# 第八章　方差分析与回归分析

本章前三节研究方差分析，讨论多个正态总体的比较，后两节研究回归分析。讨论两个变量之间的相关关系。

## §8.1　方差分析

### 8.1.1　问题的提出

上一章讨论了单个或两个正态总体的假设检验，这里讨论多个正态总体的均值比较问题。

通常为了研究某一因素对某项指标的影响情况，将该因素在多种情形下进行抽样检验，作出比较。一般将该因素称为一个因子，所检验的每种情形称为水平。在每个水平下需要考察的指标都分别构成一个总体，比较它们的总体均值是否相等。对每一个总体都分别抽取一个样本，样本容量称为重复数。

如果只对一个因子中的多个水平进行比较，称为单因子方差分析，对多个因子的水平进行比较，称为多因子方差分析。本章只进行单因子方差分析。

**例**　在饲料养鸡增肥的研究中，现有三种饲料配方：$A_1, A_2, A_3$，为比较三种饲料的效果，特选 24 只相似的雏鸡随机均分为三组，每组各喂一种饲料，60 天后观察它们的重量。实验结果如下表所示：

| 饲料 | 鸡重（g） | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $A_1$ | 1073 | 1009 | 1060 | 1001 | 1002 | 1012 | 1009 | 1028 |
| $A_2$ | 1107 | 1092 | 990 | 1109 | 1090 | 1074 | 1122 | 1001 |
| $A_3$ | 1093 | 1029 | 1080 | 1021 | 1022 | 1032 | 1029 | 1048 |

在此例中，就是要考察饲料对鸡增重的影响，需要比较三种饲料对鸡增肥的作用是否相同。这里，饲料就是一个因子，三种饲料配方就是该因子的三个水平，每种饲料喂养的雏鸡 60 天后的重量分别构成一个总体，这里共有 3 个总体，每一个总体抽取样本的重复数都是 8，比较这 3 个总体的均值是否相等。

### 8.1.2　单因子方差分析的统计模型

设因子 $A$ 有 $r$ 个水平 $A_1, A_2, \cdots, A_r$，在每个水平下需要考察的指标都构成一个总体，即有 $r$ 个总体，分别记为 $Y_1, Y_2, \cdots, Y_r$，对每一个总体都分别抽取一个样本，首先考虑重复数相等的情形，设重复数都是 $m$，总体 $Y_i$ 的样本 $Y_{i1}, Y_{i2}, \cdots, Y_{im}$，　$i = 1, 2, \cdots, r$。作出以下假定：

（1）每一个总体都服从正态分布，即 $Y_i \sim N(\mu_i, \sigma_i^2)$，　$i = 1, 2, \cdots, r$。

（2）各个总体的方差都相等，即 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$，都记为 $\sigma^2$；

（3）各个总体及抽取的样本相互独立，即 $Y_{ij}$，　$i = 1, 2, \cdots, r;\ j = 1, 2, \cdots, m$ 相互独立。

需要比较它们的总体均值是否相等，即检验的原假设为 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$。如果原假设 $H_0$ 成立，就可以认为这 $r$ 个水平下的总体均值相同，称为因子 $A$ 不显著；反之，如果原假设 $H_0$ 不成立，就称为因子 $A$ 显著。

在水平 $A_i$ 下的样品 $Y_{ij}$ 与该水平下的总体均值 $\mu_i$ 之差 $\varepsilon_{ij} = Y_{ij} - \mu_i$ 为随机误差。由于 $Y_{ij} \sim N(\mu_i, \sigma^2)$，

因此随机误差 $\varepsilon_{ij} \sim N(0,\sigma^2)$。对所有 $r$ 个水平下的总体均值求平均，即

$$\mu = \frac{1}{r}(\mu_1 + \mu_2 + \cdots + \mu_r) = \frac{1}{r}\sum_{i=1}^{r}\mu_i$$

称为总均值。每个水平 $A_i$ 下的总体期望 $\mu_i$ 与总均值 $\mu$ 之差 $a_i = \mu_i - \mu$ 称为该水平 $A_i$ 下主效应。显然所有主效应 $A$ 之和等于 0，即

$$\sum_{i=1}^{r} a_i = 0\ ,$$

检验所有水平下的总体均值是否相等，也就是检验所有主效应 $a_i$ 是否全等于 0。这样单因子方差分析在重复数相等的情形下，统计模型为

$$\begin{cases} Y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad i = 1,2,\cdots,r;\ j = 1,2,\cdots,m; \\ \sum_{i=1}^{r} a_i = 0; \\ \text{各 } \varepsilon_{ij} \text{ 相互独立，且服从相同的正态分布 } N(0,\sigma^2). \end{cases}$$

检验的原假设为 $H_0 : a_1 = a_2 = \cdots = a_r = 0$。

8.1.3　　平方和分解

一．模型参数的点估计

根据 $r$ 个总体下的试验数据 $Y_{ij}$，$i = 1,2,\cdots,r;\ j = 1,2,\cdots,m$，对统计模型中的参数 $\mu$、$\mu_i$ 与 $a_i$、$\varepsilon_{ij}$ 作出估计。记 $n = rm$ 表示总的样本容量，$T_i = \sum_{j=1}^{m} Y_{ij}$ 表示第 $i$ 个总体下试验数据总和，$\overline{Y}_{i\cdot} = \frac{T_i}{m}$ 表示第 $i$ 个总体下样本均值，$T = \sum_{i=1}^{r}\sum_{j=1}^{m} Y_{ij}$ 表示总的试验数据总和，$\overline{Y} = \frac{T}{n}$ 表示总的样本均值。

参数 $\mu$ 的点估计为 $\overline{Y}$，$\mu_i$ 的点估计为 $\overline{Y}_{i\cdot}$，$a_i = \mu_i - \mu$ 的点估计为 $\overline{Y}_{i\cdot} - \overline{Y}$，$\varepsilon_{ij} = Y_{ij} - \mu_i$ 的点估计为 $Y_{ij} - \overline{Y}_{i\cdot}$。

在单因子方差分析中通常将试验数据及基本计算结果写成实验数据计算表：

| 因子水平 | 试验数据 | | | | 和 | 和的平方 | 平方和 |
|---|---|---|---|---|---|---|---|
| $A_1$ | $Y_{11}$ | $Y_{12}$ | $\cdots$ | $Y_{1m}$ | $T_1$ | $T_1^2$ | $\sum_{j=1}^{m} Y_{1j}^2$ |
| $A_2$ | $Y_{21}$ | $Y_{22}$ | $\cdots$ | $Y_{2m}$ | $T_2$ | $T_2^2$ | $\sum_{j=1}^{m} Y_{2j}^2$ |
| $\vdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $A_r$ | $Y_{r1}$ | $Y_{r2}$ | $\cdots$ | $Y_{rm}$ | $T_r$ | $T_r^2$ | $\sum_{j=1}^{m} Y_{rj}^2$ |
| $\Sigma$ | | | | | $T$ | $\sum_{i=1}^{r} T_i^2$ | $\sum_{i=1}^{r}\sum_{j=1}^{m} Y_{ij}^2$ |

二．组内偏差与组间偏差

数据$Y_{ij}$与样本总均值$\bar{Y}$之差$Y_{ij}-\bar{Y}$称为样本总偏差，可以分成两部分之和：

$$Y_{ij}-\bar{Y}=(Y_{ij}-\bar{Y}_{i\cdot})+(\bar{Y}_{i\cdot}-\bar{Y})\,,$$

其中$\bar{Y}_{i\cdot}-\bar{Y}$是第$i$个总体的样本均值与总的样本均值的偏差，为主效应$a_i=\mu_i-\mu$的点估计，反映因子各水平之间的偏差，称为组间偏差；$Y_{ij}-\bar{Y}_{i\cdot}$是第$i$个总体内数据与该总体内样本均值的偏差，为随机误差$\varepsilon_{ij}=Y_{ij}-\mu_i$的点估计，反映随机因素造成的偏差，称为组内偏差。

三．偏差平方和及其自由度

在统计学中，对于$k$个独立数据$Y_1,Y_2,\cdots,Y_r$，平均值$\bar{Y}=\frac{1}{k}\sum_{i=1}^{k}Y_i$，称$Y_i$与$\bar{Y}$之差为偏差，所有偏差的平方和

$$Q=\sum_{i=1}^{k}(Y_i-\bar{Y})^2$$

称为这$k$个数据的偏差平方和，反映这$k$个数据的分散程度。由于所有偏差之和

$$\sum_{i=1}^{k}(Y_i-\bar{Y})=\sum_{i=1}^{k}Y_i-k\bar{Y}=0\,,$$

即这$k$个偏差由$k$个独立数据受到一个约束条件形成，可以证明它们与$k-1$个独立（随机）变量可以相互线性表示，称之为等价于$k-1$个独立（随机）变量。一般地，若$k$个独立数据受到$r$个不相关的约束条件，则它们等价于$k-r$个独立（随机）变量。在统计学中，把形成平方和的变量所等价的独立变量个数，称为该平方和的自由度，通常记为$f$。如上述偏差平方和$Q$的自由度为$k-1$，即$f_Q=k-1$。

由于平方和的大小与变量个数（或自由度）有关，为了对偏差进行比较，通常考虑偏差平方和与其自由度之商，称为均方和，记为$MS$，反映一组数据的平均分散程度，如样本方差$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$就是样本数据偏差的均方和。

四．总平方和分解公式

总偏差平方和记为$S_T$或$SST$，其自由度记为$f_T$，有

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2\,,\quad f_T=rm-1=n-1\,;$$

组内偏差平方和记为$S_e$或$SSE$，其自由度记为$f_e$，有

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_{i\cdot})^2\,,\quad f_e=rm-r=n-r\,;$$

组间偏差平方和记为$S_A$或$SSA$，其自由度记为$f_A$，有

$$S_A=\sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{Y}_{i\cdot}-\bar{Y})^2=m\sum_{j=1}^{m}(\bar{Y}_{i\cdot}-\bar{Y})^2\,,\quad f_A=r-1\,。$$

组内偏差平方和反映所有总体内的随机误差，组间偏差平方和反映所有总体的主效应。

**定理** 总偏差平方和 $S_T$ 可以分解为组内偏差平方和 $S_e$ 与组间偏差平方和 $S_A$ 之和，其自由度也可作相应的分解，即 $S_T = S_e + S_A$， $f_T = f_e + f_A$，称之为平方和分解公式。

**证明：**显然有

$$f_T = n-1 = (n-r)+(r-1) = f_e + f_A,$$

且直接展开验证可得

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{r}\sum_{j=1}^{m}[(Y_{ij}-\bar{Y}_{i\cdot})+(\bar{Y}_{i\cdot}-\bar{Y})]^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_{i\cdot})^2 + \sum_{i=1}^{r}\sum_{j=1}^{m}(\bar{Y}_{i\cdot}-\bar{Y})^2 + 2\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_{i\cdot})(\bar{Y}_{i\cdot}-\bar{Y})$$

$$= S_e + S_A + 2\sum_{i=1}^{r}\left[(\bar{Y}_{i\cdot}-\bar{Y})\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_{i\cdot})\right] = S_e + S_A + 2\sum_{i=1}^{r}\left[(\bar{Y}_{i\cdot}-\bar{Y})\times 0\right] = S_e + S_A。$$

8.1.4 检验方法

由于组内偏差平方和反映所有总体内的随机误差，组间偏差平方和反映所有总体的主效应，通过比较组内偏差平方和与组间偏差平方和检验因子的显著性。下面将证明在假设所有主效应都等于 0 成立的条件下，它们的均方和之商服从 $F$ 分布。

**定理** 在单因子方差分析模型中，组内偏差平方和 $S_e$ 与组间偏差平方和 $S_A$ 满足

（1） $E(S_e) = (n-r)\sigma^2$，且 $\dfrac{S_e}{\sigma^2} \sim \chi^2(n-r)$。

（2） $E(S_A) = (r-1)\sigma^2 + m\sum_{i=1}^{r} a_i^2$，且在 $H_0 : a_1 = a_2 = \cdots = a_r = 0$ 成立条件下，$\dfrac{S_A}{\sigma^2} \sim \chi^2(r-1)$。

（3） $S_e$ 与 $S_A$ 相互独立。

**证明：**根据第五章的定理结论知：设 $X_1, X_2, \cdots, X_n$ 相互独立且都服从正态分布 $N(\mu, \sigma^2)$，记

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{，} \quad S_0 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \text{，}$$

则 $\bar{X}$ 与 $S_0$ 相互独立，且

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \frac{S_0}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1)。$$

（1）因 $Y_{i1}, Y_{i2}, \cdots, Y_{im}$ 相互独立且都服从正态分布 $N(\mu_i, \sigma^2)$，$\bar{Y}_{i\cdot} = \dfrac{1}{m}\sum_{i=1}^{m} Y_{ij}$，则 $\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_{i\cdot})^2$ 与 $\bar{Y}_{i\cdot}$ 相互独立，且

$$\bar{Y}_{i\cdot} = \frac{1}{m}\sum_{i=1}^{m} Y_{ij} \sim N\left(\mu_i, \frac{\sigma^2}{m}\right), \quad \frac{1}{\sigma^2}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_{i\cdot})^2 \sim \chi^2(m-1)。$$

因在不同水平下的样本都相互独立，因此$\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_{i\cdot})^2,\ \ i=1,2,\cdots,r$之间相互独立。根据独立$\chi^2$分布的可加性知

$$\frac{1}{\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_{i\cdot})^2\sim\chi^2(rm-r)\,,$$

即

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

$$E(S_e)=(n-r)\sigma^2\,。$$

（2）因$\bar{Y}_1,\bar{Y}_2,\cdots,\bar{Y}_r$相互独立且

$$\bar{Y}_{i\cdot}-\mu_i\sim N\left(0,\frac{\sigma^2}{m}\right),\ \ \frac{1}{r}\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\mu_i)=\frac{1}{r}\sum_{i=1}^{r}\bar{Y}_{i\cdot}-\frac{1}{r}\sum_{i=1}^{r}\mu_i=\bar{Y}-\mu\,,$$

则

$$\frac{m}{\sigma^2}\sum_{i=1}^{r}[(\bar{Y}_{i\cdot}-\mu_i)-(\bar{Y}-\mu)]^2=\frac{m}{\sigma^2}\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y}-a_i)^2\sim\chi^2(r-1)\,,$$

因

$$\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y}-a_i)^2=\sum_{i=1}^{r}[(\bar{Y}_{i\cdot}-\bar{Y})^2-2a_i(\bar{Y}_{i\cdot}-\bar{Y})+a_i^2]=\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y})^2-2\sum_{i=1}^{r}a_i(\bar{Y}_{i\cdot}-\bar{Y})+\sum_{i=1}^{r}a_i^2\,,$$

则

$$r-1=\frac{m}{\sigma^2}E\left[\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y}-a_i)^2\right]=\frac{m}{\sigma^2}\left\{E\left[\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y})^2\right]-2\sum_{i=1}^{r}a_iE(\bar{Y}_{i\cdot}-\bar{Y})+\sum_{i=1}^{r}a_i^2\right\}$$

$$=\frac{m}{\sigma^2}\left\{E\left[\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y})^2\right]-2\sum_{i=1}^{r}a_i^2+\sum_{i=1}^{r}a_i^2\right\}$$

$$=\frac{1}{\sigma^2}\left\{E\left[m\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y})^2\right]-m\sum_{i=1}^{r}a_i^2\right\}=\frac{1}{\sigma^2}\left\{E(S_A)-m\sum_{i=1}^{r}a_i^2\right\},$$

故

$$E(S_A)=(r-1)\sigma^2+m\sum_{i=1}^{r}a_i^2\,。$$

且在$H_0:a_1=a_2=\cdots=a_r=0$成立条件下，

$$\frac{m}{\sigma^2}\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y}-a_i)^2=\frac{m}{\sigma^2}\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y})^2=\frac{S_A}{\sigma^2}\sim\chi^2(r-1)\,。$$

（3）因$S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y}_{i\cdot})^2$与$\bar{Y}_1,\bar{Y}_2,\cdots,\bar{Y}_r$相互独立，有$S_e$与$\bar{Y}=\frac{1}{r}\sum_{i=1}^{r}\bar{Y}_{i\cdot}$相互独立，且

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

故$S_e$与$S_A$相互独立。

**注：**（1）由于 $\frac{S_e}{\sigma^2}\sim\chi^2(n-r)$，在 $H_0:a_1=a_2=\cdots=a_r=0$ 成立条件下，$\frac{S_A}{\sigma^2}\sim\chi^2(r-1)$，且 $S_e$ 与 $S_A$ 相互独立，则根据 $F$ 分布的定义可知：在 $H_0$ 成立条件下，有

$$F=\frac{\frac{S_A}{\sigma^2}\Big/(r-1)}{\frac{S_e}{\sigma^2}\Big/(n-r)}=\frac{S_A/f_A}{S_e/f_e}=\frac{MS_A}{MS_e}\sim F(r-1,n-r)。$$

因 $E(S_A)=(r-1)\sigma^2+m\sum_{i=1}^{r}a_i^2$，当 $H_0$ 不成立时，$a_i$ 不全为零，$F$ 统计量的分子 $S_A$ 很可能变大，从而检验统计量观测值 $f$ 很可能变大，此检验的拒绝域在右侧，$W=\{f\geq f_{1-\alpha}(r-1,n-r)\}$。

