第五章　广义与一般线性模型

一、数据的分类

二、模型选择方式

三、广义线性模型
- 1.logistic模型
- 2.对数线性模型

四、一般线性模型
- 1.完全随机设计模型
- 2.随机单位组设计模型
- 3.析因设计模型
- 4.正交试验设计模型

一.数据的分类

变量取值方式：

（1）连续变量

如胸径、树高、生长量等

（2）“0-1”变量或称二分类变量

如实验成功、失败，有效、无效；性别：男、女

（3）有序变量（等级变量）

如施肥效果，立地质量、土壤剖面等；

二.模型选择方式

1.y为连续变量

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon = X\beta + \varepsilon$$ 一般线性模型

其中，ε 为随机误差，$E(\varepsilon) = 0$ 。

（1）当自变量为连续变量时，也就是上讲讲的**线性回归模型**，为向量，X为矩阵；

（2）当自变量x是由因素构成的哑变量，y为反应变量（实验结果），X为设计阵。模型称为**实验设计模型或方差分析模型**。

$$Y = X\beta + Z\alpha + \varepsilon$$

（3）当一部分xi是根据因素产生的哑变量，另一部分zi是变量，模型称为**协方差模型**。

X为哑变量构成的设计阵，Z为变量构成的观察阵。

2.y为0-1变量

一般用logistic回归模型来描述y与x之间的关系

3.Y为有序变量

一般用 累积比数模型和对数模型

4.y为多分类变量

对数线性模型和多分类logistic回归模型

二.模型选择方式

基本公式 $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon = X\beta + \varepsilon$$

$$E(\varepsilon) = 0, cov(\varepsilon) = \sigma^2 I$$

y / x	连续变量	0-1变量	有序变量	多分类变量	连续伴有删失
连续变量	线性回归方程	logistic回归模型	累积比数模型 对数线性模型	对数线性模型 多分类logistic回归模型	cox比例风险模型
分类变量	实验设计模型（方差分析模型）				
连续变量 分类变量	协方差分析模型				

三.广义线性模型

一般线性模型：自变量为定性变量的线性模型，如实验设计模型、方差分析模型。其基本假设是y服从正态分布或者至少y的方差 σ^2 为有限常数。

广义线性模型： 因变量为非正态分布的线性模型，如logistic回归模型、对数线性模型和Cox比例风险模型。

$$E(y) = \mu$$
$$m(\mu) = X\beta$$
$$\mathrm{cov}(y) = \sigma^2 V(\mu)$$

连接函数m(.)组成的向量将 μ 转化成 β 的线性表达式；
$V(\mu)$ 为n*n的矩阵，其中每个元素都是 μ 的函数；

当各 y_i 值相互独立时，$V(\mu)$ 为对角矩阵；
当 $m(\mu) = \mu$ $V(\mu) = I$ 时，左式为一般线性模型。

三.广义线性模型

在广义线性模型中，均假定观察值y具有指数族概率密度函数

$$f(y|\theta,\phi) = exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\} \quad (5\text{-}4)$$

其中$a(\cdot)$、$b(\cdot)$和$c(\cdot)$是三种函数形式，θ为典则参数。

广义线性模型中的常用分布族：

分布	函数	模型
正态（高斯）	$E(y) = X'\beta$	普通线性模型
二项（Binomial）	$E(y) = \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$	Logistic模型和概率模型单位模型
泊松（Poisson）	$E(y) = \exp(X'\beta)$	对数线性模型

在广义线性模型中，（5.4）式中的典则参数不仅仅是μ的函数，还是参数$\beta_0, \beta_1, \ldots, \beta_p$的线性表达式。对$\mu$作变换，则可得到这三种分布连接函数的形式

正态分布：$m(\mu) = \mu = \sum \beta_j x_j$

二项分布：$m(\mu) = log(\frac{\mu}{1-\mu}) = \sum \beta_j x_j$

Poisson 分布：$m(\mu) = log(\mu) = \sum \beta_j x_j$

广义线性模型函数 glm()的用法

glm(formula, family = gaussian, data,...)

