, \beta; \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right), \frac{\sigma^2}{s_{xx}}, \frac{-\bar{x}}{\sqrt{s_{xx}}\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{s_{xx}}}}\right); \tag{10.1.11}$$

$$(3)\ \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2). \tag{10.1.12}$$

证　(1) 只证 $\bar{Y}$ 和 $\hat{\beta}$ 相互独立, 其余证明较难, 略去.

因为 $\bar{Y} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i, \hat{\beta} = \dfrac{s_{xY}}{s_{xx}} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) \Big/ s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})Y_i \Big/ s_{xx}$, 它们都是 $Y_1, Y_2, \cdots, Y_n$ 的线性组合. 对任意常数 a, b, 则 $z = a\bar{Y} + b\hat{\beta}$ 也是 $Y_1, Y_2, \cdots, Y_n$ 的线性组合, 由正态分布可加性, 知 z 服从正态分布. 再由第五章定理 5.8, 知 $(\bar{Y}, \hat{\beta})$ 服从二维正态分布. 因此要证 $\bar{Y}$ 与 $\hat{\beta}$ 相互独立, 只需证 $\mathrm{Cov}(\bar{Y}, \hat{\beta}) = 0$.

注意 $Y_1, Y_2, \cdots, Y_n$ 相互独立且具有相同方差 σ^2, 由协方差性质, 有

$$\begin{aligned}\mathrm{Cov}(\bar{Y}, \hat{\beta}) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{j=1}^{n}\frac{(x_j-\bar{x})}{s_{xx}}Y_j\right)\\ &= \frac{1}{ns_{xx}}\sum_{i=1}^{n}\mathrm{Cov}(Y_i, (x_i-\bar{x})Y_i) = \frac{1}{ns_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})\sigma^2 = 0,\end{aligned}$$

从而证得 $\bar{Y}$ 与 $\hat{\beta}$ 相互独立.

(2) 由于 $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$, 由 (1) 中所证 $(\bar{Y}, \hat{\beta})$ 服从二维正态分布. 故由定理 5.8, 知 $\hat{\alpha}$ 和 $\hat{\beta}$ 的任意线性组合 (因为也是 $\bar{Y}$ 和 $\hat{\beta}$ 的线性组合) 服从正态分布. 从而 $(\hat{\alpha}, \hat{\beta})$ 服从二维正态分布. 故以下只需求各数字特征.

由于 $E(Y_i) = \alpha + \beta x_i$, 从而

$$\begin{aligned}E(\hat{\beta}) &= E\left(\frac{s_{xY}}{s_{xx}}\right) = E\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{s_{xx}}\right)\\ &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(\alpha+\beta x_i)}{s_{xx}}\\ &= \frac{\alpha\sum_{i=1}^{n}(x_i-\bar{x}) + \beta\sum_{i=1}^{n}(x_i-\bar{x})x_i}{s_{xx}} = \beta,\\ E(\hat{\alpha}) &= E(\bar{Y} - \hat{\beta}\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(\alpha+\beta x_i) - \beta\bar{x}\\ &= \alpha + \beta\bar{x} - \beta\bar{x} = \alpha,\\ D(\hat{\beta}) &= D\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{s_{xx}}\right] = \frac{1}{s_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2}{s_{xx}}.\end{aligned}$$

注意 $\bar{Y}$ 与 $\hat{\beta}$ 独立, 有

$$\begin{aligned}D(\hat{\alpha}) &= D(\bar{Y} - \hat{\beta}\bar{x}) = D(\bar{Y}) + \bar{x}^2 D(\hat{\beta})\\ &= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{s_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right),\end{aligned}$$

又

$$\begin{aligned}\mathrm{Cov}(\hat{\alpha},\hat{\beta}) &= \mathrm{Cov}(\bar{Y}-\hat{\beta}\bar{x},\hat{\beta})\\ &= \mathrm{Cov}(\bar{Y},\hat{\beta})-\bar{x}\mathrm{Cov}(\hat{\beta},\hat{\beta}) = -\bar{x}\frac{\sigma^2}{s_{xx}},\end{aligned}$$

从而 $\hat{\alpha}$ 与 $\hat{\beta}$ 的相关系数为

$$r=\frac{-\bar{x}\dfrac{\sigma^2}{s_{xx}}}{\sqrt{\dfrac{\sigma^2}{s_{xx}}}\sqrt{\sigma^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{s_{xx}}\right)}}=\frac{-\bar{x}}{\sqrt{s_{xx}}\sqrt{\dfrac{1}{n}+\dfrac{\bar{x}^2}{s_{xx}}}}.$$

(3) 证明略去.

由定理 9.1, 我们知 $\hat{\alpha},\hat{\beta}$ 分别为 α 和 β 的无偏估计量, $\dfrac{n}{n-2}\hat{\sigma}^2=\dfrac{1}{n-2}\sum\limits_{i=1}^{n}[Y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2$ 是 σ^2 的无偏估计量, 且 $\hat{\alpha}$ 与 $\hat{\beta}$ 都服从正态分布.

§10.1.3　线性回归方程的显著性检验

当 x 与 Y 不存在线性相关关系时, 由式 (9.1.5) 也可计算出 $\hat{\alpha}$ 和 $\hat{\beta}$, 从而也可以形式上得到一条 "回归直线". 但这条直线并不代表 x 与 Y 的相关关系趋势. 因此在实际情况中, x 与 Y 是否存在线性相关 (线性回归) 关系, 即模型中 $E(Y)=\tilde{Y}=\alpha+\beta x$ 的假定是否符合实际情况, 是需要判断的.

这种判断首先可根据有关的专业知识和实践来判断, 也可以由统计理论和方法来判断. 直观和初步的统计判断方法可根据散点图来进行.

散点图是将各样本观测值 (x_i,Y_i) $(i=1,2,\cdots,n)$ 描绘在一个直角坐标平面上得到的图. 若这些点大致散布在某条直线周围, 则可初步认定 x 与 Y 之间存在线性相关关系或线性回归关系. 图 10.3 绘出了例 10.1 中数据的散点图. 这些点大致在一条直线周围, 且随着温度的增加, 得率有线性增加的趋势. 因此可初步认定该例中温度 x 与得率 Y 有线性相关关系.

对 x 与 Y 是否存在线性相关关系 (线性回归关系) 的细致分析应是对回归模型的线性性进行显著性检验. 检验假设为

$$H_0:\beta=0,\qquad H_1:\beta\neq 0. \tag{10.1.13}$$

当 $\beta=0$ 时, 相当于 x 与 Y 不存在线性回归关系.

当 H_0 成立时, 由定理 10.1, 有

$$\hat{\beta}\sim N\left(0,\frac{\sigma^2}{s_{xx}}\right),$$

$$\frac{n\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-2),$$

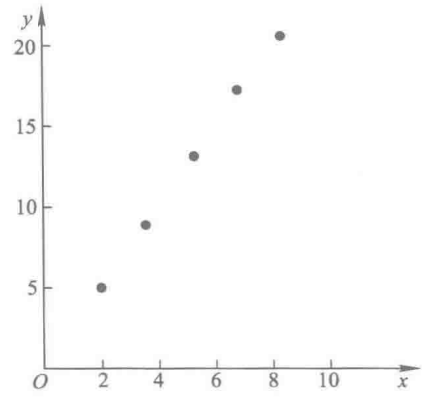

图 10.3

且 $\hat{\beta}$ 和 $\hat{\sigma}^2$ 独立. 于是

$$t=\frac{\hat{\beta}\Big/\dfrac{\sigma}{\sqrt{s_{xx}}}}{\sqrt{\dfrac{n\hat{\sigma}^2}{\sigma^2(n-2)}}}=\frac{\hat{\beta}\sqrt{s_{xx}}}{\sqrt{\dfrac{n}{n-2}\hat{\sigma}^2}}\sim t(n-2). \tag{10.1.14}$$

由于 H_1 成立时, 检验统计量 t 有 $|t|$ 取值偏大, 得到检验 (10.1.13) 的拒绝域为

$$W=\{|t|>t_{1-\alpha/2}(n-2)\}, \tag{10.1.15}$$

其中 α 为显著性水平.

关于检验统计量 t 中 $\hat{\sigma}^2$ 的计算, 我们有如下公式

$$\hat{\sigma}^2=\frac{1}{n}(s_{YY}-\hat{\beta}s_{xY}), \tag{10.1.16}$$

其中 $s_{YY}=\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2=\sum\limits_{i=1}^{n}Y_i^2-n\bar{Y}^2$.