（2）因 $E(S_e)=(n-r)\sigma^2$，即 $E\left(\frac{S_e}{n-r}\right)=\sigma^2$，故

$$\hat{\sigma}^2=\frac{S_e}{n-r}=MS_e$$

是误差方差 $\sigma^2$ 的无偏估计。

步骤：

（1）假设 $H_0:a_1=a_2=\cdots=a_r=0$。

（2）统计量 $F=\frac{S_A/f_A}{S_e/f_e}=\frac{MS_A}{MS_e}\sim F(r-1,n-r)$。

（3）拒绝域：$W=\{f\geq f_{1-\alpha}(r-1,n-r)\}$。

（4）计算检验统计量观测值 $f$ 与检验的 $p$ 值，并作出决策。

这是 $F$ 检验法。

通常将检验统计量观测值 $F$ 的计算过程列成方差分析表：

| 来源 | 平方和 | 自由度 | 均方和 | $F$ 比 | $p$ 值 |
|---|---|---|---|---|---|
| 因子 | $S_A$ | $f_A=r-1$ | $MS_A=S_A/f_A$ | $f=MS_A/MS_e$ | $p=P\{F\geq f\}$ |
| 误差 | $S_e$ | $f_e=n-r$ | $MS_e=S_e/f_e$ | | |
| 总和 | $S_T$ | $f_T=n-1$ | | | |

为了计算方便，可给出三个偏差平方和的计算公式。记

$$T_i=\sum_{j=1}^{m}Y_{ij}，\quad T=\sum_{i=1}^{r}T_i=\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}，$$

可得

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2=\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}^2-n\bar{Y}^2=\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}^2-\frac{1}{n}\left(\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}\right)^2=\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}^2-\frac{T^2}{n}，$$

$$S_A=m\sum_{i=1}^{r}(\bar{Y}_{i\cdot}-\bar{Y})^2=m\left(\sum_{i=1}^{r}\bar{Y}_{i\cdot}^2-r\bar{Y}^2\right)=m\sum_{i=1}^{r}\left(\frac{1}{m}\sum_{j=1}^{m}Y_{ij}\right)^2-mr\left(\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}\right)^2=\frac{1}{m}\sum_{i=1}^{r}T_i^2-\frac{T^2}{n}，$$

$$S_e=S_T-S_A=\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}^2-\frac{1}{m}\sum_{i=1}^{r}T_i^2。$$

**例** 在饲料养鸡增肥的研究中，现有三种饲料配方：$A_1, A_2, A_3$，为比较三种饲料的效果，特选 24 只相似的雏鸡随机均分为三组，每组各喂一种饲料，60 天后观察它们的重量。实验结果如下表所示：

| 饲料 | 鸡重（g） | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $A_1$ | 1073 | 1009 | 1060 | 1001 | 1002 | 1012 | 1009 | 1028 |
| $A_2$ | 1107 | 1092 | 990 | 1109 | 1090 | 1074 | 1122 | 1001 |
| $A_3$ | 1093 | 1029 | 1080 | 1021 | 1022 | 1032 | 1029 | 1048 |

在显著水平 $\alpha = 0.05$ 下检验这三种饲料对雏鸡增重是否有显著差异。

**解：**单因子方差分析，假设 $H_0 : a_1 = a_2 = a_3 = 0$，统计量 $F = \dfrac{S_A / f_A}{S_e / f_e} \sim F(r-1, n-r)$，显著水平 $\alpha = 0.05$，

$r = 3$， $m = 8$， $n = 24$， $f_{1-\alpha}(r-1, n-r) = f_{0.95}(2, 21) = 3.47$，右侧拒绝域 $W = \{f \geq 3.47\}$。

试验数据计算表

| 因子水平 | 试验数据 $Y_{ij}$ | | | | | | | | $T_i$ | $T_i^2$ | $\sum_{j=1}^{m} Y_{ij}^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $A_1$ | 1073 | 1009 | 1060 | 1001 | 1002 | 1012 | 1009 | 1028 | 8194 | 67141636 | 8398024 |
| $A_2$ | 1107 | 1092 | 990 | 1109 | 1090 | 1074 | 1122 | 1001 | 8585 | 73702225 | 9230355 |
| $A_3$ | 1093 | 1029 | 1080 | 1021 | 1022 | 1032 | 1029 | 1048 | 8354 | 69789316 | 8728984 |
| Σ | | | | | | | | | 25133 | 210633177 | 26357363 |

计算可得

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{m} Y_{ij}^2 - \frac{T^2}{n} = 26357363 - \frac{1}{24} \times 25133^2 = 37875.9583 ,$$

$$S_A = \frac{1}{m}\sum_{i=1}^{r} T_i^2 - \frac{T^2}{n} = \frac{1}{8} \times 210633177 - \frac{1}{24} \times 25133^2 = 9660.0833 ,$$

$$S_e = S_T - S_A = 37875.9583 - 9660.0833 = 28215.8750 。$$

方差分析表

| 来源 | 平方和 | 自由度 | 均方和 | $F$ 比 | $p$ 值 |
|---|---|---|---|---|---|
| 因子 | 9660.0833 | 2 | 4830.0417 | 3.5948 | 0.0454 |
| 误差 | 28215.8750 | 21 | 1343.6131 | | |
| 总和 | 37875.9583 | 23 | | | |

可得检验统计量观测值 $F$ 比 $f = 3.5948 \in W$，并且检验的 $p$ 值

$$p = P\{F \geq 3.5948\} = 0.0454 < \alpha = 0.05 ,$$

故拒绝 $H_0$，接受 $H_1$。可以认为这三种饲料对雏鸡增重有显著差异。

### 8.1.5 参数估计

在方差分析问题中，可对统计模型中的参数总均值 $\mu$、误差方差 $\sigma^2$ 作出估计。

当检验结果为因子不显著时，各水平下指标的总体均值与总体方差都相同，可将所有水平的指标看作

一个统一的总体，全部试验数据是来自正态总体 $Y\sim N(\mu,\sigma^2)$ 的一个容量为 $n=rm$ 的样本，因此样本均值与样本方差分别为

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij}=\frac{T}{n},\quad S^2=\frac{1}{n-1}\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\bar{Y})^2=\frac{S_T}{n-1},$$

这样总均值 $\mu$ 和误差方差 $\sigma^2$ 的点估计分别为 $\hat{\mu}=\bar{Y}$，$\hat{\sigma^2}=S^2$，置信度为 $1-\alpha$ 的置信区间分别为

$$\mu\in\left[\bar{Y}\pm t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}\right],\quad \sigma^2\in\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right]。$$

当检验结果为因子显著时，除了总均值 $\mu$、误差方差 $\sigma^2$ 外，还可进一步对主效应 $a_i$ 作参数估计。

一．点估计

因试验数据 $Y_{ij}$，$i=1,2,\cdots,r;\ j=1,2,\cdots,m$ 相互独立且都服从正态分布 $N(\mu+a_i,\sigma^2)$，根据最大似然估计法，可得到总均值 $\mu$，主效应 $a_i$ 及误差方差 $\sigma^2$ 的点估计。似然函数

$$L(\mu,a_1,a_2,\cdots,a_r,\sigma^2)=\prod_{i=1}^{r}\prod_{j=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_{ij}-\mu-a_i)^2}{2\sigma^2}}=\frac{1}{(\sqrt{2\pi})^n\sigma^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\mu-a_i)^2},$$

取对数，得

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\mu-a_i)^2,$$

令关于 $\mu$ 的偏导数等于 0，有

$$\frac{\partial\ln L}{\partial\mu}=-\frac{1}{2\sigma^2}\sum_{i=1}^{r}\sum_{j=1}^{m}2(y_{ij}-\mu-a_i)\cdot(-1)=\frac{1}{\sigma^2}\left(\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}-n\mu-0\right)=\frac{1}{\sigma^2}\left(\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}-n\mu\right)=0,$$

得

$$\mu=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}y_{ij}=\bar{y},$$

故总均值 $\mu$ 的最大似然估计为 $\hat{\mu}=\bar{Y}$。

令关于 $a_k$ 的偏导数等于 0，有

$$\frac{\partial\ln L}{\partial a_k}=-\frac{1}{2\sigma^2}\sum_{j=1}^{m}2(y_{kj}-\mu-a_k)\cdot(-1)=\frac{1}{\sigma^2}\left(\sum_{j=1}^{m}y_{kj}-m\mu-ma_k\right)=0,\quad k=1,2,\cdots,r,$$

得

$$a_k=\frac{1}{m}\sum_{j=1}^{m}y_{kj}-\mu=\bar{y}_{k\cdot}-\mu,$$

故主效应 $a_i$ 的最大似然估计为

$\hat{a}_i = \bar{Y}_{i\cdot} - \hat{\mu} = \bar{Y}_{i\cdot} - \bar{Y}$，$i = 1, 2, \cdots, r$。

令关于$\sigma^2$的偏导数等于 0，有

$$\frac{\partial \ln L}{\partial(\sigma^2)} = -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \mu - a_i)^2 = 0\ ,$$

得

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij} - \mu - a_i)^2\ ,$$

故误差方差$\sigma^2$的最大似然估计为

$$\hat{\sigma}_M^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij} - \hat{\mu} - \hat{a}_i)^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_{i\cdot})^2 = \frac{S_e}{n}\ 。$$

但由于$E(S_e) = (n-r)\sigma^2$，可知$\hat{\sigma}_M^2$不是$\sigma^2$的无偏估计，修偏得$\sigma^2$的无偏估计

$$\hat{\sigma}^2 = \frac{S_e}{n-r} = MS_e\ 。$$

二．置信区间

进一步给出总均值$\mu$，各水平的总体均值$\mu_i$及误差方差$\sigma^2$的置信区间。

总均值$\mu$的点估计$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}Y_{ij} = \bar{Y}$。因试验数据$Y_{ij}$，$i = 1, 2, \cdots, r;\ j = 1, 2, \cdots, m$相互独立且都服从正态分布$N(\mu_i, \sigma^2)$，则有$\bar{Y}$服从正态分布，且

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}E(Y_{ij}) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m}\mu_i = \frac{1}{n}\sum_{i=1}^{r}m\mu_i = \frac{1}{r}\sum_{i=1}^{r}\mu_i = \mu\ ,$$

$$\mathrm{Var}(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{r}\sum_{j=1}^{m}\mathrm{Var}(Y_{ij}) = \frac{1}{n^2}\sum_{i=1}^{r}\sum_{j=1}^{m}\sigma^2 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}\ ,$$

故$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，即$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$。但$\sigma$未知，用$\hat{\sigma} = \sqrt{\frac{S_e}{n-r}}$替换。由于$\frac{S_e}{\sigma^2} \sim \chi^2(n-r)$且$S_e$与$\bar{Y}$相互独立，则根据$t$分布的定义可得

$$T = \frac{\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S_e}{\sigma^2}\Big/(n-r)}} = \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t(n-r)\ 。$$

根据$P\left\{-t_{1-\alpha/2}(n-r) \le \frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \le t_{1-\alpha/2}(n-r)\right\} = 1-\alpha$可得总均值$\mu$的$1-\alpha$的置信区间是

$$\left[\bar{Y} \pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right]。$$

第 $i$ 个水平的总体均值 $\mu_i$ 的点估计为 $\hat{\mu}_i=\frac{1}{m}\sum_{j=1}^{m}Y_{ij}=\overline{Y}_{i\cdot}$，因试验数据 $Y_{ij}$，$i=1,2,\cdots,r;\ j=1,2,\cdots,m$ 相互独立且都服从正态分布 $N(\mu_i,\sigma^2)$，则有 $\overline{Y}_{i\cdot}\sim N\left(\mu_i,\frac{\sigma^2}{m}\right)$，即 $\frac{\overline{Y}_{i\cdot}-\mu_i}{\sigma/\sqrt{m}}\sim N(0,1)$。但 $\sigma$ 未知，用 $\hat{\sigma}=\sqrt{\frac{S_e}{n-r}}$ 替换。由于 $\frac{S_e}{\sigma^2}\sim\chi^2(n-r)$ 且 $S_e$ 与 $\overline{Y}_{i\cdot}$ 相互独立，则根据 $t$ 分布的定义可得

$$T=\frac{\frac{\overline{Y}_{i\cdot}-\mu_i}{\sigma/\sqrt{m}}}{\sqrt{\frac{S_e}{\sigma^2}\Big/(n-r)}}=\frac{\overline{Y}_{i\cdot}-\mu_i}{\hat{\sigma}/\sqrt{m}}\sim t(n-r)。$$

根据 $P\left\{-t_{1-\alpha/2}(n-r)\le\frac{\overline{Y}_{i\cdot}-\mu_i}{\hat{\sigma}/\sqrt{m}}\le t_{1-\alpha/2}(n-r)\right\}=1-\alpha$ 可得第 $i$ 个水平的总体均值 $\mu_i$ 的 $1-\alpha$ 的置信区间是

$$\left[\overline{Y}_{i\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m}}\right],\quad i=1,2,\cdots,r。$$

误差方差 $\sigma^2$ 的点估计为 $\hat{\sigma^2}=\frac{S_e}{n-r}$，且 $\frac{S_e}{\sigma^2}\sim\chi^2(n-r)$，根据 $P\left\{\chi^2_{\alpha/2}(n-r)\le\frac{S_e}{\sigma^2}\le\chi^2_{1-\alpha/2}(n-r)\right\}=1-\alpha$ 可得误差方差 $\sigma^2$ 的 $1-\alpha$ 置信区间是

$$\left[\frac{S_e}{\chi^2_{1-\alpha/2}(n-r)},\frac{S_e}{\chi^2_{\alpha/2}(n-r)}\right]。$$

**例** 由前面的鸡饲料对鸡增重问题的数据给出总均值 $\mu$，三个水平的总体均值 $\mu_1,\mu_2,\mu_3$ 及误差方差 $\sigma^2$ 的点估计和 0.95 置信区间。

**解：**经检验知因子显著，则总均值 $\mu$、三个水平的总体均值 $\mu_1,\mu_2,\mu_3$、误差方差 $\sigma^2$ 的点估计分别为

$$\hat{\mu}=\overline{Y}=\frac{T}{n}=\frac{25133}{24}=1047.2083,$$
$$\hat{\mu}_1=\overline{Y}_{1\cdot}=\frac{T_1}{m}=\frac{8194}{8}=1024.25,$$
$$\hat{\mu}_2=\overline{Y}_{2\cdot}=\frac{T_2}{m}=\frac{8585}{8}=1073.125,$$
$$\hat{\mu}_3=\overline{Y}_{3\cdot}=\frac{T_3}{m}=\frac{8354}{8}=1044.25,$$
$$\hat{\sigma^2}=\frac{S_e}{n-r}=1343.6131;$$

又因它们的 0.95 置信区间分别为

$$\mu\in\left[\overline{Y}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right]=\left[1047.2083\pm2.0796\times\frac{\sqrt{1343.6131}}{\sqrt{24}}\right]=[1031.6482,1062.7684],$$

$$\mu_1=\left[\overline{Y}_{1\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m}}\right]=\left[1024.25\pm2.0796\times\frac{\sqrt{1343.6131}}{\sqrt{8}}\right]=[997.2992,1051.2008],$$

$$\mu_2=\left[\bar{Y}_{2\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m}}\right]=\left[1073.125\pm 2.0796\times\frac{\sqrt{1343.6131}}{\sqrt{8}}\right]=[1046.1742,1100.0758]，$$

$$\mu_3=\left[\bar{Y}_{3\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m}}\right]=\left[1044.25\pm 2.0796\times\frac{\sqrt{1343.6131}}{\sqrt{8}}\right]=[1017.2992,1071.2008]，$$

$$\sigma^2\in\left[\frac{S_e}{\chi^2_{1-\alpha/2}(n-r)},\frac{S_e}{\chi^2_{\alpha/2}(n-r)}\right]=\left[\frac{28215.875}{35.4789},\frac{28215.875}{10.2829}\right]=[795.2861,2743.9608]。$$

8.1.6　重复数不等的情形

如果每个水平下试验次数不全相等，称为重复数不等的情形，其检验方法与在重复数相等的情形下类似，只是在对数据的表述和处理上有几点区别。

一．数据

设第 $i$ 个水平 $A_i$ 下的重复数为 $m_i$，所取得的样本为 $Y_{i1},Y_{i2},\cdots,Y_{im_i}$，$i=1,2,\cdots,r$。显然重复数总数为 $m_1+m_2+\cdots+m_r=n$。

二．总均值

总均值 $\mu$ 是各水平下总体均值 $\mu_i$ 的以重复数的比率 $\frac{m_i}{n}$ 为权数的加权平均，即

$$\mu=\frac{m_1}{n}\mu_1+\frac{m_2}{n}\mu_2+\cdots+\frac{m_r}{n}\mu_r=\frac{1}{n}\sum_{i=1}^{r}m_i\mu_i。$$

三．主效应约束条件

设第 $i$ 个水平 $A_i$ 下主效应 $a_i=\mu_i-\mu$，则满足

$$\sum_{i=1}^{r}m_i\mu_i=\sum_{i=1}^{r}m_ia_i-n\mu=0。$$

四．模型

单因子方差分析在重复数不等的情形下，统计模型为

$$\begin{cases}Y_{ij}=\mu+a_i+\varepsilon_{ij},\quad i=1,2,\cdots,r;\ j=1,2,\cdots,m_i;\\ \sum_{i=1}^{r}m_ia_i=0;\\ \text{各 }\varepsilon_{ij}\text{ 相互独立，且服从相同的正态分布 }N(0,\sigma^2).\end{cases}$$