formula 为公式，即为要拟合的模型

family 为分布族，包括正态分布（gaussian）、二项分布（binomial）、泊松分布（poission）和伽玛分布（gamma），分布族还可以通过选项 link=来指定使用的连接函数

data 为可选择的数据框。

（一）Logistic回归模型

1 模型的引进

2 Logistic回归模型估计

3 Logistic回归模型的评价

4 Logistic回归系数的统计推断

1 模型的引进

因变量是二分类定性变量时，考虑简单线性模型：

其中yi服从两点分布：$y_i = \beta_0 + \beta_1 \mathbf{x}_i + \varepsilon_i$

可知

$$P(y_i = 1) = \pi_i$$
$$P(y_i = 0) = 1 - \pi_i$$
$$E(y_i) = \pi_i = \beta_0 + \beta_1 x_i$$

logistic回归模型

某疾病的病例对照研究

Id	y	x1	x2	x3	…
1	1	1	1	3	…
2	1	0	3	2	…
…	…	…	…	…	…
…	1	1	0	1	…
…	0	1	4	0	…
…	0	0	6	0	…
…	…	…	…	…	…
N	0	0	1	1	…

logistic回归模型

研究目的：X_1，X_2，X_3等因素对该疾病有无影响？

建立Y与X的多重线性回归模型？

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

（取值0和1）

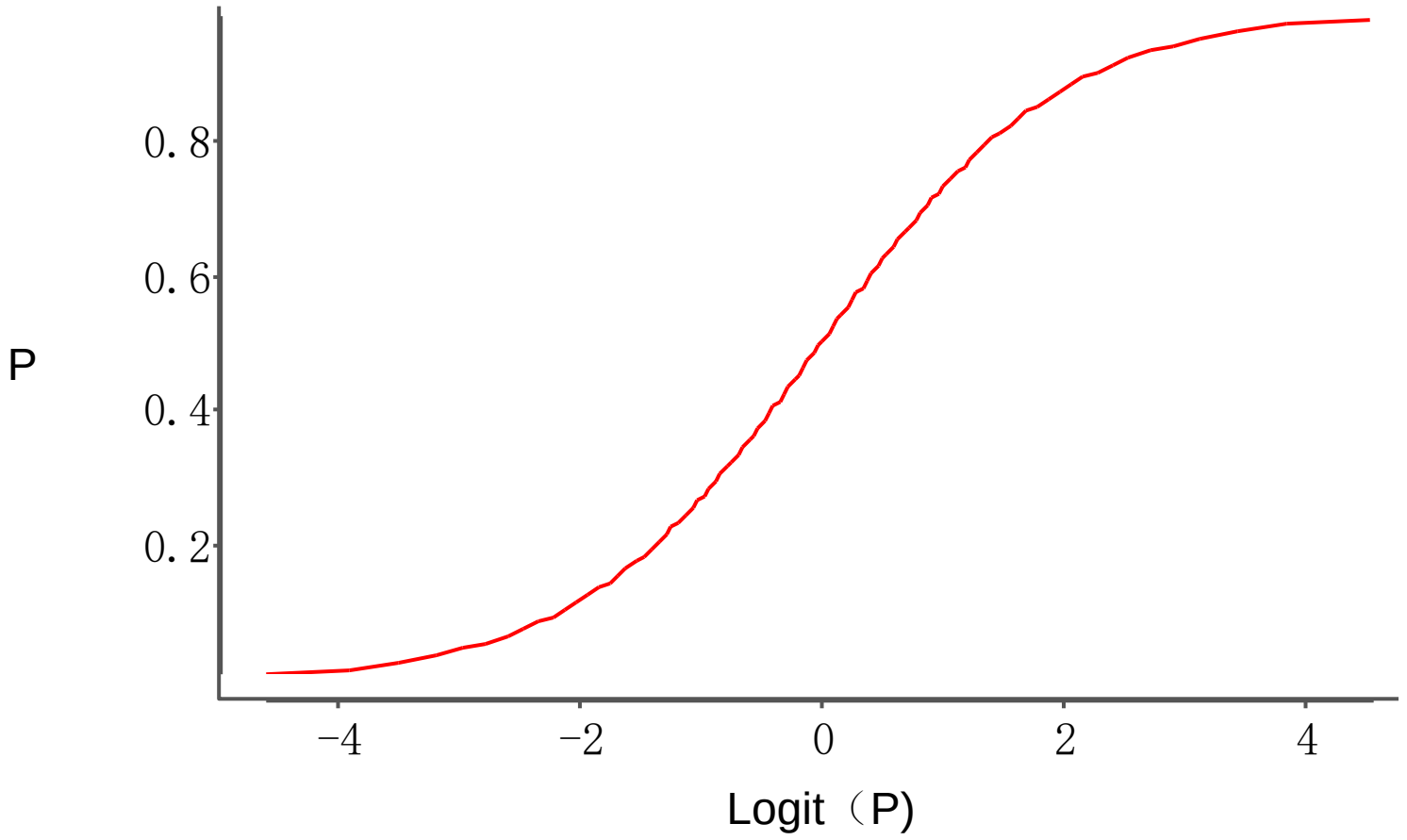
logistic回归模型

建立p(Y=1|X)与X的多重线性回归模型？

$$\mathrm{p}(Y=1|X)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3$$

(取值范围0～1)

logistic回归模型



logistic回归模型

建立logit（p）与X的多重线性回归模型:

优势(odds)

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

$$\ln\left(\frac{p(Y=1/X)}{1-p(Y=1/X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

(取值范围-∞～+∞)

logistic回归模型的一般形式

$$\text{logit（p）} = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

$$p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}$$

$$p = \frac{1}{1+e^{-\left(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k\right)}}$$

Logistic回归模型

1 模型的引进

2. Logistic回归模型估计

3. Logistic回归模型的评价

4. Logistic回归系数的统计推断

2 Logistic回归模型估计

Logistic回归模型估计的假设条件与OLS的不同

（1）logistic回归的因变量是二分类变量

（2）logistic回归的因变量与自变量之间的关系是非线性的