(10.1.16) 成立是因为

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}[Y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2.$$

记 $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$, 则由 (10.1.9) 式, 有

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 &= \sum_{i=1}^{n}[Y_i - \bar{Y} - \hat{\beta}(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) + \hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= s_{YY} - 2\hat{\beta}s_{xY} + \hat{\beta}^2 s_{xx},\end{aligned}$$

但 $\hat{\beta} = \dfrac{s_{xY}}{s_{xx}}$, 故得

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = s_{YY} - \hat{\beta}s_{xY},$$

从而证得 (10.1.16) 式.

例 10.2　在例 10.1 中, 当显著性水平 $\alpha = 0.05$ 时, 检验温度 x 与得率 Y 之间是否存在线性回归关系.

解　检验假设为

$$H_0 : \beta = 0, \qquad H_1 : \beta \neq 0.$$

由例 10.1, 有 $n = 5, \widehat{\beta} = 1.95, s_{xx} = 220 - 5 \times 6^2 = 40, s_{xy} = 480 - 5 \times 6 \times 13.4 = 78$, 又可算得 $s_{yy} = \sum_{i=1}^{5} y_i^2 - 5\bar{y}^2 = 1\ 051 - 5 \times 13.4^2 = 153.2$, 从而有 $\hat{\sigma}^2 = \dfrac{1}{5} \times (153.2 - 1.95 \times 78) = 0.22$. 从而

$$t = \frac{1.95\sqrt{40}}{\sqrt{\dfrac{5}{3} \times 0.22}} = 20.367\ 1.$$

又查表, 知 $t_{1-\frac{\alpha}{2}}(n-2) = t_{0.975}(3) = 3.182\ 4 < |t| = 20.367\ 1$. 于是拒绝 H_0, 认为 $\beta \neq 0$, 即温度 x 与得率 Y 之间存在线性回归关系.

§10.1.4　预测

当经检验后认为 x 与 Y 存在线性回归关系后, 回归方程重要的用途是预测和控制. 即在 x 的指定值 x_0 处, 对于对应的 Y_0 的平均取值和 Y_0 的 $1-\alpha$ 取值区间进行预测, 就是 Y_0 的点预测和区间预测. 而控制问题则相反, 当 Y 的取值限定在区间 (Y_l, Y_u) 中时, 要求 x 应控制在什么范围内.

本书只讨论预测问题.

按模型 (10.1.1) 的假定, 当 $x = x_0$ 时, 有

$$Y_0 = \alpha + \beta x_0 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2). \tag{10.1.17}$$

这样, $E(Y_0)=\alpha+\beta x_0$. 其估计值 $\hat{Y}_0=\hat{\alpha}+\hat{\beta}x_0$ 就作为 Y_0 的点预测值. 下面求 Y_0 的预测区间.

因为 (x_0,Y_0) 是一个新的观测值, 故 Y_0 与 $Y_1,Y_2,\cdots,Y_n$ 相互独立. 由于 $\hat{Y}_0$ 是 $Y_1,Y_2,\cdots,Y_n$ 的线性组合, 故 $\hat{Y}_0$ 与 Y_0 相互独立, 且 $\hat{Y}_0-Y_0$ 是 $Y_0,Y_1,Y_2,\cdots,Y_n$ 的线性组合. 于是

$$\begin{aligned}
E(\hat{Y}_0-Y_0)&=E(\hat{\alpha}+\hat{\beta}x_0)-E(Y_0)\\
&=(\alpha+\beta x_0)-(\alpha+\beta x_0)=0,\\
D(\hat{Y}_0-Y_0)&=D(\hat{Y}_0)+D(Y_0)\\
&=D[\bar{Y}+\hat{\beta}(x_0-\bar{x})]+D(Y_0)\\
&=D(\bar{Y})+(x_0-\bar{x})^2D(\hat{\beta})+D(Y_0)\\
&=\frac{\sigma^2}{n}+(x_0-\bar{x})^2\frac{\sigma^2}{s_{xx}}+\sigma^2\\
&=\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right],
\end{aligned}$$

从而有

$$\hat{Y}_0-Y_0\sim N\left(0,\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right]\right). \tag{10.1.18}$$

又由定理 10.1 有

$$\frac{n\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-2),$$

及 $\hat{\sigma}^2$ 与 $\bar{Y},\hat{\beta}$ 相互独立, 可得 $\hat{\sigma}^2$ 与 $\hat{Y}_0-Y_0$ 相互独立. 于是

$$\begin{aligned}
t&=\frac{\hat{Y}_0-Y_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}}}\Bigg/\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2(n-2)}}\\
&=\frac{\hat{Y}_0-Y_0}{\sqrt{\frac{n\hat{\sigma}^2}{n-2}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right]}}\sim t(n-2).
\end{aligned} \tag{10.1.19}$$

对于置信度 $1-\alpha$, 可得 Y_0 的置信度为 $1-\alpha$ 的预测区间为

$$(\hat{Y}_0-\delta(x_0),\hat{Y}_0+\delta(x_0)), \tag{10.1.20}$$

其中

$$\delta(x_0)=t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{n}{n-2}\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right]}. \tag{10.1.21}$$

由 (10.1.20) 和 (10.1.21) 可知预测区间 (10.1.20) 的长度是 x_0 的函数, 随 $|x_0-\bar{x}|$ 的增加而增加, 当 $x_0=\bar{x}$ 时预测区间最短.

例 10.3　在例 10.1 中, 求温度为 7 °C 时, 得率 Y 的置信度为 95% 的预测区间.

解　由例 10.1, 有回归方程

$$\hat{Y}=1.7+1.95x.$$

当 $x_0=7$ 时, 得 $\hat{Y}_0=1.7+1.95\times 7=15.35$.

由例 10.1 还可知 $\bar{x}=6, s_{xx}=40$. 再由例 10.2, 知 $\hat{\sigma}^2=0.22$. 查表又可知 $t_{1-\frac{\alpha}{2}}(n-2)=t_{0.975}(3)=3.182\,4$, 由 (10.1.20) 和 (10.1.21), 得

$$\hat{Y}_0-\delta(x_0)=15.35-3.182\,4\sqrt{\frac{5}{3}\times 0.22\left[1+\frac{1}{5}+\frac{(7-6)^2}{40}\right]}=13.243\,7,$$

$$\hat{Y}_0+\delta(x_0)=15.35+3.182\,4\sqrt{\frac{5}{3}\times 0.22\left[1+\frac{1}{5}+\frac{(7-6)^2}{40}\right]}=17.456\,3,$$

即 Y_0 的置信度为 95% 的预测区间为

$$(13.243\,7, 17.456\,3).$$

在实际应用中, 若样本容量 n 较大, 且 x_0 在 $\bar{x}$ 邻近时, 注意此时 $t_{1-\frac{\alpha}{2}}(n-2)\approx u_{1-\frac{\alpha}{2}}$, 预测区间 (10.1.20) 化简为

$$(\hat{Y}_0-u_{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2}, \hat{Y}_0+u_{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2}). \tag{10.1.22}$$

§10.1.5　曲线回归的线性化

在许多情况下, x 与 Y 之间没有线性相关关系. 若由经验或专业知识或从散点图初步判断 x 与 Y 呈曲线相关关系时, 求曲线回归方程一般就十分复杂. 但也有一些曲线回归函数可通过变量替代化为线性回归函数, 例如生物学中细菌培养实验, 每一时刻繁殖的细菌总量 Y 与时间 t 之间呈现指数函数相关关系, 即 $Y=\alpha e^{\beta t}\varepsilon$, 则有 $\ln Y=\ln\alpha+\beta t+\ln\varepsilon$. 此时则可由前面介绍的直线回归分析方法来讨论. 以下通过例子来说明曲线回归的线性化.

例 10.4　表 10.2 是某年美国二手轿车价格的调查资料. 今以 x 表示轿车的使用年数, Y 表示相应的平均价格 (单位: 美元). 求 Y 关于 x 的回归方程.