检验的原假设为 $H_0:a_1=a_2=\cdots=a_r=0$。

五．平方和的计算

记

$$T_i=\sum_{j=1}^{m_i}Y_{ij}，\quad \bar{Y}_{i\cdot}=\frac{1}{m_i}\sum_{j=1}^{m_i}Y_{ij}=\frac{T_i}{m_i}，\quad T=\sum_{i=1}^{r}\sum_{j=1}^{m_i}Y_{ij}=\sum_{i=1}^{r}T_i，\quad \bar{Y}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{m_i}Y_{ij}=\frac{1}{n}\sum_{i=1}^{r}T_i=\frac{T}{n}，$$

可得各平方和的计算公式

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m_i}(Y_{ij}-\bar{Y})^2=\sum_{i=1}^{r}\sum_{j=1}^{m_i}Y_{ij}^2-n\bar{Y}^2=\sum_{i=1}^{r}\sum_{j=1}^{m_i}Y_{ij}^2-\frac{T^2}{n}，$$

$$S_A=\sum_{i=1}^{r}\sum_{j=1}^{m_i}(\bar{Y}_{i\cdot}-\bar{Y})^2=\sum_{i=1}^{r}m_i(\bar{Y}_{i\cdot}-\bar{Y})^2=\sum_{i=1}^{r}m_i\bar{Y}_{i\cdot}^2-n\bar{Y}^2=\sum_{i=1}^{r}\frac{T_i^2}{m_i}-\frac{T^2}{n},$$

$$S_e=S_T-S_A=\sum_{i=1}^{r}\sum_{j=1}^{m_i}Y_{ij}^2-\sum_{i=1}^{r}\frac{T_i^2}{m_i}。$$

**例** 某食品公司对一种食品设计了四种新包装，为了考察哪种包装最受顾客欢迎，选了 10 个地段繁华程度相似、规模相近的商店做试验，其中两种包装各指定两个商店销售，另两种包装各指定三个商店销售。在试验期内各店货架排放的位置、空间都相同，营业员的促销方法也基本相同，经过一段时间，记录其销售量数据，见下表

| 包装类型 | 销售量数据 | | |
|---|---|---|---|
| $A_1$ | 12 | 18 | |
| $A_2$ | 14 | 12 | 13 |
| $A_3$ | 19 | 17 | 21 |
| $A_4$ | 24 | 30 | |

在显著水平 $\alpha=0.01$ 下检验这四种包装对销售量是否有显著影响，并给出各参数的点估计和 0.99 置信区间。

**解：** 单因子方差分析，假设 $H_0:a_1=a_2=a_3=a_4=0$，统计量 $F=\dfrac{S_A/f_A}{S_e/f_e}\sim F(r-1,n-r)$，显著水平 $\alpha=0.01$，$r=4$，$n=10$，$f_{1-\alpha}(r-1,n-r)=f_{0.99}(3,6)=9.78$，右侧拒绝域 $W=\{f\geq 9.78\}$。

销售量数据计算表

| 因子水平 | 销售量数据 $Y_{ij}$ | | | $m_i$ | $T_i$ | $T_i^2/m_i$ | $\sum_{j=1}^{m}Y_{ij}^2$ |
|---|---|---|---|---|---|---|---|
| $A_1$ | 12 | 18 | | 2 | 30 | 450 | 468 |
| $A_2$ | 14 | 12 | 13 | 3 | 39 | 507 | 509 |
| $A_3$ | 19 | 17 | 21 | 3 | 57 | 1083 | 1091 |
| $A_4$ | 24 | 30 | | 2 | 54 | 1458 | 1476 |
| Σ | | | | 10 | 180 | 3498 | 3544 |

计算可得

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{m_i}Y_{ij}^2-\frac{T^2}{n}=3544-\frac{1}{10}\times 180^2=304,$$

$$S_A=\sum_{i=1}^{r}\frac{T_i^2}{m_i}-\frac{T^2}{n}=3498-\frac{1}{10}\times 180^2=258,$$

$$S_e=S_T-S_A=304-258=46。$$

方差分析表

| 来源 | 平方和 | 自由度 | 均方和 | $F$ 比 | $p$ 值 |
|---|---|---|---|---|---|
| 因子 | 258 | 3 | 86 | 11.2174 | 0.0071 |
| 误差 | 46 | 6 | 7.6667 | | |
| 总和 | 304 | 9 | | | |

可得检验统计量观测值 $F$ 比 $f=11.2174\in W$，并且检验的 $p$ 值

$$p=P\{F\geq 11.2174\}=0.0071<\alpha=0.01,$$

故拒绝 $H_0$，接受 $H_1$。可以认为这四种包装对销售量有显著影响。

由于因子显著，则总均值 $\mu$、四个水平的总体均值 $\mu_1,\mu_2,\mu_3,\mu_4$、误差方差 $\sigma^2$ 的点估计分别为

$$\hat{\mu}=\bar{Y}=\frac{T}{n}=\frac{180}{10}=18,$$

$$\hat{\mu}_1=\bar{Y}_{1\cdot}=\frac{T_1}{m_1}=\frac{30}{2}=15,$$

$$\hat{\mu}_2=\bar{Y}_{2\cdot}=\frac{T_2}{m_2}=\frac{39}{3}=13,$$

$$\hat{\mu}_3=\bar{Y}_{3\cdot}=\frac{T_3}{m_3}=\frac{57}{3}=19,$$

$$\hat{\mu}_4=\bar{Y}_{4\cdot}=\frac{T_4}{m_4}=\frac{54}{2}=27,$$

$$\hat{\sigma^2}=\frac{S_e}{n-r}=7.6667;$$

它们的 0.99 置信区间分别为

$$\mu\in\left[\bar{Y}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right]=\left[18\pm 3.7074\times\frac{\sqrt{7.6667}}{\sqrt{10}}\right]=[14.7538,21.2462],$$

$$\mu_1=\left[\bar{Y}_{1\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m_1}}\right]=\left[15\pm 3.7074\times\frac{\sqrt{7.6667}}{\sqrt{2}}\right]=[7.7413,22.2587],$$

$$\mu_2=\left[\bar{Y}_{2\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m_2}}\right]=\left[13\pm 3.7074\times\frac{\sqrt{7.6667}}{\sqrt{3}}\right]=[7.0733,18.9267],$$

$$\mu_3=\left[\bar{Y}_{3\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m_3}}\right]=\left[19\pm 3.7074\times\frac{\sqrt{7.6667}}{\sqrt{3}}\right]=[13.0733,24.9267],$$

$$\mu_4=\left[\bar{Y}_{4\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\frac{\hat{\sigma}}{\sqrt{m_4}}\right]=\left[27\pm 3.7074\times\frac{\sqrt{7.6667}}{\sqrt{2}}\right]=[19.7413,34.2587],$$

$$\sigma^2\in\left[\frac{S_e}{\chi^2_{1-\alpha/2}(n-r)},\frac{S_e}{\chi^2_{\alpha/2}(n-r)}\right]=\left[\frac{46}{18.5476},\frac{46}{0.6757}\right]=[2.4801,68.0775]。$$

## §8.2　多重比较

上一节是将多个总体作为一个整体进行检验。如果检验结果是因子 $A$ 显著，则可以认为各水平下的总体均值 $\mu_i$ 不全相等，但却不能直接说明 $\mu_i$ 中哪些可以认为相等，哪些可以认为不等。这一节是对各个 $\mu_i$ 两两之间进行比较，对 $\mu_i-\mu_j$，也就是效应差 $a_i-a_j$ 作出估计、检验。

### 8.2.1　效应差的置信区间

效应差 $a_i-a_j=\mu_i-\mu_j$ 的点估计为 $\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}$。因试验数据 $Y_{ik}$，$i=1,2,\cdots,r;k=1,2,\cdots,m_i$ 相互独立且都服从正态分布 $N(\mu_i,\sigma^2)$，则

$$\bar{Y}_{i\cdot}=\frac{1}{m_i}\sum_{k=1}^{m_i}Y_{ik}\sim N\left(\mu_i,\frac{\sigma^2}{m_i}\right),\quad \bar{Y}_{j\cdot}=\frac{1}{m_j}\sum_{k=1}^{m_j}Y_{jk}\sim N\left(\mu_j,\frac{\sigma^2}{m_j}\right),$$

且当 $i\neq j$ 时，$\bar{Y}_{i\cdot}$ 与 $\bar{Y}_{j\cdot}$ 相互独立，可得

$$\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}\sim N\left(\mu_i-\mu_j,\sigma^2\left(\frac{1}{m_i}+\frac{1}{m_j}\right)\right),$$

即

$$\frac{(\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot})-(\mu_i-\mu_j)}{\sigma\sqrt{\dfrac{1}{m_i}+\dfrac{1}{m_j}}}\sim N(0,1),$$

但 $\sigma$ 未知，用 $\hat{\sigma}=\sqrt{\dfrac{S_e}{n-r}}$ 替换。由于 $\dfrac{S_e}{\sigma^2}\sim\chi^2(n-r)$ 且 $S_e$ 与 $\bar{Y}_{i\cdot},\bar{Y}_{j\cdot}$ 相互独立，则根据 $t$ 分布的定义可得

$$T=\frac{\dfrac{(\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot})-(\mu_i-\mu_j)}{\sigma\sqrt{\dfrac{1}{m_i}+\dfrac{1}{m_j}}}}{\sqrt{\dfrac{S_e}{\sigma^2}\Big/(n-r)}}=\frac{(\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot})-(\mu_i-\mu_j)}{\hat{\sigma}\sqrt{\dfrac{1}{m_i}+\dfrac{1}{m_j}}}\sim t(n-r),$$

故效应差 $a_i-a_j=\mu_i-\mu_j$ 的置信度为 $1-\alpha$ 的置信区间是

$$\mu_i-\mu_j\in\left[\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}\pm t_{1-\alpha/2}(n-r)\cdot\hat{\sigma}\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}\right]。$$

**例**　由前面的鸡饲料对鸡增重问题的数据给出各效应差 $\mu_i-\mu_j$ 的点估计和 0.95 置信区间。

**解：**因 $r=3$，$m_1=m_2=m_3=8$，$n=24$，且

$$\bar{Y}_{1\cdot}=\frac{T_1}{m_1}=\frac{8194}{8}=1024.25,\quad \bar{Y}_{2\cdot}=\frac{T_2}{m_2}=\frac{8585}{8}=1073.125,\quad \bar{Y}_{3\cdot}=\frac{T_3}{m_3}=\frac{8354}{8}=1044.25,$$

则各效应差 $\mu_i-\mu_j$ 的点估计分别为

$$\hat{\mu}_1-\hat{\mu}_2=\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}=1024.25-1073.125=-48.875,$$

$$\hat{\mu}_1-\hat{\mu}_3=\overline{Y}_{1\cdot}-\overline{Y}_{3\cdot}=1024.25-1044.25=-20,$$

$$\hat{\mu}_2-\hat{\mu}_3=\overline{Y}_{2\cdot}-\overline{Y}_{3\cdot}=1073.125-1044.25=28.875。$$

又因

$$t_{1-\alpha/2}(n-r)=t_{0.975}(21)=2.0796,\quad \hat{\sigma}=\sqrt{\frac{S_e}{n-r}}=\sqrt{\frac{28215.875}{21}}=36.6553,$$

则各效应差 $\mu_i-\mu_j$ 的 0.95 置信区间分别是

$$\mu_1-\mu_2\in\left[-48.875\pm2.0796\times36.6553\times\sqrt{\frac{1}{8}+\frac{1}{8}}\right]=[-86.9892,-10.7608],$$

$$\mu_1-\mu_3\in\left[-20\pm2.0796\times36.6553\times\sqrt{\frac{1}{8}+\frac{1}{8}}\right]=[-58.1142,18.1142],$$

$$\mu_2-\mu_3\in\left[28.875\pm2.0796\times36.6553\times\sqrt{\frac{1}{8}+\frac{1}{8}}\right]=[-9.2392,66.9892]。$$

**例** 由前面的食品包装对销售量影响问题的数据给出各效应差 $\mu_i-\mu_j$ 的点估计和 0.99 置信区间。

**解：**因 $r=4$，$m_1=2$，$m_2=3$，$m_3=3$，$m_4=2$，$n=10$，且

$$\overline{Y}_{1\cdot}=\frac{T_1}{m_1}=\frac{30}{2}=15,\quad \overline{Y}_{2\cdot}=\frac{T_2}{m_2}=\frac{39}{3}=13,\quad \overline{Y}_{3\cdot}=\frac{T_3}{m_3}=\frac{57}{3}=19,\quad \overline{Y}_{4\cdot}=\frac{T_4}{m_4}=\frac{54}{2}=27,$$

则各效应差 $\mu_i-\mu_j$ 的点估计分别为

$$\hat{\mu}_1-\hat{\mu}_2=\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}=15-13=2,$$

$$\hat{\mu}_1-\hat{\mu}_3=\overline{Y}_{1\cdot}-\overline{Y}_{3\cdot}=15-19=-4,$$

$$\hat{\mu}_1-\hat{\mu}_4=\overline{Y}_{1\cdot}-\overline{Y}_{4\cdot}=15-27=-12,$$

$$\hat{\mu}_2-\hat{\mu}_3=\overline{Y}_{2\cdot}-\overline{Y}_{3\cdot}=13-19=-6,$$

$$\hat{\mu}_2-\hat{\mu}_4=\overline{Y}_{2\cdot}-\overline{Y}_{4\cdot}=13-27=-14,$$

$$\hat{\mu}_3-\hat{\mu}_4=\overline{Y}_{3\cdot}-\overline{Y}_{4\cdot}=19-27=-8。$$

又因

$t_{1-\alpha/2}(n-r)=t_{0.995}(6)=3.7074$，$\hat{\sigma}=\sqrt{\frac{S_e}{n-r}}=\sqrt{\frac{46}{6}}=2.7689$，

则各效应差 $\mu_i-\mu_j$ 的 0.99 置信区间分别是

$$\mu_1-\mu_2\in\left[2\pm3.7074\times2.7689\times\sqrt{\frac{1}{2}+\frac{1}{3}}\right]=[-7.3709,11.3709]，$$

$$\mu_1-\mu_3\in\left[-4\pm3.7074\times2.7689\times\sqrt{\frac{1}{2}+\frac{1}{3}}\right]=[-13.3709,5.3709]，$$

$$\mu_1-\mu_4\in\left[-12\pm3.7074\times2.7689\times\sqrt{\frac{1}{2}+\frac{1}{2}}\right]=[-22.2653,-1.7347]，$$

$$\mu_2-\mu_3\in\left[-6\pm3.7074\times2.7689\times\sqrt{\frac{1}{3}+\frac{1}{3}}\right]=[-14.3816,2.3816]，$$

$$\mu_2-\mu_4\in\left[-14\pm3.7074\times2.7689\times\sqrt{\frac{1}{3}+\frac{1}{2}}\right]=[-23.3709,-4.6291]，$$

$$\mu_3-\mu_4\in\left[-8\pm3.7074\times2.7689\times\sqrt{\frac{1}{3}+\frac{1}{2}}\right]=[-17.3709,1.3709]。$$

8.2.2　多重比较问题

对各个 $\mu_i$ 两两之间进行比较，也就是检验任意两个水平 $A_i$ 与 $A_j$ 下的总体均值是否相等，即检验假设

$H_0^{ij}:\mu_i=\mu_j$，$i,j=1,2,\cdots,r$。

对于每一个假设 $H_0^{ij}$ 可以采取两个正态总体的均值比较方法进行检验，但这里需要同时检验 $C_r^2=\frac{r(r-1)}{2}$ 个这种假设。

设需要同时检验 $k$ 个假设 $H_0^i$，$i=1,2,\cdots,k$，每一个假设的显著水平是 $\alpha$，即在 $H_0^i$ 成立的条件下，拒绝 $H_0^i$ 的概率为小概率 $\alpha$，但在所有 $k$ 个假设 $H_0^i$ 都成立的条件下，至少拒绝其中一个假设的概率最大时可能达到 $k\alpha$，这可能就不再是小概率事件了，不再适用小概率原理进行判断。

可见，需要同时检验多个假设时，一般不应逐个检验每一个假设，而是采用多重比较方法同时检验多个假设。多重比较方法，就是针对所有假设，构造一个统一的拒绝域，再逐个进行比较判断。

这里，对于假设 $H_0^{ij}:\mu_i=\mu_j$，$1\le i<j\le r$，在 $H_0^{ij}$ 成立的条件下，$\bar{Y}_{i\cdot}$ 与 $\bar{Y}_{j\cdot}$ 不应相差太大。对每一个假设 $H_0^{ij}$，拒绝域可以取为 $W^{ij}=\{|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|\ge c_{ij}\}$ 的形式，其中 $c_{ij}$ 是常数。对所有的假设 $H_0^{ij}$，$i,j=1,2,\cdots,r$，统一的拒绝域取为

$$W=\bigcup_{1\le i<j\le r}W^{ij}=\bigcup_{1\le i<j\le r}\{|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|\ge c_{ij}\}。$$

分成重复数相等与不等两种场合进行讨论。

8.2.3　　重复数相等场合的$T$法

重复数相等时，各水平是平等的，由对称性，可以要求所有的常数$c_{ij}$相等，记为$c$，即统一的拒绝域为

$$W=\bigcup_{1\le i<j\le r}\{|\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}|\ge c\}=\left\{\max_{1\le i<j\le r}|\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}|\ge c\right\}=\left\{\max_{1\le i\le r}\overline{Y}_{i\cdot}-\min_{1\le i\le r}\overline{Y}_{i\cdot}\ge c\right\}。$$