表 10.2

使用年数 x	1	2	3	4	5	6	7	8	9	10
平均价格 Y	2 651	1 943	1 494	1 087	765	538	484	290	226	204

解 作散点图如图 10.4, 可见 x 与 Y 的相关关系呈指数下降趋势, 故令

$$Z = \ln Y.$$

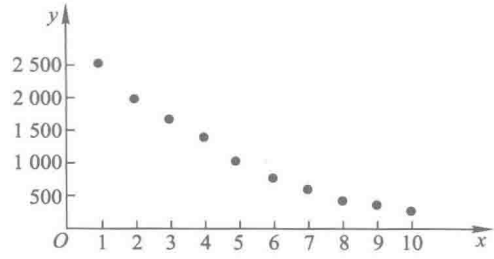

图 10.4

记 $z_i = \ln y_i$, 并作 (x_i, z_i) 的散点图如图 10.5, 可见 x 与 Z 有线性相关关系.

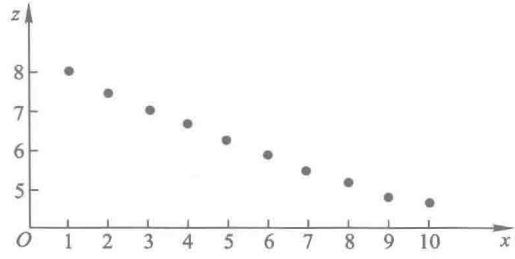

图 10.5

可设模型为

$$Z = \alpha + \beta x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).$$

做计算表 10.3.

表 10.3

x_i	1	2	3	4	5	6	7	8	9	10	$\Sigma x_i = 55$
z_i	7.88	7.57	7.31	6.99	6.64	6.29	6.18	5.67	5.42	5.32	$\Sigma z_i = 65.27$
x_i^2	1	4	9	16	25	36	49	64	81	100	$\Sigma x_i^2 = 385$
$x_i z_i$	7.88	15.14	21.93	27.96	33.2	37.74	43.26	45.36	48.78	53.2	$\Sigma x_i z_i = 334.45$

于是有 $\bar{x} = \dfrac{55}{10} = 5.5, \bar{z} = \dfrac{65.27}{10} = 6.527, s_{xz} = 334.45 - 10 \times 5.5 \times 6.527 =$

-24.535, $s_{xx}=385-10\times 5.5^2=82.5$, 从而由 (10.1.5) 有

$$\hat{\beta}=\frac{s_{xz}}{s_{xx}}=\frac{-24.535}{82.5}=-0.297\ 4,$$
$$\hat{\alpha}=\bar{z}-\hat{\beta}\bar{x}=6.527+0.297\ 4\times 5.5=8.162\ 7,$$

得 Z 关于 x 的线性回归方程为

$$\hat{Z}=8.162\ 7-0.297\ 4x.$$

再由 $\hat{\sigma}^2=\dfrac{1}{n}(s_{zz}-\hat{\beta}s_{xz})=\dfrac{1}{10}[7.352+0.297\ 4\times(-24.535)]=0.010\ 9$, 有

$$t=\frac{\hat{\beta}\sqrt{s_{xx}}}{\sqrt{\dfrac{n}{n-2}\hat{\sigma}^2}}=\frac{-0.297\ 4\sqrt{82.5}}{\sqrt{\dfrac{10}{8}\times 0.010\ 9}}=-23.141\ 9.$$

若取 $\alpha=0.05$, 查表有 $t_{1-\frac{\alpha}{2}}(n-2)=t_{0.975}(8)=2.306\ 0$. 因 $|t|=23.141\ 9>t_{0.975}(8)=2.306\ 0$, 则检验认为 x 与 Z 有显著的线性回归关系.

将所得的线性回归方程代回原变量 x 与 Y, 得到 Y 关于 x 的曲线回归方程

$$\hat{Y}=\mathrm{e}^{8.162\ 7-0.297\ 4x}.$$

一般说来, 曲线回归线性化中的变量替代是不唯一的. 通常我们要根据专业知识和散点图选择几种变量替换, 做出几个相应的线性回归方程. 再对它们作显著性检验, 选择 $|t|$ 最大的一个线性回归方程代回原变量得到曲线回归方程.

§10.2　单因素试验的方差分析

§10.2.1　单因素试验的方差分析模型

在科学实验和生产实践中, 影响一事物的因素常常有许多. 例如, 工业产品的质量受原材料、机器、工人的技术熟练水平等因素的影响;农作物的产品受种子、肥料、土壤等因素的影响. 英国的统计学家费希尔 (R. A. Fisher) 于 20 世纪 20 年代提出了方差分析方法, 可根据试验得到的数据, 分析各个因素对事物的影响是否显著.

影响一个随机变量 X 的因素可能有很多. 若试验中只让其中一个因素 A 变化, 而让其他因素保持不变, 这样的试验称为**单因素试验**. 若试验中变化的因素多于一个, 则称为**双因素**或**多因素试验**. 单因素试验中, 变化的因素 A 所处的状态叫做水平.

设因素 A 有 l 个水平 $A_1,A_2,\cdots,A_l$. 在水平 A_i 下的总体

$$X_i \sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \cdots, l. \tag{10.2.1}$$

并且总体 $X_1, X_2, \cdots, X_l$ 是相互独立的.

比如我们要分析水稻品种对水稻亩产品的影响, 试验田分别种了 5 种不同品种的水稻, 则变化的因素 A 为水稻的品种, A 共有 5 个水平, 即 5 个品种. 各品种水稻的亩产量 X_i 是相互独立的.

在总体 X_i 取得样本 $X_{i1}, X_{i2}, \cdots, X_{in_i}, i = 1, 2, \cdots, l$ 后, 单因素方差分析的模型为

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, 2, \cdots, l; j = 1, 2, \cdots, n_i. \tag{10.2.2}$$

且诸 X_{ij} 相互独立.

若因素 A 对 X_i 的影响不显著, 则应有各 μ_i 相等, 对应的检验假设为

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_l, \qquad H_1 : \mu_1, \mu_2, \cdots, \mu_l \text{ 不全相等}. \tag{10.2.3}$$

可见单因素试验的方差分析是检验方差相同的各正态总体的总体均值是否相等, 是第九章讨论的两个正态总体均值检验的推广.

§10.2.2 方差分析的原理和方法

两正态总体均值的检验是用样本均值差 $\bar{X} - \bar{Y}$ 导出的检验统计量进行检验.$\bar{X} - \bar{Y}$ 是 $\mu_1 - \mu_2$ 的无偏估计量, 反映了两总体均值差别的情况. 双侧检验结果是根据 $|\bar{X} - \bar{Y}|$ 的相对大小得出的. 而对于多个总体均值的检验, 我们可以考虑用 $\sum_{i=1}^{l} n_i(\bar{X}_i - \bar{X})^2$ 的相对大小得出检验 (10.2.3) 的结论, 其中

$$\begin{aligned} \bar{X}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \qquad i = 1, 2, \cdots, l, \\ \bar{X} &= \frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{n_i} X_{ij}, \qquad n = \sum_{i=1}^{l} n_i, \end{aligned} \tag{10.2.4}$$

即 $\bar{X}_i$ 是第 i 个样本 $X_{i1}, X_{i2}, \cdots, X_{in_i}$ 的样本均值, 习惯称为第 i 组样本的样本均值, $\bar{X}$ 为样本总均值.

下面推导检验 (10.2.3) 的方法. 记

$$S_T = \sum_{i=1}^{l}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2, \tag{10.2.5}$$

称为总离差平方和.

将 S_T 分解为

$$\begin{aligned} S_T &= \sum_{i=1}^{l} n_i(\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{l}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \\ &= S_A + S_E. \end{aligned} \tag{10.2.6}$$

(10.2.6) 成立是因为

$$\begin{aligned} S_T &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}[(X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})]^2 \\ &= \sum_{i=1}^{l}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{l}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 + \\ &\quad 2\sum_{i=1}^{l}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}), \end{aligned}$$

但

$$\sum_{i=1}^{l}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 = \sum_{i=1}^{l} n_i(\bar{X}_i - \bar{X})^2,$$

$$\begin{aligned} \sum_{i=1}^{l}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}) &= \sum_{i=1}^{l}(\bar{X}_i - \bar{X})\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i) \\ &= \sum_{i=1}^{l}(\bar{X}_i - \bar{X})(n_i\bar{X}_i - n_i\bar{X}_i) = 0. \end{aligned}$$

(10.2.5) 式中, S_A 表示各组样本均值 $\bar{X}_i$ 与样本总均值 $\bar{X}$ 的加权差异平方和, 故叫做组间平方和. 它反映的是因素 A 的不同水平下所取样本间的差异. S_E 表示各组样本中 X_{ij} 与该组样本均值 $\bar{X}_i$ 的差异平方和, 故叫做组内平方和. 它反映的是各观测随机变量 X_{ij} 由于随机因素而引起的随机误差.