因试验数据$Y_{ij}$，$i=1,2,\cdots,r;\ j=1,2,\cdots,m$相互独立且都服从正态分布$N(\mu_i,\sigma^2)$，则有$\overline{Y}_{i\cdot}\sim N\left(\mu_i,\dfrac{\sigma^2}{m}\right)$。

当所有假设$H_0^{ij}$都成立时，即$\mu_1=\mu_2=\cdots=\mu_r=\mu$，有$\overline{Y}_{i\cdot}\sim N\left(\mu,\dfrac{\sigma^2}{m}\right)$，则

$$\frac{\overline{Y}_{i\cdot}-\mu}{\sigma/\sqrt{m}}\sim N(0,1)。$$

但$\sigma$未知，用$\hat{\sigma}=\sqrt{\dfrac{S_e}{n-r}}$替换。由于$\dfrac{S_e}{\sigma^2}\sim\chi^2(n-r)$且$S_e$与$\overline{Y}_{i\cdot}$相互独立，则根据$t$分布的定义可得

$$T=\frac{\dfrac{\overline{Y}_{i\cdot}-\mu}{\sigma/\sqrt{m}}}{\sqrt{\dfrac{S_e}{\sigma^2}\Big/(n-r)}}=\frac{\overline{Y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}\sim t(n-r)=t(f_e)，$$

统一的拒绝域$W$的形式可改写为

$$W=\left\{\max_{1\le i\le r}\overline{Y}_{i\cdot}-\min_{1\le i\le r}\overline{Y}_{i\cdot}\ge c\right\}=\left\{\max_{1\le i\le r}\frac{\overline{Y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}-\min_{1\le i\le r}\frac{\overline{Y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}\ge\frac{c}{\hat{\sigma}/\sqrt{m}}\right\}，$$

其中

$$Q=\max_{1\le i\le r}\frac{\overline{Y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}-\min_{1\le i\le r}\frac{\overline{Y}_{i\cdot}-\mu}{\hat{\sigma}/\sqrt{m}}=\frac{\max\limits_{1\le i\le r}\overline{Y}_{i\cdot}-\min\limits_{1\le i\le r}\overline{Y}_{i\cdot}}{\hat{\sigma}/\sqrt{m}}$$

是从分布为$t(f_e)$的总体中抽取容量为$r$的样本所得的最大与最小顺序统计量之差（极差），称之为$t$化极差统计量，其分布记为$q(r,f_e)$。显然，$t$化极差统计量$Q$的分布$q(r,f_e)$只与水平个数$r$以及$t$分布的自由度$f_e$有关，而与参数$\mu,\sigma^2$及重复数$m$无关。

分布$q(r,f_e)$的准确形式比较复杂，通常采用随机模拟方法得到其分位数$q_{1-\alpha}(r,f_e)$。对于给定的容量$r$及自由度$f_e$，随机模拟方法是

（1）随机生成$r$个标准正态分布$N(0,1)$随机数$x_1,x_2,\cdots,x_r$，将这$r$个随机数按由小到大的顺序排列，得到其最小随机数$x_{(1)}$和最大随机数$x_{(r)}$。

（2）随机生成 1 个自由度为 $f_e$ 的 $\chi^2$ 分布 $\chi^2(f_e)$ 随机数 $y$。

（3）计算 $q=\dfrac{x_{(r)}-x_{(1)}}{\sqrt{y/f_e}}$。

（4）重复（1）至（3）步 $N$ 次，得到 $t$ 化极差统计量 $Q$ 的 $N$ 个观测值，只要 $N$ 非常大，就可根据样本分位数得到 $q(r,f_e)$ 分布的各种分位数 $q_{1-\alpha}(r,f_e)$ 的近似值。

当显著水平为 $\alpha$ 时，拒绝域

$$W=\left\{q\geq\frac{c}{\hat{\sigma}/\sqrt{m}}\right\}=\{q\geq q_{1-\alpha}(r,f_e)\},$$

有 $q_{1-\alpha}(r,f_e)=\dfrac{c}{\hat{\sigma}/\sqrt{m}}$，可得

$$c=q_{1-\alpha}(r,f_e)\cdot\frac{\hat{\sigma}}{\sqrt{m}},$$

再逐个将 $|\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}|$ 与 $c$ 比较，得出每一对 $\mu_i$ 与 $\mu_j$ 是否有显著差异的结论。

步骤：

（1）假设 $H_0^{ij}:\mu_i=\mu_j$，$i,j=1,2,\cdots,r$。

（2）统计量 $Q=\dfrac{\max\limits_{1\leq i\leq r}\overline{Y}_{i\cdot}-\min\limits_{1\leq i\leq r}\overline{Y}_{i\cdot}}{\hat{\sigma}/\sqrt{m}}$。

（3）右侧拒绝域 $W=\left\{q\geq\dfrac{c}{\hat{\sigma}/\sqrt{m}}\right\}=\{q\geq q_{1-\alpha}(r,f_e)\}$。

（4）计算 $c=q_{1-\alpha}(r,f_e)\cdot\dfrac{\hat{\sigma}}{\sqrt{m}}$，逐个将 $|\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}|$ 与 $c$ 比较，作出决策。

**例** 由前面的鸡饲料对鸡增重影响问题的数据对各因子作多重比较（$\alpha=0.05$）。

**解：**单因子方差分析多重比较，假设 $H_0^{ij}:\mu_i=\mu_j$，$i,j=1,2,3$，统计量 $Q=\dfrac{\max\limits_{1\leq i\leq r}\overline{Y}_{i\cdot}-\min\limits_{1\leq i\leq r}\overline{Y}_{i\cdot}}{\hat{\sigma}/\sqrt{m}}$，显著水平 $\alpha=0.05$，$r=3$，$f_e=n-r=21$，$q_{1-\alpha}(r,f_e)=q_{0.95}(3,21)=3.57$，右侧拒绝域 $W=\{q\geq3.57\}$。又因

$$\hat{\sigma}=\sqrt{\frac{S_e}{n-r}}=\sqrt{\frac{28215.875}{21}}=36.6553,$$

$$c=q_{1-\alpha}(r,f_e)\cdot\frac{\hat{\sigma}}{\sqrt{m}}=3.57\times\frac{36.6553}{\sqrt{8}}=46.2658。$$

因

$$|\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}|=|1024.25-1073.125|=48.875>c,$$

$$|\overline{Y}_{1\cdot}-\overline{Y}_{3\cdot}|=|1024.25-1044.25|=20<c,$$

$$|\bar{Y}_{2\cdot}-\bar{Y}_{3\cdot}|=|1073.125-1044.25|=28.875<c\,,$$

故 $\mu_1$ 与 $\mu_2$ 有显著差异，$\mu_1$ 与 $\mu_3$、$\mu_2$ 与 $\mu_3$ 没有显著差异。

### 8.2.4 重复数不等场合的 $S$ 法

重复数不等时，因

$$T_{ij}=\frac{(\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot})-(\mu_i-\mu_j)}{\hat{\sigma}\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}}\sim t(n-r)=t(f_e)\,,$$

当所有假设 $H_0^{ij}$ 都成立时，即 $\mu_1=\mu_2=\cdots=\mu_r=\mu$，有

$$T_{ij}=\frac{\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}}{\hat{\sigma}\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}}\sim t(f_e)\,,\quad F_{ij}=T_{ij}^2=\frac{(\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot})^2}{\hat{\sigma}^2\left(\frac{1}{m_i}+\frac{1}{m_j}\right)}\sim F(1,f_e)\,。$$

因 $\dfrac{\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}}{\hat{\sigma}\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}}$ 的分布 $t(f_e)$ 与 $i,j$ 无关，从而统一的拒绝域可以取为

$$W=\bigcup_{1\le i<j\le r}\{|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|\ge c_{ij}\}=\bigcup_{1\le i<j\le r}\left\{|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|\ge c\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}\right\}=\bigcup_{1\le i<j\le r}\left\{\frac{|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|}{\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}}\ge c\right\}$$

$$=\left\{\max_{1\le i<j\le r}\frac{|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|}{\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}}\ge c\right\}=\left\{\max_{1\le i<j\le r}\frac{(\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot})^2}{\hat{\sigma}^2\left(\frac{1}{m_i}+\frac{1}{m_j}\right)}\ge\frac{c^2}{\hat{\sigma}^2}\right\}=\left\{\max_{1\le i<j\le r}F_{ij}\ge\frac{c^2}{\hat{\sigma}^2}\right\}\,,$$

可以证明，$F=\dfrac{\max\limits_{1\le i<j\le r}F_{ij}}{r-1}\dot{\sim}F(r-1,f_e)$。当显著水平为 $\alpha$ 时，拒绝域

$$W=\left\{f\ge\frac{c^2}{(r-1)\hat{\sigma}^2}\right\}=\{f\ge f_{1-\alpha}(r-1,f_e)\}\,,$$

有 $f_{1-\alpha}(r-1,f_e)=\dfrac{c^2}{(r-1)\hat{\sigma}^2}$，可得

$$c=\hat{\sigma}\sqrt{(r-1)f_{1-\alpha}(r-1,f_e)}\,,$$

因此

$$c_{ij}=c\sqrt{\frac{1}{m_i}+\frac{1}{m_j}}=\hat{\sigma}\sqrt{(r-1)f_{1-\alpha}(r-1,f_e)\left(\frac{1}{m_i}+\frac{1}{m_j}\right)}\,,$$

再逐个将 $|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|$ 与 $c_{ij}$ 比较，得出每一对 $\mu_i$ 与 $\mu_j$ 是否有显著差异的结论。

步骤：

（1）假设 $H_0^{ij}:\mu_i=\mu_j$， $i,j=1,2,\cdots,r$。

（2）统计量 $F=\dfrac{\max\limits_{1\le i<j\le r}F_{ij}}{r-1}\sim F(r-1,f_e)$。

（3）右侧拒绝域 $W=\left\{f\ge\dfrac{c^2}{(r-1)\hat{\sigma}^2}\right\}=\{f\ge f_{1-\alpha}(r-1,f_e)\}$。

（4）计算 $c_{ij}=c\sqrt{\dfrac{1}{m_i}+\dfrac{1}{m_j}}=\hat{\sigma}\sqrt{(r-1)f_{1-\alpha}(r-1,f_e)\left(\dfrac{1}{m_i}+\dfrac{1}{m_j}\right)}$，逐个将 $|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}|$ 与 $c$ 比较，作出决策。

**例** 由前面的食品包装对销售量影响问题的数据对各因子作多重比较（$\alpha=0.01$）。

**解：**单因子方差分析多重比较，假设 $H_0^{ij}:\mu_i=\mu_j$， $i,j=1,2,3,4$，统计量 $F=\dfrac{\max\limits_{1\le i<j\le r}F_{ij}}{r-1}\sim F(r-1,f_e)$，显著水平 $\alpha=0.01$， $r=4$， $f_e=n-r=6$， $f_{1-\alpha}(r-1,f_e)=f_{0.99}(3,6)=9.78$，右侧拒绝域 $W=\{f\ge 9.78\}$。

又因 $m_1=m_4=2$， $m_2=m_3=3$，且

$$\hat{\sigma}=\sqrt{\frac{S_e}{n-r}}=\sqrt{\frac{46}{6}}=2.7689,$$

则

$$c_{12}=c_{13}=c_{24}=c_{34}=2.7689\times\sqrt{3\times 9.78\times\left(\frac{1}{2}+\frac{1}{3}\right)}=13.6914,$$

$$c_{14}=2.7689\times\sqrt{3\times 9.78\times\left(\frac{1}{2}+\frac{1}{2}\right)}=14.9981,\quad c_{23}=2.7689\times\sqrt{3\times 9.78\times\left(\frac{1}{3}+\frac{1}{3}\right)}=12.2459,$$

因

$$|\bar{Y}_{1\cdot}-\bar{Y}_{2\cdot}|=|15-13|=2<c_{12},$$

$$|\bar{Y}_{1\cdot}-\bar{Y}_{3\cdot}|=|15-19|=4<c_{13},$$

$$|\bar{Y}_{1\cdot}-\bar{Y}_{4\cdot}|=|15-27|=12<c_{14},$$

$$|\bar{Y}_{2\cdot}-\bar{Y}_{3\cdot}|=|13-19|=6<c_{23},$$

$$|\bar{Y}_{2\cdot}-\bar{Y}_{4\cdot}|=|13-27|=14>c_{24},$$

$$|\bar{Y}_{3\cdot}-\bar{Y}_{4\cdot}|=|19-27|=8<c_{34},$$

故 $\mu_2$ 与 $\mu_4$ 有显著差异，其他两两之间没有显著差异。

## §8.3 方差齐性检验

在单因子方差分析统计模型中，总是假设各个水平下的总体方差都相等，即 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2$，称之为方差齐性。但方差齐性不一定自然成立，需要对其进行检验，检验的原假设为 $H_0:\sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2$，称为方差齐性检验。

各水平下总体方差 $\sigma_i^2$ 的点估计分别是对应水平下的样本方差 $S_i^2$，以由 $S_1^2,S_2^2,\cdots,S_r^2$ 构成的函数作为检验的统计量。

分成重复数相等与不等两种场合进行讨论。

### 8.3.1 重复数相等场合的 Hartley 检验法

重复数相等时，各水平下样本方差 $S_i^2$ 是平等的，以 $r$ 个水平下样本方差 $S_i^2$，$i=1,2,\cdots,r$ 的最大值与最小值之比作为检验的统计量 $H$，即

$$H=\frac{\max\{S_1^2,S_2^2,\cdots,S_r^2\}}{\min\{S_1^2,S_2^2,\cdots,S_r^2\}}。$$

统计量 $H$ 的分布只与水平个数 $r$ 及样本方差 $S_i^2$ 的自由度 $f=m-1$ 有关，记为 $H(r,f)$。分布 $H(r,f)$ 的准确形式比较复杂，通常采用随机模拟方法得到其分位数 $H_{1-\alpha}(r,f)$。

显然 $H\geq 1$，在方差齐性成立的条件下，各 $S_i^2$ 应相差不大，即统计量 $H$ 的观测值应接近于 1 才“合理”。可见当 $H$ 的观测值接近于 1 时，应接受 $H_0$；当 $H$ 的观测值远大于 1 时，应拒绝 $H_0$。取右侧拒绝域，形式为 $W=\{H\geq H_{1-\alpha}(r,f)\}$。

步骤：

（1）假设 $H_0:\sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2$。

（2）统计量 $H=\dfrac{\max\{S_1^2,S_2^2,\cdots,S_r^2\}}{\min\{S_1^2,S_2^2,\cdots,S_r^2\}}$。

（3）右侧拒绝域 $W=\{H\geq H_{1-\alpha}(r,f)\}$。

（4）计算检验统计量观测值 $H$ 与检验的 $p$ 值，并作出决策。

这称之为 Hartley 检验法。

各水平下样本方差的计算公式为

$$S_i^2=\frac{1}{m-1}\sum_{j=1}^{m}(Y_{ij}-\overline{Y}_{i\cdot})^2=\frac{1}{m-1}\left[\sum_{j=1}^{m}Y_{ij}^2-m\overline{Y}_{i\cdot}^2\right]=\frac{1}{m-1}\left[\sum_{j=1}^{m}Y_{ij}^2-\frac{T_i^2}{m}\right],\quad i=1,2,\cdots,r。$$

**例** 由前面的鸡饲料对鸡增重影响问题的数据采用 Hartley 检验法进行方差齐性检验（$\alpha=0.05$）。

**解：**方差齐性检验。假设 $H_0:\sigma_1^2=\sigma_2^2=\sigma_3^2$，统计量 $H=\dfrac{\max\{S_1^2,S_2^2,S_3^2\}}{\min\{S_1^2,S_2^2,S_3^2\}}$，显著水平 $\alpha=0.05$，$r=3$，

$f=m-1=7$，$H_{1-\alpha}(r,f)=H_{0.95}(3,7)=6.94$，右侧拒绝域$W=\{H\geq 6.94\}$。

根据试验数据计算表，可得

$$T_1=8194，T_2=8585，T_3=8354，\sum_{j=1}^{m}y_{1j}^2=8398024，\sum_{j=1}^{m}y_{2j}^2=9230355，\sum_{j=1}^{m}y_{3j}^2=8728984，$$

则

$$s_1^2=\frac{1}{7}\left(8398024-\frac{8194^2}{8}\right)=759.9286，$$

$$s_2^2=\frac{1}{7}\left(9230355-\frac{8585^2}{8}\right)=2510.9821，$$

$$s_3^2=\frac{1}{7}\left(8728984-\frac{8354^2}{8}\right)=759.9286，$$

可得

$$H=\frac{2510.9821}{759.9286}=3.3042\notin W，$$

故接受$H_0$，拒绝$H_1$，可以认为三个水平下的总体方差满足方差齐性。

### 8.3.2 重复数不等场合大样本情形的 Bartlett 检验法

重复数不等时，各水平下样本方差$S_i^2$不平等，比较时应考虑对应的重复数$m_i$或其自由度$m_i-1$。因一组正数的算术平均大于等于其几何平均，二者相等当且仅当这组正数全相等。