定理 10.2　当 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_l$ 成立时, 在模型 (10.2.2) 下, 有

$$(1)\quad \frac{S_T}{\sigma^2} \sim \chi^2(n-1); \tag{10.2.7}$$

$$(2)\quad \frac{S_E}{\sigma^2} \sim \chi^2(n-l), \qquad \frac{S_A}{\sigma^2} \sim \chi^2(l-1). \tag{10.2.8}$$

且 S_A 与 S_E 独立.

证 (1) 当 H_0 成立时, 有各 X_{ij} 独立同分布于 $N(\mu,\sigma^2)$, 由第七章定理 7.4 后的说明, 有

$$\frac{S_T}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{l}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2\sim\chi^2(n-1).$$

(2) 同理, 有

$$\frac{\sum\limits_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2}{\sigma^2}\sim\chi^2(n_i-1).$$

再由各 X_{ij} 相互独立及 χ^2 分布可加性, 得

$$\frac{S_E}{\sigma^2}=\sum_{i=1}^{l}\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2\sim\chi^2\left(\sum_{i=1}^{l}(n_i-1)\right),$$

即 $\dfrac{S_E}{\sigma^2}\sim\chi^2(n-l)$.

由于 $\dfrac{S_T}{\sigma^2}\sim\chi^2(n-1)$, 由 χ^2 分布定义, 知道存在独立同服从于 $N(0,1)$ 分布的 $Z_1,Z_2,\cdots,Z_{n-1}$, 使得

$$\frac{S_T}{\sigma^2}=\sum_{i=1}^{n-1}Z_i^2.$$

同样由 χ^2 分布定义, 知 $\dfrac{S_E}{\sigma^2}$ 可表为 $n-l$ 个独立同服从于 $N(0,1)$ 分布的随机变量平方和, 且是 $\dfrac{S_T}{\sigma^2}$ 的一部分, 故而不妨令

$$\frac{S_E}{\sigma^2}=\sum_{i=1}^{n-l}Z_i^2.$$

因为 $S_T=S_A+S_E$, 则有

$$\frac{S_A}{\sigma^2}=\sum_{i=n-l+1}^{n-1}Z_i^2,$$

由 χ^2 分布定义, 故有

$$\frac{S_A}{\sigma^2}\sim\chi^2(l-1),$$

再由 Z_i 的相互独立性得 $\dfrac{S_A}{\sigma^2}$ 与 $\dfrac{S_E}{\sigma^2}$ 独立.

由定理 10.2 及 F 分布定义, 当 H_0 成立时, 有

$$F=\frac{S_A}{l-1}\Big/\frac{S_E}{n-l}\sim F(l-1,n-l),\tag{10.2.9}$$

而且当 H_0 不成立时, S_A 有偏大的倾向, 故得检验 (10.2.3) 的拒绝域为

$$W = \{F > F_{1-\alpha}(l-1, n-l)\}, \tag{10.2.10}$$

其中 α 为显著性水平.

习惯记 $M_A = \dfrac{S_A}{l-1}, M_E = \dfrac{S_E}{n-l}$, 检验中各种计算结果可列为分析表 (如表 10.4):

表 10.4　方差分析表

方差来源	S	自由度	MS	F
总	S_T	$n-1$		
组间	S_A	$l-1$	$\dfrac{S_A}{l-1}$	$\dfrac{M_A}{M_E}$
组内	S_E	$n-l$	$\dfrac{S_E}{n-l}$	

计算中, 由第七章 (7.3.6) 式, 可得

$$S_T = \sum_{i=1}^{l}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2, \tag{10.2.11}$$

$$S_E = \sum_{i=1}^{l}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{l} n_i \bar{X}_i^2. \tag{10.2.12}$$

例 10.5　某年度某湖不同季节湖水氯化物含量 (单位: mg/L) 测定结果如表 10.5, 试比较不同季节湖水氯化物含量的差别有无显著性 ($\alpha = 0.05$)?

表 10.5　某年度某湖不同季节湖水氯化物含量

春	夏	秋	冬
22.6	19.1	18.9	19.0
22.8	22.8	13.6	16.9
21.0	24.5	17.2	17.6
16.9	18.0	15.1	14.8
24.0	15.2	16.6	13.1
21.9	18.4	14.2	16.9
21.5	20.1	16.7	16.2
21.2	21.2	19.6	14.8

解　由表 10.5 中数据可得计算表 10.6:

表 10.6 某年度某湖不同季节湖水氯化物含量

	春	夏	秋	冬	合计	
	22.6	19.1	18.9	19.0		
	22.8	22.8	13.6	16.9		
	21.0	24.5	17.2	17.6		
	16.9	18.0	15.1	14.8		
	24.0	15.2	16.6	13.1		
	21.9	18.4	14.2	16.9		
	21.5	20.1	16.7	16.2		
	21.2	21.2	19.6	14.8		
$\sum_{j=1}^{8} x_{ij}$	171.9	159.3	131.9	129.3	592.4	$(\sum x)$
$\overline{x}_{i\cdot}$	21.49	19.91	16.49	16.16	18.51	$(\overline{x})$
$\sum_{j=1}^{8} x_{ij}^2$	3 724.51	3 231.95	2 206.27	2 114.11	11 276.84	$(\sum x^2)$

从而由 (10.2.11), (10.2.12) 有

$$
\begin{aligned}
S_T &= 11\ 276.84 - 32 \times 18.51^2 = 312.996\ 8, \\
S_E &= 11\ 276.84 - 8 \times 21.49^2 - 8 \times 19.91^2 - 8 \times 16.49 - 8 \times 16.16^2 \\
&= 146.488\ 8,
\end{aligned}
$$

故

$$
S_A = 312.996\ 8 - 146.488\ 8 = 166.508\ 0.
$$

于是

$$
\begin{aligned}
M_A &= \frac{166.508\ 0}{3} = 55.502\ 7, \\
M_E &= \frac{146.488\ 8}{28} = 5.231\ 7, \\
F &= \frac{55.502\ 7}{5.231\ 7} = 10.608\ 9.
\end{aligned}
$$

且查表, 有 $F_{1-\alpha}(l-1, n-l) = F_{0.95}(3, 28) = 2.95$, 得本例方差分析表 10.7:

表 10.7　方差分析表

方差来源	S	自由度	MS	F
总	312.996 8	31		
组间	166.508 0	3	55.502 7	10.608 9
组内	146.488 8	28	5.231 7	

因 $F = 10.608\ 9 > 2.95 = F_{0.95}(3, 28)$, 故拒绝 H_0, 认为不同季节的湖水氯化物含量有显著差别.

§10.3　双因素无重复试验的方差分析

若影响随机变量 X 的因素有两个, 要判断这两个因素对 X 的影响是否显著, 则需要进行双因素的方差分析. 根据每对因素水平的搭配是否进行了重复观察, 可以把双因素方差分析分为无重复试验的方差分析和有重复试验的方差分析. 我们首先讨论无重复试验的方差分析.

§10.3.1　双因素无重复试验的方差分析模型

设因素 A 有 l 个水平 $A_1, A_2, \cdots, A_l$, 因素 B 有 m 个水平 $B_1, B_2, \cdots, B_m$. 在因素 A 与 B 的各水平每种搭配 (A_i, B_j) 下只做一次试验, 试验的结果为随机变量 X_{ij}, 试验数据如表 10.8:

表 10.8　双因素无重复试验的数据

数据 / 因素	因素	因素 B 各水平			
因素		B_1	B_2	$\cdots$	B_m
因素 A 各水平	A_1	x_{11}	x_{12}	$\cdots$	x_{1m}
	A_2	x_{21}	x_{22}	$\cdots$	x_{2m}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_l	x_{l1}	x_{l2}	$\cdots$	x_{lm}

模型设置为

$$X_{ij} \sim N(\mu_{i\cdot} + \mu_{\cdot j}, \sigma^2), \quad i = 1, 2, \cdots, l;\ j = 1, 2, \cdots, m. \tag{10.3.1}$$

且诸 X_{ij} 相互独立. 这个模型叫做**双因素无重复试验的可加模型**.