取各水平下样本方差$S_i^2$以其自由度$f_i=m_i-1$为权数的加权算术平均与加权几何平均之比。因

$$S_i^2=\frac{1}{m_i-1}\sum_{j=1}^{m_i}(Y_{ij}-\overline{Y}_{i\cdot})^2=\frac{1}{m_i-1}\left[\sum_{j=1}^{m_i}Y_{ij}^2-m_i\overline{Y}_{i\cdot}^2\right]=\frac{1}{m_i-1}\left[\sum_{j=1}^{m_i}Y_{ij}^2-\frac{T_i^2}{m_i}\right],\quad i=1,2,\cdots,r，$$

可得

$$S_e=\sum_{i=1}^{r}\sum_{j=1}^{m}(Y_{ij}-\overline{Y}_{i\cdot})^2=\sum_{i=1}^{r}(m_i-1)S_i^2=\sum_{i=1}^{r}f_iS_i^2，$$

$$f_e=n-r=\sum_{i=1}^{r}(m_i-1)=\sum_{i=1}^{r}f_i，$$

则组内偏差均方和

$$MS_e=\frac{S_e}{f_e}=\frac{1}{f_e}\sum_{i=1}^{r}f_iS_i^2=\sum_{i=1}^{r}\frac{f_i}{f_e}S_i^2，$$

即组内偏差均方和$MS_e$等于样本方差$S_1^2,S_2^2,\cdots,S_r^2$以其自由度$f_i=m_i-1$为权数的加权算术平均。再考虑对应的加权几何平均，记

$$GMS_e=\prod_{i=1}^{r}(S_i^2)^{\frac{f_i}{f_e}}，$$

可以证明，大样本场合，在方差齐性成立的条件下，$MS_e$ 与 $GMS_e$ 之商的函数

$$B=\frac{f_e}{C}\ln\frac{MS_e}{GMS_e}=\frac{1}{C}\left[f_e\ln(MS_e)-\sum_{i=1}^{r}f_i\ln(S_i^2)\right]\dot{\sim}\chi^2(r-1)\,,$$

其中常数

$$C=1+\frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i}-\frac{1}{f_e}\right)。$$

由于算术平均必然大于等于几何平均，有 $\frac{MS_e}{GMS_e}\geq 1$，即 $B\geq 0$，在方差齐性成立的条件下，$\frac{MS_e}{GMS_e}$ 应接近于 1，即 $B$ 接近于 0。因此拒绝域为右侧拒绝域 $W=\{B\geq\chi^2_{1-\alpha}(r-1)\}$。

步骤：

（1）假设 $H_0:\sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2$。

（2）统计量 $B=\frac{f_e}{C}\ln\frac{MS_e}{GMS_e}\dot{\sim}\chi^2(r-1)$，其中 $GMS_e=\prod_{i=1}^{r}(S_i^2)^{\frac{f_i}{f_e}}$，$C=1+\frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i}-\frac{1}{f_e}\right)$。

（3）右侧拒绝域 $W=\{B\geq\chi^2_{1-\alpha}(r-1)\}$。

（4）计算检验统计量观测值 $B$ 与检验的 $p$ 值，并作出决策。

这称之为 Bartlett 检验法。适用于每一个样本容量 $m_i$ 都不小于 5 的情形，重复数相等或不等均可采用。

**例** 由前面的鸡饲料对鸡增重影响问题的数据采用 Bartlett 检验法进行方差齐性检验（$\alpha=0.05$）。

**解：**方差齐性检验。假设 $H_0:\sigma_1^2=\sigma_2^2=\sigma_3^2$，统计量 $B=\frac{f_e}{C}\ln\frac{MS_e}{GMS_e}\dot{\sim}\chi^2(r-1)$，显著水平 $\alpha=0.05$，$r=3$，$\chi^2(r-1)=\chi^2(2)=5.9915$，右侧拒绝域 $W=\{B\geq 5.9915\}$。

根据试验数据计算表，可得

$s_1^2=759.9286$，$s_2^2=2510.9821$，$s_3^2=759.9286$，

$f_1=f_2=f_3=m-1=7$，$f_e=n-r=21$，$MS_e=1343.6131$，

$GMS_e=(s_1^2)^{\frac{f_1}{f_e}}(s_2^2)^{\frac{f_2}{f_e}}(s_3^2)^{\frac{f_3}{f_e}}=759.9268^{\frac{1}{3}}\times 2510.9821^{\frac{1}{3}}\times 759.9268^{\frac{1}{3}}=1131.8696$，

$C=1+\frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i}-\frac{1}{f_e}\right)=1+\frac{1}{3\times 2}\left(\frac{1}{7}+\frac{1}{7}+\frac{1}{7}-\frac{1}{21}\right)=1.0635$，

可得

$$B=\frac{f_e}{C}\ln\frac{MS_e}{GMS_e}=\frac{21}{1.0635}\ln\frac{1343.6131}{1131.8696}=3.8363\notin W\,,$$

故接受 $H_0$，拒绝 $H_1$，可以认为三个水平下的总体方差满足方差齐性。

### 8.3.3 重复数不等场合小样本情形的修正 Bartlett 检验法

Bartlett 检验法只能适用于每一个样本容量 $m_i$ 都不小于 5 的情形。当样本容量小于 5 时，Box 提出了修正 Bartlett 检验法。沿用 Bartlett 检验法的记号，修正的 Bartlett 检验统计量为

$$B' = \frac{r_2 BC}{r_1(A - BC)},$$

其中

$$r_1 = r - 1,\quad r_2 = \frac{r+1}{(C-1)^2},\quad A = \frac{r_2}{2 - C + \dfrac{2}{r_2}},$$

可以证明，在方差齐性成立的条件下，

$$B' = \frac{r_2 BC}{r_1(A - BC)} \dot{\sim} F(r_1, r_2),$$

显然 $B' \geq 0$，并且是 $B$ 的单调增加函数，在方差齐性成立的条件下，$B$ 接近于 0，$B'$ 也接近于 0。因此拒绝域为右侧拒绝域 $W = \{B' \geq f_{1-\alpha}(r_1, r_2)\}$。

步骤：

（1）假设 $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$。

（2）统计量 $B' = \dfrac{r_2 BC}{r_1(A - BC)} \dot{\sim} F(r_1, r_2)$。

（3）右侧拒绝域 $W = \{B' \geq f_{1-\alpha}(r_1, r_2)\}$。

（4）计算检验统计量观测值 $B'$ 与检验的 $p$ 值，并作出决策。

这称之为修正 Bartlett 检验法。不论重复数相等或不等，样本容量 $m_i$ 是大还是小都适用。

**例** 前面的食品包装对销售量影响问题的数据采用修正 Bartlett 检验法进行方差齐性检验（$\alpha = 0.01$）。

**解：**方差齐性检验。假设 $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$，统计量 $B' = \dfrac{r_2 BC}{r_1(A - BC)} \dot{\sim} F(r_1, r_2)$，显著水平 $\alpha = 0.01$，

$r = 4$，$f_1 = f_4 = m_1 - 1 = 1$，$f_2 = f_3 = m_2 - 1 = 2$，$f_e = n - r = 6$，有

$$C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{f_i} - \frac{1}{f_e}\right) = 1 + \frac{1}{3\times 3}\left(\frac{1}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{1} - \frac{1}{6}\right) = 1.3148,$$

$$r_1 = r - 1 = 3,\quad r_2 = \frac{r+1}{(C-1)^2} = \frac{5}{0.3148^2} = 50.4498,$$

则 $f_{1-\alpha}(r_1, r_2) = f_{0.99}(3, 50.4498) = 4.1954$，右侧拒绝域 $W = \{B' \geq 4.1954\}$。

根据试验数据计算表，可得

$$\frac{T_1^2}{m_1}=450,\ \frac{T_2^2}{m_2}=507,\ \frac{T_3^2}{m_3}=1083,\ \frac{T_4^2}{m_4}=1458,$$

$$\sum_{j=1}^{m_1}y_{1j}^2=468,\ \sum_{j=1}^{m_2}y_{2j}^2=509,\ \sum_{j=1}^{m_3}y_{3j}^2=1091,\ \sum_{j=1}^{m_4}y_{4j}^2=1476,$$

$$s_1^2=468-450=18,\ s_2^2=\frac{1}{2}(509-507)=1,\ s_3^2=\frac{1}{2}(1091-1083)=4,\ s_1^2=1476-1458=18,$$

$$MS_e=7.6667,\ GMS_e=(s_1^2)^{\frac{f_1}{f_e}}(s_2^2)^{\frac{f_2}{f_e}}(s_3^2)^{\frac{f_3}{f_e}}(s_4^2)^{\frac{f_4}{f_e}}=18^{\frac{1}{6}}\times1^{\frac{2}{6}}\times4^{\frac{2}{6}}\times18^{\frac{1}{6}}=4.1602,$$

$$B=\frac{f_e}{C}\ln\frac{MS_e}{GMS_e}=\frac{6}{1.3148}\ln\frac{7.6667}{4.1602}=2.7897,$$

$$A=\frac{r_2}{2-C+\frac{2}{r_2}}=\frac{50.4498}{2-1.3148+\frac{2}{50.4498}}=69.6024,$$

可得

$$B'=\frac{r_2BC}{r_1(A-BC)}=\frac{50.4498\times2.7897\times1.3148}{3\times(69.6024-2.7897\times1.3148)}=0.9355\notin W,$$

故接受$H_0$，拒绝$H_1$，可以认为四个水平下的总体方差满足方差齐性。

## §8.4　一元线性回归

### 8.4.1　变量之间的关系

实际工作中，通常需要考虑两个（随机）变量之间的关系，如圆的半径与面积的关系，人的身高与体重的关系，一个国家的 GDP 与年份的关系等等。

通常变量之间的关系分成两类：确定性关系与相关关系。确定性关系是指给定其中一个变量的值，就能确定另一个变量的值，如圆的半径与面积的关系，通常可以用函数表示。相关关系是指两个变量的取值有一些的联系，但不能由一个变量完全确定另一个变量，如人的身高与体重的关系。

对于具有相关关系的两个变量一般不能给出二者确切的函数关系，但可以在平均意义下给出二者的近似关系。如人的身高与体重之间没有确切的函数关系，但在平均意义下，近似有

体重（kg）= 身高（cm）– 105，或体重（kg）= 24 × 身高（m）$^2$。

回归分析就是分析相关关系的两个变量在平均意义下的函数关系表达式——回归函数。

对于具有相关关系的两个变量，类似于函数关系，也是以其中一个为自变量，另一个为因变量。因变量是随机变量，而自变量可以是普通变量，也可以是随机变量。但不论自变量是普通变量还是随机变量，进行回归分析时，总是将其看作可控的，称为可控变量，一般不再看作随机变量。

设可控变量 $x$ 与随机变量 $Y$ 具有相关关系，以 $x$ 为自变量，$Y$ 为因变量。当给定变量 $x$ 的值时，不能确定变量 $Y$ 的值，$Y$ 是一个与 $x$ 取值有关的随机变量。用 $Y$ 在 $x$ 每一取值下的数学期望作为理论回归函数

$$f(x)=E(Y;x)=\int_{-\infty}^{+\infty} yp(y;x)dy\,,$$

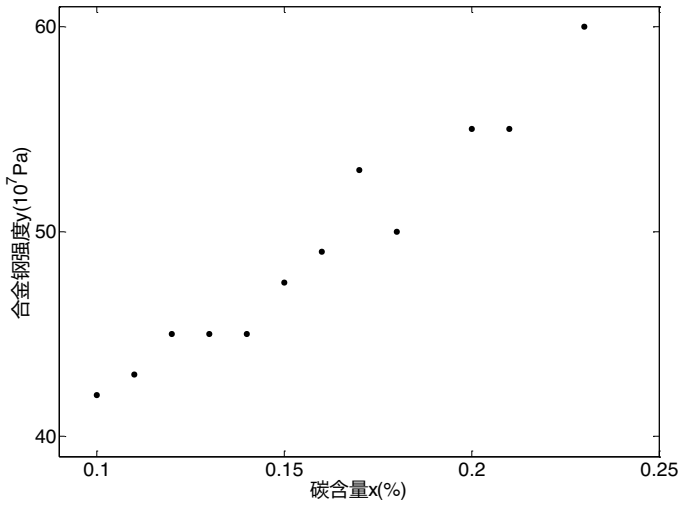
且因变量 $Y=f(x)+\varepsilon$，其中 $\varepsilon$ 是随机误差。精确的理论回归函数 $f(x)$ 一般很难得到，通常是首先根据观测数据作出散点图，再选择一个合适的回归函数形式，进一步估计其中的某些参数。

**例**　合金的强度 $Y$（$\times 10^7$ Pa）与合金中碳的含量 $x$（%）有关。为了掌握这两个变量的关系，收集了 12 对数据 $(x_i, y_i),\ \ i=1,2,\cdots,12$。作出散点图，并选择一个合适的回归函数形式。

**合金钢强度 $y$ 与碳含量 $x$ 的数据**

| 序号 | $x$/% | $y/10^7 Pa$ | 序号 | $x$/% | $y/10^7 Pa$ | 序号 | $x$/% | $y/10^7 Pa$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.10 | 42.0 | 5 | 0.14 | 45.0 | 9 | 0.18 | 50.0 |
| 2 | 0.11 | 43.0 | 6 | 0.15 | 47.5 | 10 | 0.20 | 55.0 |
| 3 | 0.12 | 45.0 | 7 | 0.16 | 49.0 | 11 | 0.21 | 55.0 |
| 4 | 0.13 | 45.0 | 8 | 0.17 | 53.0 | 12 | 0.23 | 60.0 |

**解：**根据观测数据作 $(x, y)$ 的散点图



可以看出这些点近似位于一条直线上，回归函数取为线性函数

$$Y=\beta_0+\beta_1 x+\varepsilon,$$

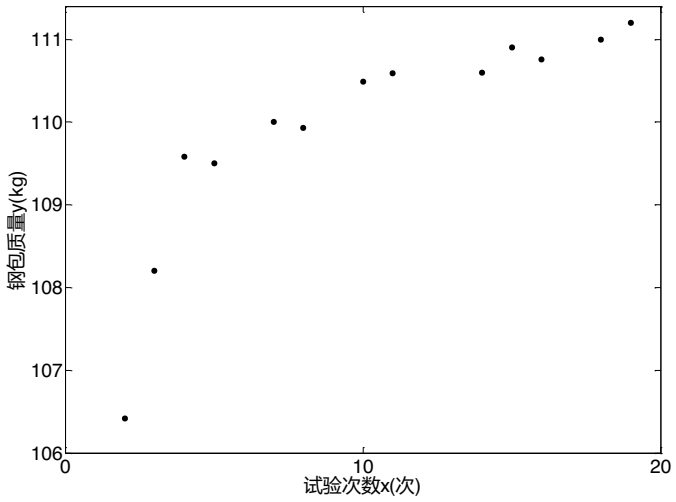
其中 $\beta_0,\beta_1$ 为未知参数，$\varepsilon$ 为随机误差。

**例** 炼钢厂出钢水时用的钢包，在使用过程中由于钢水的侵蚀，其容积不断增大。钢包容积用盛满钢水时的质量 $Y$（kg）表示，相应的试验次数用 $x$ 表示。根据表中的数据，作出散点图，并选择一个合适的回归函数形式。

**钢包的质量 $y$ 与试验次数 $x$ 的数据**

| 序号 | $x$（次） | $y$（kg） | 序号 | $x$（次） | $y$（kg） |
|---|---|---|---|---|---|
| 1 | 2 | 106.42 | 8 | 11 | 110.59 |
| 2 | 3 | 108.20 | 9 | 14 | 110.60 |
| 3 | 4 | 109.58 | 10 | 15 | 110.90 |
| 4 | 5 | 109.50 | 11 | 16 | 110.76 |
| 5 | 7 | 110.00 | 12 | 18 | 111.00 |
| 6 | 8 | 109.93 | 13 | 19 | 110.20 |
| 7 | 10 | 110.49 | | | |

**解：**根据观测数据作 $(x,y)$ 的散点图




可以看出这些点并不是位于一条直线上，根据散点图，回归函数可以取为非线性函数，如

$$Y=a+\frac{b}{x}+\varepsilon,$$

其中 $a,b$ 为未知参数，$\varepsilon$ 为随机误差。

如果回归函数 $f(x)$ 是一个线性函数，就称为线性回归，否则称为非线性回归。这一节讨论线性回归问题，下一节在讨论一些特殊的非线性回归问题。

8.4.2 一元线性回归模型

如果根据观测数据所作的散点图中，各点近似位于一条直线上，回归函数取为线性函数

$$Y=\beta_0+\beta_1 x+\varepsilon,$$

其中 $\beta_0,\beta_1$ 为未知参数，$\varepsilon$ 为随机误差。假定 $\varepsilon$ 服从均值为 0，方差为 $\sigma^2$ 的正态分布，即

$$\varepsilon \sim N(0,\sigma^2)\,,$$

可得

$$Y \sim N(\beta_0+\beta_1 x,\sigma^2)\,。$$

对于每一组数据 $(x_i,Y_i)$，有

$$Y_i=\beta_0+\beta_1 x_i+\varepsilon_i \sim N(\beta_0+\beta_1 x_i,\sigma^2)\,。$$

进一步假定收集数据时，每一次观测都是独立进行的，且误差方差 $\sigma^2$ 与 $x$ 无关。这样，各个 $\varepsilon_i$ 相互独立，且服从相同的正态分布 $N(0,\sigma^2)$。

一元线性回归模型

$$\begin{cases} Y_i=\beta_0+\beta_1 x_i+\varepsilon_i, \quad i=1,2,\cdots,n; \\ 各\ \varepsilon_i\ 相互独立，且服从相同的正态分布 N(0,\sigma^2). \end{cases}$$

根据观测数据 $(x_i,y_i)$，对参数 $\beta_0,\beta_1$ 作出估计，得到 $\hat{\beta}_0,\hat{\beta}_1$，取

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 x\,,$$

称为 $Y$ 关于 $x$ 的经验回归函数，也称为回归方程。若给定 $x$ 的值 $x_0$，可得 $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1 x_0$，称为随机变量 $Y$ 在 $x_0$ 处的回归值或预测值。

### 8.4.3 回归系数的最小二乘估计

一般采用最小二乘法估计回归参数 $\beta_0,\beta_1$。方法是选取 $\beta_0,\beta_1$ 的值，使得总的误差平方和达到最小，所得 $\beta_0,\beta_1$ 的值作为其估计值 $\hat{\beta}_0,\hat{\beta}_1$，称为最小二乘估计（Least Squares Estimation）。

对于 $n$ 对观测数据 $(x_i,y_i),\quad i=1,2,\cdots,n$，总的误差平方和

$$Q=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\,,$$

选取 $\beta_0,\beta_1$ 的值，使得 $Q$ 达到最小。令 $Q$ 关于 $\beta_0,\beta_1$ 的偏导数等于 0，得

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0}=\sum\limits_{i=1}^{n}2(y_i-\beta_0-\beta_1 x_i)\cdot(-1)=0; \\ \dfrac{\partial Q}{\partial \beta_0}=\sum\limits_{i=1}^{n}2(y_i-\beta_0-\beta_1 x_i)\cdot(-x_i)=0. \end{cases}$$

称为正规方程组，经过整理，可得

$$\begin{cases} n\beta_0+\beta_1\sum x_i=\sum y_i; \\ \beta_0\sum x_i+\beta_1\sum x_i^2=\sum x_i y_i. \end{cases}$$

这里为了简便，在作回归分析时，一般将求和号“$\sum_{i=1}^{n}$”简记为“$\sum$”，并记$\bar{x}=\frac{1}{n}\sum x_i$，$\bar{y}=\frac{1}{n}\sum y_i$，有

$$\begin{cases}\beta_0+\beta_1\bar{x}=\bar{y};\\ n\beta_0\bar{x}+\beta_1\sum x_i^2=\sum x_iy_i.\end{cases}$$

记$l_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-n\bar{x}^2$，$l_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_iy_i-n\bar{x}\bar{y}$，$l_{yy}=\sum(y_i-\bar{y})^2=\sum y_i^2-n\bar{y}^2$，求解正规方程组，可得

$$\begin{cases}\beta_1=\dfrac{\sum x_iy_i-n\bar{x}\bar{y}}{\sum x_i^2-n\bar{x}^2}=\dfrac{l_{xy}}{l_{xx}};\\ \beta_0=\bar{y}-\beta_1\bar{x}.\end{cases}$$

故取$\beta_0,\beta_1$的最小二乘估计为

$$\begin{cases}\hat{\beta}_1=\dfrac{l_{xY}}{l_{xx}};\\ \hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{x}.\end{cases}$$

**例** 根据前例中合金钢强度和碳含量数据，求回归方程。

**解：**根据试验数据得出计算表：

试验数据计算表

| | | |
|---|---|---|
| $\sum x_i=1.9$ | $n=12$ | $\sum y_i=589.5$ |
| $\bar{x}=0.1583$ | | $\bar{y}=49.125$ |
| $\sum x_i^2=0.3194$ | $\sum x_iy_i=95.805$ | $\sum y_i^2=29304.25$ |
| $n\bar{x}^2=0.3008$ | $n\bar{x}\bar{y}=93.3375$ | $n\bar{y}^2=28959.1875$ |
| $l_{xx}=0.018567$ | $l_{xy}=2.4675$ | $l_{yy}=345.0625$ |
| | $\hat{\beta}_1=l_{xy}/l_{xx}=132.8995$<br>$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=28.0826$ | |

故回归方程为

$\hat{Y}=28.0826+132.8995x$。

**定理** 线性回归模型中参数$\beta_0,\beta_1$和随机变量$Y$的最小二乘估计$\hat{\beta}_0,\hat{\beta}_1$和$\hat{Y}$的分布为

（1）$\hat{\beta}_0\sim N\left(\beta_0,\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right)$，$\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{l_{xx}}\right)$。

（2）$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)=-\frac{\bar{x}}{l_{xx}}\sigma^2$。

（3）对给定的$x_0$，$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1x_0\sim N\left(\beta_0+\beta_1x_0,\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\sigma^2\right)$。

**证明：**（1）因

$\sum(x_i-\bar{x})=\sum x_i-n\bar{x}=0$，

有

$$l_{xx}=\sum(x_i-\bar{x})^2=\sum(x_i-\bar{x})x_i-\bar{x}\sum(x_i-\bar{x})=\sum(x_i-\bar{x})x_i\,,$$

$$l_{xY}=\sum(x_i-\bar{x})(Y_i-\bar{Y})=\sum(x_i-\bar{x})Y_i-\bar{Y}\sum(x_i-\bar{x})=\sum(x_i-\bar{x})Y_i\,,$$

则

$$\hat{\beta}_1=\frac{l_{xY}}{l_{xx}}=\frac{\sum(x_i-\bar{x})Y_i}{l_{xx}}=\sum\frac{x_i-\bar{x}}{l_{xx}}Y_i\,,$$

$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{x}=\frac{1}{n}\sum Y_i-\sum\frac{x_i-\bar{x}}{l_{xx}}Y_i\cdot\bar{x}=\sum\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right]Y_i\,。$$

可见 $\hat{\beta}_0,\hat{\beta}_1$ 都是独立正态变量 $Y_1,Y_2,\cdots,Y_n$ 的线性组合，即 $\hat{\beta}_0,\hat{\beta}_1$ 都服从正态分布。

因

$$Y_i=\beta_0+\beta_1x_i+\varepsilon_i\sim N(\beta_0+\beta_1x_i,\sigma^2)\,,$$

有 $E(Y_i)=\beta_0+\beta_1x_i$，$\mathrm{Var}(Y_i)=\sigma^2$，则

$$E(\hat{\beta}_1)=\sum\frac{x_i-\bar{x}}{l_{xx}}E(Y_i)=\sum\frac{x_i-\bar{x}}{l_{xx}}(\beta_0+\beta_1x_i)=\frac{\beta_0}{l_{xx}}\sum(x_i-\bar{x})+\frac{\beta_1}{l_{xx}}\sum(x_i-\bar{x})x_i=\frac{\beta_1}{l_{xx}}\cdot l_{xx}=\beta_1\,,$$

$$\mathrm{Var}(\hat{\beta}_1)=\sum\left(\frac{x_i-\bar{x}}{l_{xx}}\right)^2\mathrm{Var}(Y_i)=\sum\frac{(x_i-\bar{x})^2}{l_{xx}^2}\sigma^2=\frac{\sigma^2}{l_{xx}^2}\sum(x_i-\bar{x})^2=\frac{\sigma^2}{l_{xx}^2}\cdot l_{xx}=\frac{\sigma^2}{l_{xx}}\,,$$

$$E(\hat{\beta}_0)=E(\bar{Y}-\hat{\beta}_1\bar{x})=\frac{1}{n}\sum E(Y_i)-E(\hat{\beta}_1)\bar{x}=\frac{1}{n}\sum(\beta_0+\beta_1x_i)-\beta_1\bar{x}=\beta_0+\beta_1\bar{x}-\beta_1\bar{x}=\beta_0\,,$$

$$\mathrm{Var}(\hat{\beta}_0)=\sum\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right]^2\mathrm{Var}(Y_i)=\sum\left[\frac{1}{n^2}+\frac{(x_i-\bar{x})^2\bar{x}^2}{l_{xx}^2}-\frac{2(x_i-\bar{x})\bar{x}}{nl_{xx}}\right]\sigma^2$$

$$=n\cdot\frac{\sigma^2}{n^2}+\frac{\bar{x}^2\sigma^2}{l_{xx}^2}\sum(x_i-\bar{x})^2-\frac{2\bar{x}\sigma^2}{nl_{xx}}\sum(x_i-\bar{x})=\frac{\sigma^2}{n}+\frac{\bar{x}^2\sigma^2}{l_{xx}}-0=\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\,,$$

故

$$\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{l_{xx}}\right),\quad\hat{\beta}_0\sim N\left(\beta_0,\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right)。$$

（2）因 $Y_1,Y_2,\cdots,Y_n$ 相互独立，当 $i\neq j$ 时，有 $\mathrm{Cov}(Y_i,Y_j)=0$，故

$$\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1)=\mathrm{Cov}\left(\sum\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right]Y_i,\sum\frac{x_i-\bar{x}}{l_{xx}}Y_i\right)=\sum\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right]\frac{x_i-\bar{x}}{l_{xx}}\mathrm{Cov}(Y_i,Y_i)$$

$$=\sum\left[\frac{x_i-\bar{x}}{nl_{xx}}-\frac{(x_i-\bar{x})^2\bar{x}}{l_{xx}^2}\right]\sigma^2=\frac{\sigma^2}{nl_{xx}}\sum(x_i-\bar{x})-\frac{\bar{x}\sigma^2}{l_{xx}^2}\sum(x_i-\bar{x})^2=-\frac{\bar{x}}{l_{xx}}\sigma^2\,。$$

（3）对给定的 $x_0$，有

$$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1x_0=\sum\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{l_{xx}}\right]Y_i+\sum\frac{x_i-\bar{x}}{l_{xx}}Y_i\cdot x_0=\sum\left[\frac{1}{n}+\frac{(x_i-\bar{x})(x_0-\bar{x})}{l_{xx}}\right]Y_i\,,$$

可见 $\hat{Y}_0$ 也是独立正态变量 $Y_1,Y_2,\cdots,Y_n$ 的线性组合，即 $\hat{Y}_0$ 服从正态分布。

因

$$E(\hat{Y}_0)=E(\hat{\beta}_0)+E(\hat{\beta}_1)x_0=\beta_0+\beta_1x_0\,,$$

$$\begin{aligned}\mathrm{Var}(\hat{Y}_0)&=\sum\left[\frac{1}{n}+\frac{(x_i-\bar{x})(x_0-\bar{x})}{l_{xx}}\right]^2\mathrm{Var}(Y_i)\\&=\sum\left[\frac{1}{n^2}+\frac{(x_i-\bar{x})^2(x_0-\bar{x})^2}{l_{xx}^2}+\frac{2(x_i-\bar{x})(x_0-\bar{x})}{nl_{xx}}\right]\sigma^2\\&=n\cdot\frac{\sigma^2}{n^2}+\frac{(x_0-\bar{x})^2\sigma^2}{l_{xx}^2}\cdot\sum(x_i-\bar{x})^2+\frac{2(x_0-\bar{x})\sigma^2}{nl_{xx}}\sum(x_i-\bar{x})\\&=\frac{\sigma^2}{n}+\frac{(x_0-\bar{x})^2\sigma^2}{l_{xx}^2}\cdot l_{xx}+0=\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\sigma^2\,,\end{aligned}$$

故

$$\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1x_0\sim N\left(\beta_0+\beta_1x_0,\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\sigma^2\right)。$$

8.4.4　回归方程的显著性检验

由前面回归系数的最小二乘法可见，对于任意 $n$ 对数据 $(x_i,y_i),\ \ i=1,2,\cdots,n$，都可得到一个回归方程 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1x$，但实际上这两个变量并不一定具有相关关系，得到的回归方程不一定有实际意义，因此需要对回归方程进行显著性检验。

线性回归方程的目的是寻找变量 $Y$ 随变量 $x$ 的线性变化的规律。对于线性回归问题 $Y=\beta_0+\beta_1x+\varepsilon$，如果 $\beta_1=0$，则变量 $Y$ 的变化只是由随机误差 $\varepsilon$ 造成，与变量 $x$ 的变化无关，表明 $Y$ 与 $x$ 没有相关关系；反之，如果 $\beta_1\neq0$，则变量 $Y$ 的变化与变量 $x$ 的变化有关，表明 $Y$ 与 $x$ 具有相关关系。因此可通过检验假设 $H_0:\beta_1=0\ \ vs\ \ H_1:\beta_1\neq0$，判断 $Y$ 与 $x$ 是否具有相关关系。如果接受 $H_0$，则回归方程不显著；如果拒绝 $H_0$，则回归方程显著。

有三种等价的检验方法：$F$ 检验，$t$ 检验，$r$ 检验。检验时，可采用任一检验方法。

一．$F$ 检验

采用方差分析的思想进行检验。设数据为 $(x_i,y_i),\ \ i=1,2,\cdots,n$，线性回归模型 $Y=\beta_0+\beta_1x+\varepsilon$，回归系数 $\beta_0,\beta_1$ 的最小二乘估计为 $\hat{\beta}_0,\hat{\beta}_1$，且 $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$ 为 $Y_i$ 的回归值。

记 $\bar{Y}=\frac{1}{n}\sum Y_i$，称 $Y_i-\bar{Y}$ 为偏差，$Y_i-\hat{Y}_i$ 为残差，$\hat{Y}_i-\bar{Y}$ 为回归差，显然有

$$Y_i-\bar{Y}=(Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y})。$$

称 $S_T=\sum(Y_i-\bar{Y})^2=l_{YY}$ 为总偏差平方和，$S_e=\sum(Y_i-\hat{Y}_i)^2$ 为残差平方和，$S_R=\sum(\hat{Y}_i-\bar{Y})^2$ 为回归平方和。

**结论** （平方和分解）总偏差平方和 $S_T$ 可以分解为残差平方和 $S_e$ 与回归平方和 $S_R$ 之和，即

$$S_T=S_e+S_R。$$

**证明：**因 $\hat{\beta}_0,\hat{\beta}_1$ 是正规方程

$$\begin{cases}\sum(Y_i-\beta_0-\beta_1x_i)=0;\\ \sum(Y_i-\beta_0-\beta_1x_i)\cdot x_i=0.\end{cases}$$

的解，可知

$$\begin{cases}\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=\sum(Y_i-\hat{Y}_i)=0;\\ \sum(Y_i-\beta_0-\beta_1x_i)\cdot x_i=\sum(Y_i-\hat{Y}_i)\cdot x_i=0.\end{cases}$$

故

$$S_T=\sum(Y_i-\bar{Y})^2=\sum[(Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y})]^2=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\bar{Y})^2+2\sum(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})$$

$$=S_e+S_R+2\sum(Y_i-\hat{Y}_i)(\hat{\beta}_0+\hat{\beta}_1x_i-\bar{Y})$$

$$=S_e+S_R+2(\hat{\beta}_0-\bar{Y})\sum(Y_i-\hat{Y}_i)+2\hat{\beta}_1\sum(Y_i-\hat{Y}_i)\cdot x_i=S_e+S_R。$$

**定理** 线性回归模型中，残差平方和 $S_e$ 与回归平方和 $S_R$ 满足

（1）$E(S_R)=\sigma^2+\beta_1^2l_{xx}$，且在 $H_0:\beta_1=0$ 成立条件下，$\frac{S_R}{\sigma^2}\sim\chi^2(1)$。

（2）$E(S_e)=(n-2)\sigma^2$，且 $\frac{S_e}{\sigma^2}\sim\chi^2(n-2)$。

（3）$S_e$、$S_R$、$\bar{Y}$ 相互独立。

**证明：**（1）因 $\bar{Y}=\hat{\beta}_0+\hat{\beta}_1\bar{x}$，则

$$S_R=\sum(\hat{Y}_i-\bar{Y})^2=\sum(\hat{\beta}_0+\hat{\beta}_1x_i-\hat{\beta}_0-\hat{\beta}_1\bar{x})^2=\hat{\beta}_1^2\sum(x_i-\bar{x})^2=\hat{\beta}_1^2l_{xx}。$$

又因 $\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{l_{xx}}\right)$，可得

$$E(S_R)=E(\hat{\beta}_1^2)l_{xx}=\{\mathrm{Var}(\hat{\beta}_1)+[E(\hat{\beta}_1)]^2\}l_{xx}=\left(\frac{\sigma^2}{l_{xx}}+\beta_1^2\right)l_{xx}=\sigma^2+\beta_1^2l_{xx}。$$

且在 $H_0:\beta_1=0$ 成立条件下，有 $\hat{\beta}_1\sim N\left(0,\frac{\sigma^2}{l_{xx}}\right)$，即 $\frac{\hat{\beta}_1\sqrt{l_{xx}}}{\sigma}\sim N(0,1)$，故

$$\frac{S_R}{\sigma^2}=\frac{\hat{\beta}_1^2 l_{xx}}{\sigma^2}=\left(\frac{\hat{\beta}_1\sqrt{l_{xx}}}{\sigma}\right)^2\sim\chi^2(1)。$$

（2）因

$$S_T=\sum(Y_i-\bar{Y})^2=\sum Y_i^2-n\bar{Y}^2,$$

则

$$\sum Y_i^2=n\bar{Y}^2+S_T=n\bar{Y}^2+S_R+S_e=n\bar{Y}^2+\hat{\beta}_1^2 l_{xx}+S_e。$$

因 $\bar{Y}=\frac{1}{n}\sum Y_i$，$\hat{\beta}_1=\frac{l_{xY}}{l_{xx}}=\sum\frac{x_i-\bar{x}}{l_{xx}}Y_i$，有

$$n\bar{Y}^2=\left(\sum\frac{1}{\sqrt{n}}Y_i\right)^2,\quad \hat{\beta}_1^2 l_{xx}=\left(\sum\frac{x_i-\bar{x}}{\sqrt{l_{xx}}}Y_i\right)^2。$$