如果因素 A 的影响不显著, 则各 $\mu_{i\cdot}$ 应相等. 同样, 因素 B 的影响不显著

时, 则各 $\mu_{\cdot j}$ 应相等. 因此显著性检验的假设为

$$\begin{aligned} &H_{0A}: \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{l\cdot}, \qquad && H_{1A}: \text{ 各 } \mu_{i\cdot} \text{ 不全相等;} \\ &H_{0B}: \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot m}, && H_{1B}: \text{ 各 } \mu_{\cdot j} \text{ 不全相等.} \end{aligned} \tag{10.3.2}$$

§10.3.2 方差分析方法

记

$$\begin{cases} \bar{X} = \dfrac{1}{lm}\displaystyle\sum_{i=1}^{l}\sum_{j=1}^{m} X_{ij}, \\ \bar{X}_{i\cdot} = \dfrac{1}{m}\displaystyle\sum_{j=1}^{m} X_{ij}, \\ \bar{X}_{\cdot j} = \dfrac{1}{l}\displaystyle\sum_{i=1}^{l} X_{ij}, \end{cases} \tag{10.3.3}$$

再记

$$\begin{cases} S_A = \displaystyle\sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{X}_{i\cdot} - \bar{X})^2 = m\sum_{i=1}^{l}(\bar{X}_{i\cdot} - \bar{X})^2, \\ S_B = \displaystyle\sum_{i=1}^{l}\sum_{j=1}^{m}(\bar{X}_{\cdot j} - \bar{X})^2 = l\sum_{j=1}^{m}(\bar{X}_{\cdot j} - \bar{X})^2, \\ S_E = \displaystyle\sum_{i=1}^{l}\sum_{j=1}^{m}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2. \end{cases} \tag{10.3.4}$$

由于 $\bar{X}$ 是样本总均值, $\bar{X}_{i\cdot}$ 是因素 A 在水平 A_i 下的样本均值, $\bar{X}_{\cdot j}$ 是因素 B 在水平 B_j 下的样本均值, 则 S_A 表示了因素 A 的不同水平下所取样本的差异, S_B 表示因素 B 的不同水平下所取样本的差异, 而 S_E 表示各 X_{ij} 的随机误差.

与单因素方差分析类似, 我们可得总离差平方和 S_T 方差分解, 即

$$S_T = \sum_{i=1}^{l}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2 = S_A + S_B + S_E. \tag{10.3.5}$$

定理 10.3 当 $H_{0A}: \mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{l\cdot}, H_{0B}: \mu_{\cdot 1} = \mu_{\cdot 2} = \cdots = \mu_{\cdot m}$ 成立时, 有

$$(1)\ \frac{S_T}{\sigma^2} \sim \chi^2(lm-1); \tag{10.3.6}$$

$$(2)\ \frac{S_A}{\sigma^2} \sim \chi^2(l-1), \frac{S_B}{\sigma^2} \sim \chi^2(m-1), \frac{S_E}{\sigma^2} \sim \chi^2((l-1)(m-1)). \tag{10.3.7}$$

且 S_A, S_B, S_E 相互独立.

定理 10.3 的证明略去.

由定理 10.3 及 F 分布定义, 当 H_{0A} 成立时, 有

$$F_A = \frac{S_A}{l-1} \Big/ \frac{S_E}{(l-1)(m-1)} \sim F(l-1,(l-1)(m-1)); \tag{10.3.8}$$

当 H_{0B} 成立时, 有

$$F_B = \frac{S_B}{l-1} \Big/ \frac{S_E}{(l-1)(m-1)} \sim F(m-1,(l-1)(m-1)). \tag{10.3.9}$$

而且当 H_{0A} 与 H_{0B} 不成立时, S_A 和 S_B 都有偏大的倾向, 故得检验 (10.3.2) 的拒绝域为

$$\begin{aligned} W_A &= \{F_A > F_{1-\alpha}(l-1,(l-1)(m-1))\}, \\ W_B &= \{F_B > F_{1-\alpha}(m-1,(l-1)(m-1))\}, \end{aligned} \tag{10.3.10}$$

其中 α 为显著性水平.

习惯记 $M_A = \dfrac{S_A}{l-1}, M_B = \dfrac{S_B}{m-1}, M_E = \dfrac{S_E}{(l-1)(m-1)}$, 检验中各种计算结果可列为方差分析表 10.9:

表 10.9　双因素无重复试验的方差分析表

方差来源	S	自由度	MS	F
总	S_T	$lm-1$		
因素 A	S_A	$l-1$	$\dfrac{S_A}{l-1}$	$F_A = \dfrac{M_A}{M_E}$
因素 B	S_B	$m-1$	$\dfrac{S_B}{m-1}$	$F_B = \dfrac{M_B}{M_E}$
误差	S_E	$(l-1)(m-1)$	$\dfrac{S_E}{(l-1)(m-1)}$	

计算中, 可用公式

$$\begin{cases} S_T = \displaystyle\sum_{i=1}^{l}\sum_{j=1}^{m} X_{ij}^2 - lm\bar{X}^2, \\ S_A = m\displaystyle\sum_{i=1}^{l} \bar{X}_{i\cdot}^2 - lm\bar{X}^2, \\ S_B = l\displaystyle\sum_{j=1}^{m} \bar{X}_{\cdot j}^2 - lm\bar{X}^2, \\ S_E = S_T - S_A - S_B. \end{cases} \tag{10.3.11}$$

例 10.6 一火箭使用 4 种燃料、3 种推进器做射程试验, 每种燃料与每种推进器的配伍做一次试验, 得火箭射程数据 (单位: km) 如表 10.10 的左上部分. 试问推进器之间、燃料之间是否存在显著差异 ($\alpha = 0.05$)?

表 10.10 火 箭 射 程

燃料 A \ 推进器 B	B_1	B_2	B_3	$\sum_{j=1}^{3} x_{ij}$	$\bar{x}_{i\cdot}$
A_1	58.2	56.2	65.3	179.7	59.9
A_2	49.1	54.1	51.6	154.8	51.6
A_3	60.1	70.9	39.2	170.2	56.7
A_4	75.8	58.2	48.7	182.7	60.9
$\sum_{i=1}^{4} x_{ij}$	243.2	239.4	204.8	687.4	$(\sum x)$
$\bar{x}_{\cdot j}$	60.8	59.85	51.2	57.27	$(\bar{x})$
$\sum_{i=1}^{4} x_{ij}^2$	15 155.7	14 499.3	10 834.98	40 489.98	$(\sum x^2)$

解 表 10.10 的下部和右部给出基本计算表, 从而由 (10.3.11) 可得

$$S_T = 40\,489.98 - 12 \times 57.28^2 = 1\,118.00,$$
$$S_A = 3 \times (59.9^2 + 51.6^2 + 56.7^2 + 60.9^2) - 12 \times 57.28^2 = 150.83,$$
$$S_B = 4 \times (60.8^2 + 59.85^2 + 51.2^2) - 12 \times 57.28^2 = 228.43,$$
$$S_E = 1\,118.00 - 150.83 - 228.43 = 738.74.$$

从而

$$M_A = \frac{150.83}{3} = 50.28, \quad M_B = \frac{228.43}{2} = 114.22,$$
$$M_E = \frac{738.74}{6} = 123.12,$$
$$F_A = \frac{50.28}{123.12} = 0.408\,4, \qquad F_B = \frac{114.22}{123.12} = 0.927\,7.$$

本例的计算结果列为方差分析表 10.11:

表 10.11 方差分析表

方差来源	S	自由度	MS	F
总	1 118.00	11		
因素 A	150.83	3	50.28	0.408 4
因素 B	228.43	2	114.22	0.927 7
误差	738.74	6	123.12	

查表, 得 $F_{1-\alpha}(l-1,(l-1)(m-1))=F_{0.95}(3,6)=4.76>F_A=0.408\,4$, $F_{1-\alpha}(m-1,(l-1)(m-1))=F_{0.95}(2,6)=5.14>F_B=0.927\,7$. 故都不拒绝 H_0, 从而认为燃料之间、推进器之间都不存在显著差异.

§10.4　双因素有重复试验的方差分析

§10.4.1　双因素有重复试验的方差分析模型

若在双因素 A 与 B 各水平的每种搭配 (A_i,B_j) 下分别做 n 次试验 $(n>1)$, 则称为双因素有重复试验的方差分析. 试验的结果记为随机变量 X_{ijk}, 试验的数据如表 10.12:

表 10.12　双因素有重复试验的数据

数据 / 因素	因素	因素 B 各水平			
因素		B_1	B_2	$\cdots$	B_m
因素 A 各水平	A_1	$x_{111},x_{112},\cdots,x_{11n}$	$x_{121},x_{122},\cdots,x_{12n}$	$\cdots$	$x_{1m1},x_{1m2},\cdots,x_{1mn}$
	A_2	$x_{211},x_{212},\cdots,x_{21n}$	$x_{221},x_{222},\cdots,x_{22n}$	$\cdots$	$x_{2m1},x_{2m2},\cdots,x_{2mn}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_l	$x_{l11},x_{l12},\cdots,x_{l1n}$	$x_{l21},x_{l22},\cdots,x_{l2n}$	$\cdots$	$x_{lm1},x_{lm2},\cdots,x_{lmn}$

模型设置为

$$X_{ijk}\sim N(\mu_{i\cdot\cdot}+\mu_{\cdot j\cdot}+v_{ij},\sigma^2), \tag{10.4.1}$$

其中 $i=1,2,\cdots,l, j=1,2,\cdots,m, k=1,2,\cdots,n$.

如果因素 A 的影响不显著, 则各 $\mu_{i\cdot\cdot}$ 应相等. 同样, 因素 B 的影响不显著时, 各 $\mu_{\cdot j\cdot}$ 应相等. 与双因素无重复试验不同的是, 此时我们还需检验因素 A 与因素 B 的组合是否产生交互效应, 即 A 与 B 的配伍试验结果是否有显著影响. 当交互作用没有显著影响时, 有 $v_{ij}=0$. 因此显著性检验的假设为:

$$\begin{aligned}
&H_{0A}:\mu_{1\cdot\cdot}=\mu_{2\cdot\cdot}=\cdots=\mu_{l\cdot\cdot}, \quad H_{1A}:\text{各 }\mu_{i\cdot\cdot}\text{ 不全相等};\\
&H_{0B}:\mu_{\cdot1\cdot}=\mu_{\cdot2\cdot}=\cdots=\mu_{\cdot m\cdot}, \quad H_{1B}:\text{各 }\mu_{\cdot j\cdot}\text{ 不全相等};\\
&H_{0AB}:v_{ij}=0, i=1,2,\cdots,l; j=1,2,\cdots,m, \quad H_{1AB}:\text{某个 } v_{ij}\neq 0.
\end{aligned} \tag{10.4.2}$$

§10.4.2 方差分析方法

类似地, 记

$$\begin{cases} \overline{X} = \dfrac{1}{lmn}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{m}\sum\limits_{k=1}^{n} X_{ijk}, \\ \overline{X}_{i\cdot\cdot} = \dfrac{1}{mn}\sum\limits_{j=1}^{m}\sum\limits_{k=1}^{n} X_{ijk}, \\ \overline{X}_{\cdot j\cdot} = \dfrac{1}{ln}\sum\limits_{i=1}^{l}\sum\limits_{k=1}^{n} X_{ijk}, \\ \overline{X}_{ij\cdot} = \dfrac{1}{n}\sum\limits_{k=1}^{n} X_{ijk}. \end{cases} \tag{10.4.3}$$

再记

$$\begin{cases} S_A = mn\sum\limits_{i=1}^{l}(\overline{X}_{i\cdot\cdot} - \overline{X})^2, \\ S_B = ln\sum\limits_{j=1}^{m}(\overline{X}_{\cdot j\cdot} - \overline{X})^2, \\ S_{AB} = n\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{m}(\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X})^2, \\ S_E = \sum\limits_{i=1}^{l}\sum\limits_{j=1}^{m}\sum\limits_{k=1}^{n}(\overline{X}_{ijk} - \overline{X}_{ij\cdot})^2. \end{cases} \tag{10.4.4}$$

由于 $\overline{X}$ 是样本总均值, $\overline{X}_{i\cdot\cdot}$ 是因素 A 在水平 A_i 下的样本均值, $\overline{X}_{\cdot j\cdot}$ 是因素 B 在水平 B_j 下的样本均值, $\overline{X}_{ij\cdot}$ 是因素 A 和因素 B 在配伍 (A_i, B_j) 下的样本均值, 则有 S_A 表示因素 A 在不同水平下所取样本的差异, S_B 表示在因素 B 在不同水平下所取样本的差异, 而 S_{AB} 表示因素 A 与因素 B 在不同配伍 (A_i, B_j) 下的差异, S_E 表示随机误差.

与双因素无重复试验的方差分解类似, 我们可得离差总平方和的分解式:

$$\begin{aligned} S_T &= \sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n}(X_{ijk} - \overline{X})^2 \\ &= S_A + S_B + S_{AB} + S_E. \end{aligned} \tag{10.4.5}$$

定理 10.4 当 H_{0A}, H_{0B}, H_{0AB} 成立时, 有

(1) $\dfrac{S_T}{\sigma^2} \sim \chi^2(lmn-1)$; (10.4.6)

(2) $\dfrac{S_A}{\sigma^2}\sim\chi^2(l-1),\dfrac{S_B}{\sigma^2}\sim\chi^2(m-1),$

$$\frac{S_{AB}}{\sigma^2}\sim\chi^2((l-1)(m-1)),\frac{S_E}{\sigma^2}\sim\chi^2(lm(n-1)). \tag{10.4.7}$$

且 S_A,S_B,S_{AB},S_E 相互独立,

定理 10.4 的证明略去.

由定理 10.4 及 F 分布的定义, 当 H_{0A} 成立时, 有

$$F_A=\frac{S_A}{l-1}\Big/\frac{S_E}{lm(n-1)}\sim F(l-1,lm(n-1)), \tag{10.4.8}$$

当 H_{0B} 成立时, 有

$$F_B=\frac{S_B}{m-1}\Big/\frac{S_E}{lm(n-1)}\sim F(m-1,lm(n-1)), \tag{10.4.9}$$

当 H_{0AB} 成立时, 有

$$F_{AB}=\frac{S_{AB}}{(l-1)(m-1)}\Big/\frac{S_E}{lm(n-1)}\sim F((l-1)(m-1),lm(n-1)), \tag{10.4.10}$$

而且当 H_{0A},H_{0B},H_{0AB} 成立时, 分别有 S_A,S_B,S_{AB} 取偏大的值的倾向, 故检验 (10.4.2) 的拒绝域为

$$\begin{aligned}W_A&=\{F_A>F_{1-\alpha}(l-1,lm(n-1))\}\\W_B&=\{F_B>F_{1-\alpha}(m-1,lm(n-1))\}\\W_{AB}&=\{F_{AB}>F_{1-\alpha}((l-1)(m-1),lm(n-1))\}\end{aligned} \tag{10.4.11}$$

其中 α 为显著性水平.

习惯记 $M_A=\dfrac{S_A}{l-1},M_B=\dfrac{S_B}{m-1},M_{AB}=\dfrac{S_{AB}}{(l-1)(m-1)},M_E=\dfrac{S_E}{lm(n-1)}$, 检验中的各种计算结果可列为方差分析表 10.13:

表 10.13　双因素有重复试验的方差分析表

方差来源	S	自由度	MS	F
总	S_T	$lm(n-1)$		
因素 A	S_A	$l-1$	$\dfrac{S_A}{l-1}$	$F_A=\dfrac{M_A}{M_E}$
因素 B	S_B	$m-1$	$\dfrac{S_B}{m-1}$	$F_B=\dfrac{M_B}{M_E}$
A 与 B 交互效应	S_{AB}	$(l-1)(m-1)$	$\dfrac{S_{AB}}{(l-1)(m-1)}$	$F_{AB}=\dfrac{M_{AB}}{M_E}$
误差	S_E	$lm(n-1)$	$\dfrac{S_E}{lm(n-1)}$	

例 10.7 某超市将同一种商品做了 3 种不同包装, 并摆放在 3 个不同的货架区进行销售试验. 随机抽取了 3 天的销售量作为样本, 具体数据见表 10.14. 要求检验商品包装、摆放位置及其搭配对销售量是否有显著影响 ($\alpha = 0.05$).

表 10.14 观 察 资 料

观察值		包装 B		
		B_1	B_2	B_3
摆放位置 A	A_1	5, 6, 4	6, 8, 7	4, 3, 5
	A_2	7, 8, 8	5, 5, 6	3, 6, 4
	A_3	3, 2, 4	6, 6, 5	8, 9, 6

解 这是一个两因素三水平的试验问题, 并且每个因素的搭配各作了 3 次观察, 所以属于有重复试验的方差分析.