令

$$\alpha_1=\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}},\cdots,\frac{1}{\sqrt{n}}\right)^T,\quad \alpha_2=\left(\frac{x_1-\bar{x}}{\sqrt{l_{xx}}},\frac{x_2-\bar{x}}{\sqrt{l_{xx}}},\cdots,\frac{x_n-\bar{x}}{\sqrt{l_{xx}}}\right)^T,$$

有

$$\|\alpha_1\|^2=n\cdot\frac{1}{n}=1,\quad \|\alpha_2\|^2=\frac{1}{l_{xx}}\sum(x_i-\bar{x})^2=1,\quad \alpha_1^T\alpha_2=\frac{1}{\sqrt{nl_{xx}}}\sum(x_i-\bar{x})=0,$$

可知 $\alpha_1,\alpha_2$ 是两个相互正交的单位向量，可将 $\alpha_1,\alpha_2$ 扩充为 $R^n$ 中的一组标准正交基 $\alpha_1,\alpha_2,\alpha_3,\cdots,\alpha_n$，令正交阵 $C=(\alpha_1,\alpha_2,\cdots,\alpha_n)$，正交变换 $\vec{Y}=C\vec{Z}$，即

$$\vec{Z}=(Z_1,Z_2,\cdots,Z_n)^T=C^T\vec{Y}=C^T(Y_1,Y_2,\cdots,Y_n)^T。$$

因 $Y_1,Y_2,\cdots,Y_n$ 相互独立且都服从方差同为 $\sigma^2$ 的正态分布，由§5.4 节引理可知 $Z_1,Z_2,\cdots,Z_n$ 相互独立且都服从方差同为 $\sigma^2$ 的正态分布。

因

$$\sum_{i=1}^n Y_i^2=\vec{Y}^T\vec{Y}=(C\vec{Z})^T C\vec{Z}=\vec{Z}^T C^T C\vec{Z}=\vec{Z}^T\vec{Z}=\sum_{i=1}^n Z_i^2,$$

$$\vec{Z}=\begin{pmatrix}Z_1\\Z_2\\\vdots\\Z_n\end{pmatrix}=C^T\vec{Y}=\begin{pmatrix}\alpha_1^T\\\alpha_2^T\\\vdots\\\alpha_n^T\end{pmatrix}\vec{Y}=\begin{pmatrix}\alpha_1^T\vec{Y}\\\alpha_2^T\vec{Y}\\\vdots\\\alpha_n^T\vec{Y}\end{pmatrix},$$

可知 $Z_i=\alpha_i^T\vec{Y}$，特别是

$$Z_1=\alpha_1^T\vec{Y}=\sum\frac{1}{\sqrt{n}}Y_i=\sqrt{n}\cdot\frac{1}{n}\sum Y_i=\sqrt{n}\bar{Y},$$

$$Z_2=\alpha_2^T\vec{Y}=\sum\frac{x_i-\bar{x}}{\sqrt{l_{xx}}}Y_i=\sqrt{l_{xx}}\cdot\sum\frac{x_i-\bar{x}}{l_{xx}}Y_i=\sqrt{l_{xx}}\hat{\beta}_1,$$

则

$$S_e=\sum Y_i^2-n\bar{Y}^2-\hat{\beta}_1^2 l_{xx}=\sum Z_i^2-Z_1^2-Z_2^2=\sum_{i=3}^{n}Z_i^2\ 。$$

因

$$E(\vec{Y})=\begin{pmatrix}E(Y_1)\\E(Y_2)\\\vdots\\E(Y_n)\end{pmatrix}=\begin{pmatrix}\beta_0+\beta_1x_1\\\beta_0+\beta_1x_2\\\vdots\\\beta_0+\beta_1x_n\end{pmatrix}=\begin{pmatrix}\beta_0+\beta_1\bar{x}+\beta_1(x_1-\bar{x})\\\beta_0+\beta_1\bar{x}+\beta_1(x_2-\bar{x})\\\vdots\\\beta_0+\beta_1\bar{x}+\beta_1(x_n-\bar{x})\end{pmatrix}=(\beta_0+\beta_1\bar{x})\sqrt{n}\alpha_1+\beta_1\sqrt{l_{xx}}\alpha_2\ ,$$

则当$i\geq 3$时，有

$$E(Z_i)=\alpha_i^T E(\vec{Y})=\alpha_i^T[(\beta_0+\beta_1\bar{x})\sqrt{n}\alpha_1+\beta_1\sqrt{l_{xx}}\alpha_2]=(\beta_0+\beta_1\bar{x})\sqrt{n}\alpha_i^T\alpha_1+\beta_1\sqrt{l_{xx}}\alpha_i^T\alpha_2=0\ ,$$

可知$Z_3,\cdots,Z_n$相互独立且都服从正态分布$N(0,\sigma^2)$，故

$$\frac{S_e}{\sigma^2}=\sum_{i=3}^{n}\left(\frac{Z_i}{\sigma}\right)^2\sim\chi^2(n-2)\ 。$$

且$E\left(\frac{S_e}{\sigma^2}\right)=n-2$，即$E(S_e)=(n-2)\sigma^2$。

（3）因$S_e=\sum_{i=3}^{n}Z_i^2$，$S_R=Z_2^2$，$\bar{Y}=\frac{1}{\sqrt{n}}Z_1$，且$Z_1,Z_2,\cdots,Z_n$相互独立，故$S_e$、$S_R$、$\bar{Y}$相互独立。

**注：**（1）由于$\frac{S_e}{\sigma^2}\sim\chi^2(n-2)$，在$H_0:\beta_1=0$成立条件下，$\frac{S_R}{\sigma^2}\sim\chi^2(1)$，且$S_e$与$S_R$相互独立，则根据$F$分布的定义可知：在$H_0:\beta_1=0$成立条件下，有

$$F=\frac{\frac{S_R}{\sigma^2}}{\frac{S_e}{\sigma^2}\Big/(n-2)}=\frac{S_R}{S_e/(n-2)}\sim F(1,n-2)\ 。$$

因$E(S_R)=\sigma^2+\beta_1^2 l_{xx}$，当$H_0$不成立时，$\beta_1\neq 0$，$F$统计量的分子$S_R$很可能变大，从而检验统计量观测值$f$很可能变大，此检验的拒绝域在右侧，$W=\{f\geq f_{1-\alpha}(1,n-2)\}$。

（2）因$E(S_e)=(n-2)\sigma^2$，即$E\left(\frac{S_e}{n-2}\right)=\sigma^2$，故

$$\hat{\sigma^2}=\frac{S_e}{n-2}$$

是误差方差$\sigma^2$的无偏估计。

步骤：

（1）假设$H_0:\beta_1=0\ \ vs\ \ H_1:\beta_1\neq 0$。

（2）检验统计量$F=\frac{S_R}{S_e/(n-2)}\sim F(1,n-2)$。

（3）拒绝域：$W=\{f\geq f_{1-\alpha}(1,n-2)\}$。

（4）计算检验统计量观测值 $f$ 与检验的 $p$ 值，并作出决策。

这是 $F$ 检验法。

将检验统计量观测值 $F$ 的计算过程列成方差分析表：

| 来源 | 平方和 | 自由度 | 均方和 | $F$ 比 | $p$ 值 |
|---|---|---|---|---|---|
| 回归 | $S_R$ | $f_R=1$ | $MS_R=S_R$ | $f=MS_R/MS_e$ | $p=P\{F\geq f\}$ |
| 残差 | $S_e$ | $f_e=n-2$ | $MS_e=S_e/f_e$ | | |
| 总和 | $S_T$ | $f_T=n-1$ | | | |

计算公式：

$$S_R=\hat{\beta}_1^2 l_{xx}\text{，}\quad S_T=l_{yy}\text{，}\quad S_e=S_T-S_R=l_{yy}-\hat{\beta}_1^2 l_{xx}\text{。}$$

二．$T$ 检验

因 $\hat{\beta}_1\sim N\left(\beta_1,\dfrac{\sigma^2}{l_{xx}}\right)$，$\dfrac{S_e}{\sigma^2}\sim\chi^2(n-2)$，且 $\hat{\beta}_1=\sqrt{\dfrac{S_R}{l_{xx}}}$ 与 $S_e$ 相互独立，有 $\dfrac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{l_{xx}}}\sim N(0,1)$，则根据 $t$ 分布的定义可知：

$$T=\frac{\dfrac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{l_{xx}}}}{\sqrt{\dfrac{S_e}{\sigma^2}\Big/(n-2)}}=\frac{(\hat{\beta}_1-\beta_1)\sqrt{l_{xx}}}{\sqrt{\dfrac{S_e}{n-2}}}=\frac{(\hat{\beta}_1-\beta_1)\sqrt{l_{xx}}}{\hat{\sigma}}\sim t(n-2)\text{。}$$

步骤：

（1）假设 $H_0:\beta_1=0$ $vs$ $H_1:\beta_1\neq 0$。

（2）统计量 $T=\dfrac{(\hat{\beta}_1-\beta_1)\sqrt{l_{xx}}}{\hat{\sigma}}\sim t(n-2)$，检验统计量 $T=\dfrac{\hat{\beta}_1\sqrt{l_{xx}}}{\hat{\sigma}}$。

（3）拒绝域：$W=\{|t|\geq t_{1-\alpha/2}(n-2)\}$。

（4）计算检验统计量观测值 $t$ 与检验的 $p$ 值，并作出决策。

这是 $T$ 检验法。

注意到

$$T^2=\frac{\hat{\beta}_1^2 l_{xx}}{\hat{\sigma}^2}=\frac{S_R}{S_e/(n-2)}=F\text{，}$$

可见 $T$ 检验与 $F$ 检验本质上是等价的。

三．相关系数 $r$ 检验

对应于总体相关系数

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}=\frac{E\{[X-E(X)][Y-E(Y)]\}}{\sqrt{E[X-E(X)]^2}\sqrt{E[Y-E(Y)]^2}}\text{，}$$

定义样本相关系数

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{l_{XY}}{\sqrt{l_{XX}}\sqrt{l_{YY}}}\text{。}$$

可得$|r|\le 1$，且当$|r|=1$时，$X_i$与$Y_i$具有完全的线性关系，即存在常数$a,b$，使得$Y_i=aX_i+b$。如果$|r|$越接近 1，表明$X_i$与$Y_i$的线性关系越强；如果$|r|$越接近 0，表明$X_i$与$Y_i$的线性关系越弱。

对于假设$H_0:\beta_1=0$，可将拒绝域取为$W=\{|r|\ge c\}$的形式。

步骤：

（1）假设$H_0:\beta_1=0\ \ vs\ \ H_1:\beta_1\ne 0$。

（2）检验统计量$r=\dfrac{l_{xY}}{\sqrt{l_{xx}}\sqrt{l_{YY}}}$。

（3）拒绝域：$W=\{|r|\ge r_{1-\alpha}(n-2)\}$。

（4）计算检验统计量观测值$r$，并作出决策。

这是$r$检验法。

因$r=\dfrac{l_{xY}}{\sqrt{l_{xx}}\sqrt{l_{YY}}}$，则

$$r^2=\frac{l_{xY}^2}{l_{xx}l_{YY}}=\frac{\hat\beta_1^2 l_{xx}}{l_{YY}}=\frac{S_R}{S_T}=\frac{S_R}{S_e+S_R}=\frac{\dfrac{S_R}{S_e/(n-2)}}{n-2+\dfrac{S_R}{S_e/(n-2)}}=\frac{F}{n-2+F},$$

可见相关系数$r$检验与$F$检验本质上是等价的。根据$F$分布的分位数，可得$r$检验的分位数

$$r_{1-\alpha}(n-2)=\sqrt{\frac{F_{1-\alpha}(1,n-2)}{n-2+F_{1-\alpha}(1,n-2)}}。$$

**例** 根据前例中合金钢强度和碳含量数据，对回归方程作显著性检验（$\alpha=0.01$）。

**解：**根据试验数据可得

$$l_{xx}=0.018567，\ l_{xy}=2.4675，\ l_{yy}=345.0625，$$

$$\hat\beta_1=\frac{l_{xy}}{l_{xx}}=132.8995，\ \hat\beta_0=\bar y-\hat\beta_1\bar x=28.0826，$$

故回归方程为

$$\hat Y=28.0826+132.8995x。$$

$F$检验，假设$H_0:\beta_1=0\ \ vs\ \ H_1:\beta_1\ne 0$，检验统计量$F=\dfrac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，显著水平$\alpha=0.01$，

$n=12$，$F_{1-\alpha}(1,n-2)=F_{0.99}(1,10)=10.04$，右侧拒绝域$W=\{f\ge 10.04\}$。

因

$$S_R=\hat\beta_1^2 l_{xx}=132.8995^2\times 0.018567=327.9294，\ S_T=l_{yy}=345.0625，\ S_e=S_T-S_R=17.1331，$$

则

$$f=\frac{S_R}{S_e/(n-2)}=\frac{327.9294}{17.1331/10}=191.4013\in W\ ,$$

$$p=P\{F\geq 191.4013\}=7.5853\times 10^{-8}<\alpha=0.01\ ,$$

故拒绝 $H_0$，回归方程显著。

方差分析表

| 来源 | 平方和 | 自由度 | 均方和 | $F$ 比 | $p$ 值 |
|---|---|---|---|---|---|
| 回归 | 327.9294 | 1 | 327.9294 | 191.4013 | $7.5853\times 10^{-8}$ |
| 残差 | 17.1331 | 10 | 1.7133 | | |
| 总和 | 345.0625 | 11 | | | |

8.4.5　估计与预测

当经检验回归方程为显著时，可对回归系数 $\beta_1$ 与 $\beta_0$，误差方差 $\sigma^2$，点 $x_0$ 处函数值的期望 $E(Y_0)$ 分别作出估计，以及对函数值 $Y_0$ 作出预测。

一．估计

参数 $\beta_1,\beta_0,\sigma^2,E(Y_0)=\beta_0+\beta_1x_0$ 点估计分别是

$$\hat{\beta}_1=\frac{l_{xY}}{l_{xx}}\ ,\quad \hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{x}\ ,\quad \hat{\sigma^2}=\frac{S_e}{n-2}\ ,\quad \hat{E(Y_0)}=\hat{\beta}_0+\hat{\beta}_1x_0\ 。$$

因 $\hat{\beta}_1\sim N\left(\beta_1,\frac{\sigma^2}{l_{xx}}\right)$，$\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{l_{xx}}}\sim N(0,1)$，用 $\hat{\sigma}=\sqrt{\frac{S_e}{n-2}}$ 替换 $\sigma$，有 $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}/\sqrt{l_{xx}}}\sim t(n-2)$，可得 $\beta_1$ 的 $1-\alpha$ 置信区间为

$$\left[\hat{\beta}_1\pm t_{1-\alpha/2}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{l_{xx}}}\right]。$$

因 $\hat{\beta}_0\sim N\left(\beta_0,\left(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}\right)\sigma^2\right)$，$\frac{\hat{\beta}_0-\beta_0}{\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}}}\sim N(0,1)$，用 $\hat{\sigma}=\sqrt{\frac{S_e}{n-2}}$ 替换 $\sigma$，有 $\frac{\hat{\beta}_0-\beta_0}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}}}\sim t(n-2)$，可得 $\beta_0$ 的 $1-\alpha$ 置信区间为

$$\left[\hat{\beta}_0\pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}}\right]。$$

因 $\frac{S_e}{\sigma^2}=\frac{(n-2)\hat{\sigma^2}}{\sigma^2}\sim\chi^2(n-2)$，可得 $\sigma^2$ 的 $1-\alpha$ 置信区间为

$$\left[\frac{(n-2)\hat{\sigma^2}}{\chi^2_{1-\alpha/2}(n-2)},\frac{(n-2)\hat{\sigma^2}}{\chi^2_{\alpha/2}(n-2)}\right]。$$

因 $E(\hat{Y_0})=\hat{\beta}_0+\hat{\beta}_1x_0\sim N\left(\beta_0+\beta_1x_0,\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\sigma^2\right)$，有 $\frac{(\hat{\beta}_0+\hat{\beta}_1x_0)-(\beta_0+\beta_1x_0)}{\sigma\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\sim N(0,1)$，用 $\hat{\sigma}=\sqrt{\frac{S_e}{n-2}}$ 替换 $\sigma$，有 $\frac{(\hat{\beta}_0+\hat{\beta}_1x_0)-(\beta_0+\beta_1x_0)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\sim t(n-2)$，可得 $E(Y_0)=\beta_0+\beta_1x_0$ 的 $1-\alpha$ 置信区间为

$$\left[\hat{\beta}_0+\hat{\beta}_1x_0\pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right]。$$

二．$Y_0$ 的预测区间

因函数值 $Y_0=\beta_0+\beta_1x_0+\varepsilon$ 的预测值为 $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$，与期望值 $E(Y_0)=\beta_0+\beta_1x_0$ 的点估计 $E(\hat{Y_0})$ 相同，但 $Y_0$ 的预测区间需要考虑随机误差 $\varepsilon$ 的影响，因而不同于 $E(Y_0)$ 的置信区间。期望值 $E(Y_0)$ 的置信区间是对 $\beta_0+\beta_1x_0$ 作出估计的取值范围，而函数值 $Y_0$ 的预测区间是对 $\beta_0+\beta_1x_0+\varepsilon$ 作出预测的取值范围。

因

$$Y_0=\beta_0+\beta_1x_0+\varepsilon\sim N(\beta_0+\beta_1x_0,\sigma^2)，\ \hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1x_0\sim N\left(\beta_0+\beta_1x_0,\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\sigma^2\right),$$

且 $Y_0$ 与 $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1x_0$ 相互独立，则

$$Y_0-\hat{Y}_0=Y_0-\hat{\beta}_0-\hat{\beta}_1x_0\sim N\left(0,\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\sigma^2\right)，\ \frac{Y_0-\hat{\beta}_0-\hat{\beta}_1x_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\sim N(0,1)，$$

用 $\hat{\sigma}=\sqrt{\frac{S_e}{n-2}}$ 替换 $\sigma$，有

$$\frac{Y_0-\hat{\beta}_0-\hat{\beta}_1x_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\sim t(n-2)，$$

可得 $Y_0$ 的 $1-\alpha$ 预测区间为

$$\left[\hat{\beta}_0+\hat{\beta}_1x_0\pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right]。$$

将 $Y_0$ 的预测区间与 $E(Y_0)$ 的置信区间比较，就是根号中多了个 1，这是由随机误差 $\varepsilon$ 造成的。预测区

间在 $x_0=\bar{x}$ 处区间长度最短。当 $n$ 很大时，有

$$t_{1-\alpha/2}(n-2)\approx u_{1-\alpha/2}\,,\quad \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\approx 1\,,$$

即预测区间近似为 $[\hat{\beta}_0+\hat{\beta}_1 x_0 \pm u_{1-\alpha/2}\cdot\hat{\sigma}]$。

**例** 为了考察某企业产量与成本的关系，调查获得 5 组数据：

| 产量（吨） | 25 | 28 | 30 | 32 | 35 |
|---|---|---|---|---|---|
| 成本（万元） | 384 | 395 | 412 | 417 | 430 |

求：

（1）产量与成本的线性回归方程。

（2）对回归方程作显著性检验（$\alpha=0.01$）。

（3）回归系数 $\beta_1,\beta_0$，误差方差 $\sigma^2$ 以及产量为 40（吨）时平均成本 $E(Y_0)$ 的置信区间（$\alpha=0.01$）。

（4）产量为 40（吨）时成本 $Y_0$ 的预测区间（$\alpha=0.01$）。

**解**：（1）根据试验数据得到计算表：

试验数据计算表

| | | |
|---|---|---|
| $\sum x_i=150$ | $n=5$ | $\sum y_i=2038$ |
| $\bar{x}=30$ | | $\bar{y}=407.6$ |
| $\sum x_i^2=4558$ | $\sum x_i y_i=61414$ | $\sum y_i^2=832014$ |
| $n\bar{x}^2=4500$ | $n\bar{x}\bar{y}=61140$ | $n\bar{y}^2=830688.8$ |
| $l_{xx}=58$ | $l_{xy}=274$ | $l_{yy}=1325.2$ |
| | $\hat{\beta}_1=l_{xy}/l_{xx}=4.7241$<br>$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}=265.8759$ | |

故回归方程为

$$\hat{Y}=265.8759+4.7241x\,。$$

（2）$F$ 检验：假设 $H_0:\beta_1=0$ *vs* $H_1:\beta_1\neq 0$，检验统计量 $F=\dfrac{S_R}{S_e/(n-2)}\sim F(1,n-2)$，显著水平 $\alpha=0.01$，$n=5$，$F_{1-\alpha}(1,n-2)=F_{0.99}(1,3)=34.12$，右侧拒绝域 $W=\{f\geq 34.12\}$。

因

$$S_R=\hat{\beta}_1^2 l_{xx}=4.7241^2\times 58=1294.4138\,,\quad S_T=l_{yy}=1325.2\,,\quad S_e=S_T-S_R=30.7862\,,$$

则

$$f=\frac{S_R}{S_e/(n-2)}=\frac{1294.4138}{30.7862/3}=126.1358\in W\,,$$

$$p=P\{F\geq 126.1358\}=0.0015<\alpha=0.01\,,$$

故拒绝 $H_0$，回归方程显著。

方差分析表

| 来源 | 平方和 | 自由度 | 均方和 | $F$ 比 | $p$ 值 |
|---|---|---|---|---|---|
| 回归 | 1294.4138 | 1 | 1294.4138 | 126.1358 | 0.0015 |
| 残差 | 30.7862 | 3 | 10.2621 | | |
| 总和 | 1325.2 | 4 | | | |

（3）因显著水平 $\alpha=0.01$，$n=5$，有 $t_{1-\alpha/2}(n-2)=t_{0.995}(3)=5.8409$，$\chi^2_{1-\alpha/2}(n-2)=\chi^2_{0.005}(3)=0.0717$，$\chi^2_{1-\alpha/2}(n-2)=\chi^2_{0.995}(3)=14.8603$，且

$$\bar{x}=30,\ l_{xx}=58,\ \hat{\sigma}=\sqrt{\frac{S_e}{n-2}}=\sqrt{\frac{30.7862}{3}}=3.2034,$$

故回归系数 $\beta_1$ 的 0.99 置信区间为

$$\left[\hat{\beta}_1\pm t_{1-\alpha/2}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{l_{xx}}}\right]=\left[4.7241\pm5.8409\times\frac{3.2034}{\sqrt{58}}\right]=[2.2673,7.1809];$$

回归系数 $\beta_0$ 的 0.99 置信区间为

$$\left[\hat{\beta}_0\pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}}}\right]=\left[265.8759\pm5.8409\times3.2034\times\sqrt{\frac{1}{5}+\frac{30^2}{58}}\right]$$

$$=[191.6961,340.0556]。$$

误差方差 $\sigma^2$ 的 0.99 置信区间为

$$\left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-2)},\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-2)}\right]=\left[\frac{3\times3.2034^2}{14.8603},\frac{3\times3.2034^2}{0.0717}\right]=[2.0717,429.3753]。$$

产量为 $x_0=40$（吨）时平均成本 $E(Y_0)$ 的 0.99 置信区间为

$$\left[\hat{\beta}_0+\hat{\beta}_1x_0\pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right]$$

$$=\left[265.8759+4.7241\times40\pm5.8409\times3.2034\times\sqrt{\frac{1}{5}+\frac{(40-30)^2}{58}}\right]$$

$$=[428.8867,480.7960]。$$

（4）产量为 $x_0=40$（吨）时成本 $Y_0$ 的 0.99 预测区间为

$$\left[\hat{\beta}_0+\hat{\beta}_1x_0\pm t_{1-\alpha/2}(n-2)\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}\right]$$

$$=\left[265.8759+4.7241\times40\pm5.8409\times3.2034\times\sqrt{1+\frac{1}{5}+\frac{(40-30)^2}{58}}\right]$$

$$=[422.8453,486.8374]。$$

## §8.5　一元非线性回归

回归分析时，首先根据观测数据作出散点图，如果由散点图判断出回归函数不是线性函数时，则需要选取适当的非线性函数，通常是多项式或者可将其化为线性函数，常见的有双曲线函数$\frac{1}{y}=a+\frac{b}{x}$，幂函数$y=ax^b$，指数函数$y=ae^{bx}$，$y=ae^{\frac{b}{x}}$，对数函数$y=a+b\ln x$，S 形曲线$y=\frac{1}{a+be^{-x}}$等。

如果是多项式，则可令$x_1=x, x_2=x^2, \cdots, x_m=x^m$，再采用多元线性回归进行处理。如果可化为线性函数，则先换元化为线性函数，再采用一元线性回归进行处理。如双曲线函数$\frac{1}{y}=a+\frac{b}{x}$，令$u=\frac{1}{x}$，$v=\frac{1}{y}$，化为线性函数$v=a+bu$；幂函数$y=ax^b$，有$\ln y=\ln a+b\ln x$，令$u=\ln x$，$v=\ln y$，化为线性函数$v=\ln a+bu$；指数函数$y=ae^{bx}$，有$\ln y=\ln a+bx$，令$u=x$，$v=\ln y$，化为线性函数$v=\ln a+bu$；指数函数$y=ae^{\frac{b}{x}}$，有$\ln y=\ln a+\frac{b}{x}$，令$u=\frac{1}{x}$，$v=\ln y$，化为线性函数$v=\ln a+bu$；对数函数$y=a+b\ln x$，令$u=\ln x$，$v=y$，化为线性函数$v=a+bu$；S 形曲线$y=\frac{1}{a+be^{-x}}$，有$\frac{1}{y}=a+be^{-x}$，令$u=e^{-x}$，$v=\frac{1}{y}$，化为线性函数$v=a+bu$。

这一节讨论可化为线性函数的情形。

### 8.5.1　确定可能的函数形式

根据$(x_i, y_i),\ \ i=1,2,\cdots,n$的散点图，估计函数形式$y=f(x)$，化为线性函数$v=a+bu$，再根据$(u_i, v_i)$的散点图，判断是否近似在一条直线上。若是，则对$(u, v)$作一元线性回归；否则，更换函数形式。

### 8.5.2　参数估计

对$(u, v)$作一元线性回归，得回归方程$\hat{v}=\hat{a}+\hat{b}u$，再化为$(x, y)$的非线性回归方程$\hat{y}=\hat{f}(x)$。

### 8.5.3　曲线回归方程的比较

对于非线性回归问题，通常需选取多个非线性函数，再根据结果进行比较。但此时$(u, v)$的线性拟合程度并不能完全反映$(x, y)$的非线性拟合程度，而应该直接根据$y_i$的值进行判定。

称$Y_i$的观测值$y_i$与回归值$\hat{y}_i=\hat{f}(x_i)$之差$y_i-\hat{y}_i$为残差，$\sum(y_i-\hat{y}_i)^2$为残差平方和，显然残差平方和越小，表明拟合程度越高。

在一元线性回归问题中，相关系数的平方为

$$r^2=\frac{l_{xy}^2}{l_{xx}l_{yy}}=\frac{S_R}{S_T}=1-\frac{S_e}{S_T}=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2},$$

而在一元非线性回归问题中，称

$$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2},$$

为决定系数，决定系数越接近 1，表明拟合程度越高。

**注：**由于一元非线性回归问题中平方和分解不成立，可能出现决定系数小于 0 的情况。

在一元线性回归问题中，误差标准差

$$\hat{\sigma}=\sqrt{\frac{S_e}{n-2}}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}},$$

而在一元非线性回归问题中，称

$$s=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}$$

为剩余标准差，剩余标准差越小，表明拟合程度越高。

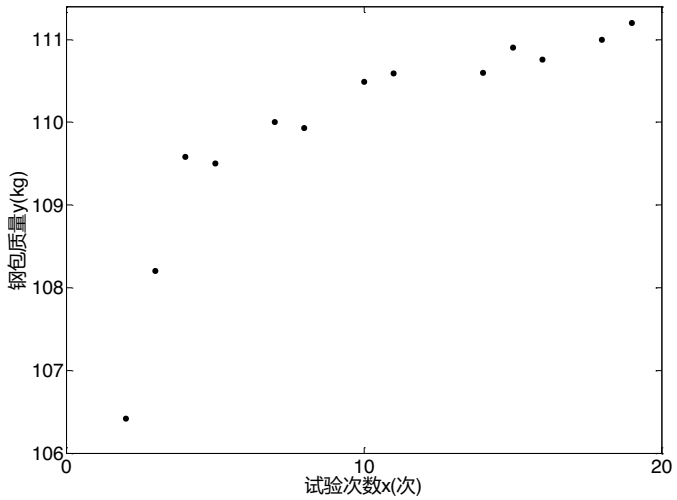
通常用决定系数或剩余标准差反映非线性回归方程的拟合程度。

**例** 炼钢厂出钢水时用的钢包，在使用过程中由于钢水的侵蚀，其容积不断增大。钢包容积用盛满钢水时的质量 $Y$（kg）表示，相应的试验次数用 $x$ 表示。根据表中的数据，作出散点图，并选择一个合适的回归函数形式。

**钢包的质量 $y$ 与试验次数 $x$ 的数据**

| 序号 | $x$（次） | $y$（kg） | 序号 | $x$（次） | $y$（kg） |
|---|---|---|---|---|---|
| 1 | 2 | 106.42 | 8 | 11 | 110.59 |
| 2 | 3 | 108.20 | 9 | 14 | 110.60 |
| 3 | 4 | 109.58 | 10 | 15 | 110.90 |
| 4 | 5 | 109.50 | 11 | 16 | 110.76 |
| 5 | 7 | 110.00 | 12 | 18 | 111.00 |
| 6 | 8 | 109.93 | 13 | 19 | 110.20 |
| 7 | 10 | 110.49 | | | |

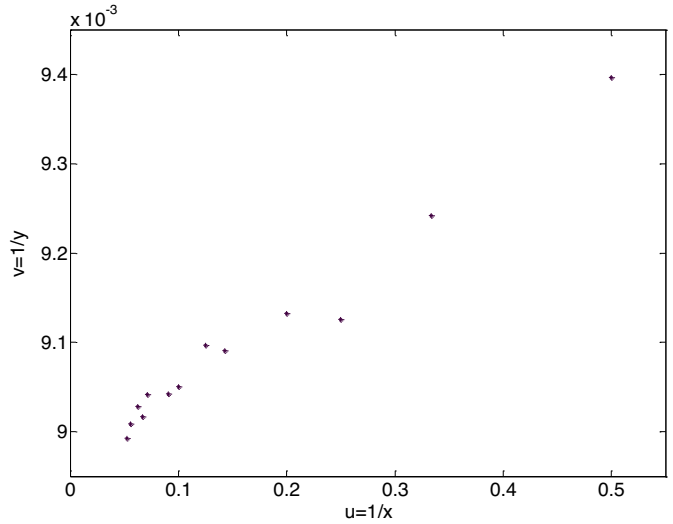
**解：**根据观测数据作 $(x, y)$ 的散点图

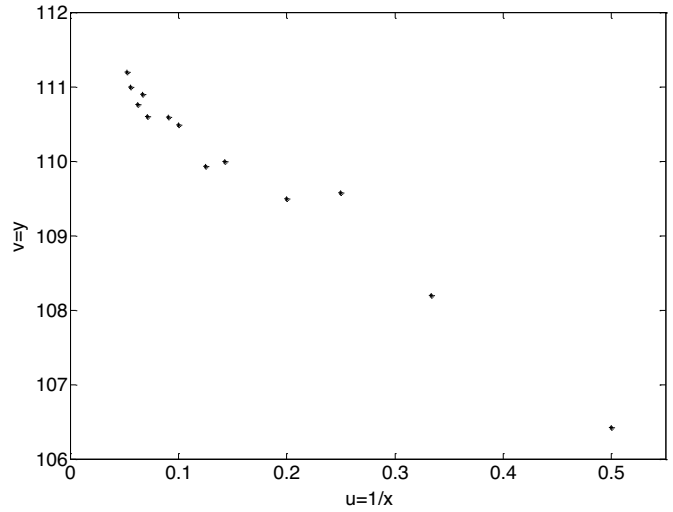



可以看出这些点并不是位于一条直线上，根据散点图，回归函数可以取为以下几个非线性函数：

$$\frac{1}{y}=a+\frac{b}{x},\quad y=a+\frac{b}{x},\quad y=a+b\ln x,\quad y=a+b\sqrt{x},\quad y-100=ae^{-\frac{x}{b}},$$

其中 $a, b$ 为未知参数。将 $(x, y)$ 的非线性函数化为 $(u, v)$ 的线性函数后，作 $(u, v)$ 散点图，只有 $\frac{1}{y} = a + \frac{b}{x}$ 或 $y = a + \frac{b}{x}$ 比较合适。




对于 $\frac{1}{y} = a + \frac{b}{x}$，令 $u = \frac{1}{x}$，$v = \frac{1}{y}$，化为线性函数 $v = a + bu$，根据 $(u_i, v_i)$ 作一元线性回归，得：

$\hat{a} = 0.00896663$，$\hat{b} = 0.00082917$，

回归方程为 $\hat{v} = 0.00896663 + 0.00082917u$，即 $\frac{1}{\hat{y}} = 0.00896663 + \frac{0.00082917}{x}$，故

$$\hat{y} = \frac{x}{0.00896663x + 0.00082917}$$。

又因 $\sum (y_i - \hat{y}_i)^2 = 0.5743$，$\sum (y_i - \bar{y})^2 = 21.2105$，故决定系数和剩余标准差分别为

$$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}=1-\frac{0.5743}{21.2105}=0.9729,$$

$$s=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{0.5743}{11}}=0.2285。$$

对于 $y=a+\frac{b}{x}$，令 $u=\frac{1}{x}$，$v=y$，化为线性函数 $v=a+bu$，根据 $(u_i,v_i)$ 作一元线性回归，得：

$$\hat{a}=111.4875，\ \hat{b}=-9.8334，$$

回归方程为 $\hat{v}=111.4875-9.8334u$，故

$$\hat{y}=111.4875-\frac{9.8334}{x}。$$

又因 $\sum(y_i-\hat{y}_i)^2=0.5496$，$\sum(y_i-\bar{y})^2=21.2105$，故决定系数和剩余标准差分别为

$$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}=1-\frac{0.5496}{21.2105}=0.9741,$$

$$s=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{0.5496}{11}}=0.2235。$$

经过比较，选取 $y=a+\frac{b}{x}$ 的形式更好，回归方程为

$$\hat{y}=111.4875-\frac{9.8334}{x}。$$