提出假设

H_{0A} : 摆放位置对商品销售影响不显著

H_{1A} : 摆放位置对商品销售影响显著

H_{0B} : 包装对商品销售影响不显著

H_{1B} : 包装对商品销售影响显著

H_{0AB} : 摆放位置与包装对商品销售影响不显著

H_{1AB} : 摆放位置与包装对商品销售影响显著

计算离差平方和

$$
\begin{aligned}
S_T &= \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}(X_{ijk}-\overline{X})^2 \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}X_{ijk}^2 - \frac{1}{3\times 3\times 3}\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}X_{ijk}\right)^2 \\
&= 5^2+6^2+4^2+\cdots+8^2+9^2+6^2 - \frac{1}{3\times 3\times 3}\times \\
&\quad (5+6+4+\cdots+8+9+6)^2 \\
&= 84.74, \\
S_A &= 3\times 3\sum_{i=1}^{3}(\overline{X}_{i\cdot\cdot}-\overline{X})^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{3\times 3}\sum_{i=1}^{3}\left(\sum_{j=1}^{3}\sum_{k=1}^{3}X_{ijk}\right)^2 - \frac{1}{3\times 3\times 3}\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}X_{ijk}\right)^2\\
&= \frac{1}{3\times 3}\times[(5+6+4+6+8+7+4+3+5)^2+\\
&\quad (7+8+8+5+5+6+3+6+4)^2\\
&\quad +(3+2+4+6+6+5+8+9+6)^2]\\
&\quad -\frac{1}{3\times 3\times 3}(5+6+4+\cdots+8+9+6)^2\\
&= 0.96,\\
S_B &= 3\times 3\sum_{j=1}^{3}(\overline{X}_{\cdot j\cdot}-\overline{X})^2\\
&= \frac{1}{3\times 3}\sum_{j=1}^{3}\left(\sum_{i=1}^{3}\sum_{k=1}^{3}X_{ijk}\right)^2 - \frac{1}{3\times 3\times 3}\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}X_{ijk}\right)^2\\
&= \frac{1}{3\times 3}\times[(5+6+7+7+8+8+3+2+4)^2\\
&\quad +(6+8+7+5+5+6+6+6+5)^2\\
&\quad +(4+3+5+3+6+4+8+9+6)^2]\\
&\quad -\frac{1}{3\times 3\times 3}\times(5+6+4+\cdots+8+9+6)^2\\
&= 3.18,\\
S_E &= \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}(X_{ijk}-\overline{X}_{ij\cdot})^2\\
&= \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}X_{ijk}^2 - \frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\sum_{k=1}^{3}X_{ijk}\right)^2\\
&= 5^2+6^2+4^2+\cdots+8^2+9^2+6^2-\frac{1}{3}\times[(5+6+4)^2\\
&\quad +(6+8+7)^2+(4+3+5)^2+\cdots+(8+9+6)^2]\\
&= 19.33,\\
S_{AB} &= S_T - S_A - S_B - S_E\\
&= 84.74-0.96-3.18-19.33\\
&= 61.27.
\end{aligned}
$$

由 $\alpha=0.05$, 查表得 $F_{0.95}(2,18)=3.55, F_{0.95}(4,18)=2.93$. 分析结果列在表 10.15 的最后一列中.

表 10.15 方差分析资料计算表

方差来源	平方和	自由度	均方	均方比	显著性
摆放位置影响	0.96	2	0.48	0.45	不显著
包装影响	3.18	2	1.59	1.49	不显著
摆放位置与包装影响	61.27	4	15.32	14.32	显著
误差影响	19.33	18	1.07		
总离差	84.74	26			

§10.5 复习分析题

例 10.8 在线性回归分析中, 记 $S_T=\sum_{i=1}^{n}(Y_i-\bar{Y})^2, S_R=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2, S_E=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$, 分别叫做总平方和、回归平方和与残差平方和. 则关于 $H_0:\beta=0; H_1:\beta\neq 0$ 的 t 检验有等价的 F 检验, 即检验统计量为

$$F=\frac{S_R}{S_E/n-2}, \tag{10.5.1}$$

对显著性水平 α, 拒绝域为 $W=\{F>F_{1-\alpha}(1,n-2)\}$.

分析 要证明当 H_0 成立时, 有

$$F=\frac{S_R}{S_E/n-2}\sim F(1,n-2),$$

且说明当 H_1 成立时, F 有取值偏大的趋势.

证 由 (10.1.9) 式, 知

$$\begin{aligned}S_R&=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2=\sum_{i=1}^{n}\left[\bar{Y}+\hat{\beta}(x_i-\bar{x})-\bar{Y}\right]^2\\&=\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2=\hat{\beta}^2 s_{xx}.\end{aligned}$$

再由 (10.1.8) 式, 知

$$S_E=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)=\sum_{i=1}^{n}[Y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2=n\hat{\sigma}^2.$$

故

$$F=\frac{S_R}{S_E/n-2}=\frac{\hat{\beta}^2 s_{xx}}{n\hat{\sigma}^2/n-2}=t^2,$$

其中 t^2 由 (10.1.14) 确定. 且在 H_0 成立时, 有 $t \sim t(n-2)$.

由 t 分布及 F 分布定义, 可知 H_0 成立时, 有 $F = t^2 \sim F(1, n-2)$.

而当 $H_1 : \beta \neq 0$ 成立时, 由 (10.1.14) 后的解释知 $|t|$ 取值偏大, 从而 $F = t^2$ 取值偏大. 故检验拒绝域为

$$W = \{F > F_{1-\alpha}(1, n-2)\}.$$

例 10.9　对例 10.8 定义的 S_T, S_R 和 S_E, 有

$$S_T = S_R + S_E. \tag{10.5.2}$$

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

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

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

故只需证交叉项和为零.

证　由分析, 要证

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

但由 (10.1.9) 式, 有

$$\hat{Y}_i = \bar{Y} + \hat{\beta}(x_i - \bar{x}),$$

从而

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

$$= \sum_{i=1}^{n}[Y_i - \bar{Y} - \hat{\beta}(x_i - \bar{x})][\bar{Y} + \hat{\beta}(x_i - \bar{x}) - \bar{Y}]$$

$$= \sum_{i=1}^{n}[(Y_i - \bar{Y}) - \hat{\beta}(x_i - \bar{x})]\hat{\beta}(x_i - \bar{x})$$

$$= \hat{\beta}\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}s_{xY} - \hat{\beta}^2 s_{xx}.$$

但由 (10.1.8) 式, 有 $\hat{\beta}=\dfrac{s_{xY}}{s_{xx}}$, 从而

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})&=\hat{\beta}s_{xY}-\hat{\beta}^2s_{xx}\\&=\frac{s_{xY}^2}{s_{xx}}-\frac{s_{xY}^2}{s_{xx}}=0.\end{aligned}$$

例 10.10 在线性回归模型中, 记样本相关系数为

$$R=\frac{\sum\limits_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{s_{xY}}{\sqrt{s_{xx}\cdot s_{YY}}},$$

则样本相关系数 R 与例 10.8 中 F 统计量有关系

$$|R|=\sqrt{\frac{F}{F+n-2}}. \tag{10.5.3}$$

分析 将 R 与 F 都表成 s_{xY},s_{xx},s_{YY} 决定的量则可得其关系.

证 由例 10.9 及 (10.1.8) 式知

$$\begin{aligned}S_R&=\hat{\beta}^2s_{xx}=\frac{s_{xY}^2}{s_{xx}},\\S_E&=S_T-S_R=S_{YY}-\frac{S_{xY}^2}{S_{xx}},\end{aligned}$$

从而

$$\begin{aligned}F&=\frac{(n-2)S_R}{S_E}\\&=\frac{(n-2)\dfrac{S_{xY}^2}{S_{xx}}}{S_{YY}-\dfrac{S_{xY}^2}{S_{xx}}}=\frac{(n-2)\dfrac{S_{xY}^2}{S_{xx}S_{YY}}}{1-\dfrac{S_{xY}^2}{S_{xx}S_{YY}}}\\&=\frac{(n-2)R^2}{1-R^2},\end{aligned}$$

从而有解

$$|R|=\sqrt{\frac{F}{F+n-2}}.$$

注 由于 Y 关于 x 的直线回归关系又称为 Y 关于 x 的线性相关关系. 故样本相关系数描述了样本 $(x_1,Y_1),\cdots,(x_n,Y_n)$ 的直线相关关系. 所以样本相关系

数的绝对值 $|R|$ 越接近于 1, 就越有理由相信 Y 关于 x 的线性相关关系存在. 所以由 $|R|$ 的大小, 可得线性相关关系的检验方法, 其临界值 C 可由 $F_{1-\alpha}(1,n-2)$ 代入 (10.5.3) 得到. 拒绝域为 $W=\{|R|>C\}$.

例 10.11　一只红铃虫的产卵数与温度有关, 观测数据如下表:

温度 x/°C	21	23	25	27	29	32	35
产卵数 Y	7	11	24	26	66	115	325

假定 x 与 Y 有回归模型

$$Y=\beta_0\exp\{\beta_1 x+\varepsilon\},\qquad \varepsilon\sim N(0,\sigma^2).$$

(1) 由此模型及数据求曲线回归方程;

(2) 对此曲线回归关系做显著性检验 ($\alpha=0.05$);

(3) 求温度为 30°C 时, 产卵数的预测区间 ($\alpha=0.05$).

分析　对模型两端取对数, 有

$$\ln Y=\ln\beta_0+\beta_1 x+\varepsilon,\qquad \varepsilon\sim N(0,\sigma^2).$$

令 $Z=\ln Y, a=\ln\beta_0, b=\beta_1$, 则模型换为等价的线性回归模型

$$Z=a+bx+\varepsilon,\qquad \varepsilon\sim N(0,\sigma^2),$$

从而可对此线性回归模型做相应回归分析.

解　(1) 由分析, 得等价线性回归模型的基本数据计算表:

x_i	21	23	25	27	29	32	35	$\Sigma x_i=192$
Y_i	7	11	24	26	66	115	325	
z_i	1.945 9	2.397 9	3.044 5	3.178 1	4.189 7	4.744 9	5.783 8	$\Sigma z_i=25.284\ 8$
x_i^2	441	529	625	729	841	1 024	1 225	$\Sigma x_i^2=5\ 414$
x_iz_i	40.863 9	55.151 7	76.112 5	85.808 7	121.501 3	151.836 8	202.433 0	$\Sigma x_iz_i=733.707\ 9$

从而有

$$\bar{x}=\frac{192}{7}=27.4,\quad \bar{z}=\frac{25.284\ 8}{7}=3.612\ 1.$$

由 (10.1.5), (10.1.6) 和 (9.1.7), 有

$$\hat{b}=\frac{\sum_{i=1}^{7}x_iz_i-7\bar{x}\bar{y}}{\sum_{i=1}^{7}x_i^2-7\bar{x}^2}=\frac{733.707\ 9-7\times 27.4\times 3.612\ 1}{5\ 414-7\times 27.4^2}=0.272\ 1,$$

$$\hat{a}=\bar{z}-\hat{b}\bar{x}=3.612\ 1-0.272\ 1\times 27.4=-3.843\ 4,$$

从而有 $Z=\ln Y$ 关于 x 的线性回归方程

$$\hat{Z}=-3.843\ 4+0.272\ 1x,$$

故有 Y 关于 x 的曲线回归方程

$$\begin{aligned}\hat{Y}&=\mathrm{e}^{-3.843\ 4}\mathrm{e}^{0.272\ 1x}\\&=0.021\ 4\mathrm{e}^{0.272\ 1x}.\end{aligned}$$

(2) 对曲线回归方程做显著性检验相当于对等价的线性回归方程做检验 H_0 : $b=0$.

由于可算得

$$\begin{aligned}s_{zz}&=\sum_{i=1}^{7}z_i^2-7\bar{z}^2=102.425\ 7-7\times 3.612\ 1^2\\&=11.094\ 8,\\s_{xx}&=\sum_{i=1}^{7}x_i^2-7\bar{x}^2=5\ 414-7\times 27.4^2\\&=147.7,\\s_{xx}&=\sum_{i=1}^{7}x_iz_i-7\bar{x}\bar{z}=733.707\ 9-7\times 27.4\times 3.612\ 1\\&=40.182\ 0.\end{aligned}$$

故由 (10.1.16) 式, 有

$$\begin{aligned}\hat{\sigma}^2&=\frac{1}{n}(s_{zz}-\hat{b}s_{xz})=\frac{1}{7}(11.094\ 8-0.272\ 1\times 40.182\ 0)\\&=0.022\ 9.\end{aligned}$$

再由 (10.1.14) 式, 有

$$\begin{aligned}t&=\frac{\hat{b}\sqrt{s_{xx}}}{\sqrt{\frac{n}{n-2}\hat{\sigma}^2}}=\frac{0.272\ 1\sqrt{147.7}}{\sqrt{\frac{7}{5}\times 0.022\ 9}}\\&=18.468\ 7.\end{aligned}$$

因为 $\alpha=0.05, t_{1-\frac{\alpha}{2}}(n-2)=t_{0.975}(5)=2.570\,6$, 故 $t=18.468\,7>2.570\,6=t_{0.975}(5)$, 从而认为 Z 关于 x 存在线性回归关系, 即 Y 关于 x 存在曲线回归关系.

(3) 当 $x=30$ 时, 对 Z 做线性回归的预测区间, 有

$$\hat{Z}_0=\hat{a}+\hat{b}x_0=-3.843\,4+0.272\,1\times 30=4.319\,6.$$

由 (10.1.21) 式, 有

$$\begin{aligned}\delta(x_0)&=t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{n}{n-2}\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right]}\\&=2.570\,6\sqrt{\frac{7}{5}\times 0.022\,9\left[1+\frac{1}{7}+\frac{(30-27.4)^2}{147.7}\right]}\\&=0.501\,8,\end{aligned}$$

故由 (10.1.20) 式, Z 关于 $x=30$ 的预测区间为

$$(4.319\,6-0.501\,8, 4.319\,6+0.501\,8),$$

从而 Y 关于 $x=30$ 的预测区间为

$$(\mathrm{e}^{4.319\,6-0.501\,8}, \mathrm{e}^{4.319\,6+0.501\,8})=(45.504\,0, 124.138\,8).$$

习　题　十

1. 某工厂采用一种选矿工艺, 对原矿铅进行选矿, 测得原矿铅氧化率 x(单位:%) 与选矿铅的回收率 y(单位:%) 列于下表:

x	15.55	5.79	7.07	6.42	8.71	10.97	13.08	18.50	22.50	23.15	18.24	17.53
y	82.40	87.22	86.24	88.20	80.17	83.50	80.18	74.87	77.11	74.25	78.24	78.53

求回收率 Y 与氧化率 x 的线性回归方程.

2. 为了研究钢线含碳量 (单位: %) 对于电阻 (单位: Ω) 在 20℃ 下的影响, 做了 7 次试验, 得到数据如下:

含碳量 x_i	0.10	0.30	0.40	0.55	0.70	0.80	0.95
电阻 y_i	15	18	19	21	22.6	23.8	26

(1) 检验电阻 Y 与含碳量 x 之间线性相关关系是否显著 ($\alpha=0.05$); 如果显著, 求 Y 关于 x 的线性回归方程.

(2) 当含碳量为 0.50 时, 求电阻的置信水平为 95% 的预测区间.

3. 某地区近六年来轻工业产品利润总额 Y (单位: 亿元) 与年次 x 有如下数据:

年次 x	1	2	3	4	5	6
利润总额 y	11.35	11.85	12.44	13.07	13.59	14.41

已知年次与利润之间呈指数关系

$$Y = ab^x\varepsilon, \quad \text{且} \quad \ln \varepsilon \sim N(0, \sigma^2),$$

求 Y 关于 x 的曲线回归方程.

4. 观察五台机器的粘接强度, 得数据 (单位: $\mathrm{kg/cm^2}$) 如下表:

机器	强度测定值									
1	40	47	48	46	45	43				
2	43	45	40	40	43	45	47	44		
3	42	39	45	31	38	40	38	33	36	35
4	42	32	38	35	35	35	36	36		
5	31	35	34	31	37	34				

问在显著性水平 $\alpha = 0.01$ 下, 这五台机器的粘接强度有无显著差异?

5. 某厂试制了三台同样的专用机床, 为鉴别这三台样机的生产效率有无显著差异, 由四名技术工人分别在这三台样机上各操作一天, 记录其产量如下表所示:

工人 \ 机床	1	2	3
甲	80	106	84
乙	74	88	64
丙	74	94	62
丁	86	96	76

试鉴定这三台样机的效率有无显著差异? 四名技术工人的操作熟练水平有无显著差异 ($\alpha = 0.05$)?