（3）logistic回归中无相同分布的假设

（4）logistic回归没有关于自变量“分布”的假设（离散，连续，虚拟）

最大似然估计

假设n个样本观测值$y_1,y_2\cdots\cdots y_n$，得到一个观测值的概率为

$$P(Y = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

其中 $y_i = 1$ 或 $y_i = 0$

由于各项观测相互独立，其联合分布为：

$$L(\theta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

选择上式作为n个观测的似然函数

$$\ln L(\theta) = \ln\left(\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\right)$$

$$= \ln\left(\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{-y_i}(1-p_i)\right)$$

$$= \ln\left(\prod_{i=1}^{n} \left(\frac{p_i}{1-p_i}\right)^{y_i}(1-p_i)\right)$$

$$= \sum_{i=1}^{n}\left[y_i(\beta_0+\beta_1 x_i) - \ln\left(1+e^{\beta_0+\beta_1 x_i}\right)\right]$$

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x$$

$$1\text{-}p = \frac{1}{1+e^{\beta_0+\beta_1 x}}$$

分别对参数求偏导，然后令它等于0：

$$\frac{\partial \ln L(\theta)}{\partial \beta_0} = \sum_{i=1}^{n}\left[y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right] = 0$$

$$\frac{\partial \ln L(\theta)}{\partial \beta_1} = \sum_{i=1}^{n}\left[y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right] x_i = 0$$

求得 β_0，β_1 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$，从而得到 $\hat{p}_i$ (p_i的极大似然估计)，这个值是在给定x_i的条件下y_i=1的条件概率的估计，它代表了Logistic回归模型的拟合值。

Logistic回归模型

1. 模型的引进
2. Logistic回归模型估计
3. Logistic回归模型的评价
4. Logistic回归系数的参数检验

3 Logistic回归模型的评价

- 拟合优度检验（Goodness of fit）
  - 似然比检验（Likelihood Ratio Test）
  - Hosmer-Lemeshow检验

- **似然比检验的思想：**

  建立logistic回归模型后，再向模型中引入另外的变量，重新拟合模型。两模型的$-2\ln L$值之差即为似然比统计量LR。

似然比检验

似然比检验用公式表示为:

$$LR = \ln\left(\frac{L_s}{L_0}\right)^2 \sim \chi^2(p) = 2\ln L_s - 2\ln L_0$$

$2\ln L_0$为只有截距项的零假设模型的2lnL,

$2\ln L_s$为设定模型的2lnL，当样本含量较大时,

LR服从卡方分布，自由度为设定模型与零假设模型自由度之差。

Hosmer-Lemeshow检验

该方法根据模型预测概率的大小将所有观察单位分为十组，然后根据每一组中因变量各种取值的实际值与理论值计算Pearson卡方：

$$HL = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g}$$

其中G代表分组数。O_g为第g组的观测频数，E_g为第g组的预测频数。

例题：高中毕业生继续进入大学学习的可能性的影响因素

如果一个高中毕业生升入了大学，则y=1；如果没有升入大学，则y=0。P为高中毕业后升入大学的概率。

自变量为性别Gender（1为男性，0为女性），高中类型Keysch（1为重点中学、0为普通中学），高中成绩Meangr。前两个为虚拟变量，Meangr为连续变量。

Logistic回归模型为：

$$\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = \beta_0 + \beta_1 \mathrm{Gender} + \beta_2 \mathrm{Keysch} + \beta_3 \mathrm{Meangr}$$

数据

	keysch	grade	college	meangr	gender
1	.00	82.20	.00	-1.50	.00
2	.00	91.80	1.00	8.10	.00
3	.00	89.80	1.00	6.10	.00
4	.00	87.30	1.00	3.60	.00
5	.00	81.70	.00	-2.00	.00
6	.00	79.00	.00	-4.70	.00
7	.00	84.70	.00	1.00	.00
8	.00	79.30	.00	-4.40	.00
9	.00	78.40	.00	-5.30	.00
10	.00	80.50	.00	-3.20	.00
11	.00	79.80	.00	-3.90	.00

$$\ln\left(\frac{p}{1-p}\right) = -1.757 + 0.866\text{Gende r} + 0.913\text{Keysc h} + 0.425\text{Meang r}$$

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.803	8	.558

Contingency Table for Hosmer and Lemeshow Test

		college = .00		college = 1.00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	99	98.948	1	1.052	100
	2	97	96.424	3	3.576	100
	3	94	94.660	8	7.340	102
	4	85	88.687	16	12.313	101
	5	78	79.555	22	20.445	100
	6	73	69.368	27	30.632	100
	7	61	57.568	40	43.432	101
	8	49	43.557	50	55.443	99
	9	19	26.535	81	73.465	100
	10	10	9.698	87	87.302	97

Logistic回归模型

1. 模型的引进
2. Logistic回归模型估计
3. Logistic回归模型的评价
4. Logistic回归系数的参数检验

4 Logistic回归系数的参数检验

- $\widehat{\beta_i}$的检验

H_0: $\widehat{\beta_i} = 0$； H_1: $\widehat{\beta_i} \neq 0$

检验统计量：$Z = \frac{\widehat{\beta_i}}{Se(\widehat{\beta_i})} \sim N(0,1)$

如果$Z < Z_a$, 认为β_i=0； 否则认为$\beta_i \neq 0$。

- β_i的可信区间

β_i的可信区间为$\widehat{\beta_i} \pm Z_\alpha Se(\widehat{\beta_i})$。

例5-1：（1）建立全变量logistic回归模型：

```
d5.1=read.table("clipboard",header=T) #读取例5.1数据
logit.glm<-glm(y~x1+x2+x3,family=binomial,data=d5.1) #Logistic回归模型
summary(logit.glm) #Logistic回归模型结果
```

```
Call:
glm(formula = y ~ x1 + x2 + x3, family = binomial, data = d5.1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5636  -0.9131  -0.7892   0.9637   1.6000

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.597610   0.894831   0.668   0.5042
x1          -1.496084   0.704861  -2.123   0.0338 *
x2          -0.001595   0.016758  -0.095   0.9242
x3           0.315865   0.701093   0.451   0.6523
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.183  on 44  degrees of freedom
Residual deviance: 57.026  on 41  degrees of freedom
AIC: 65.026

Number of Fisher Scoring iterations: 4
```

可得初步的Logistic回归模型：

$$p=\frac{\exp(0.597610-1.496084x_1-0.001595x_2+0.315865x_3)}{1+\exp(0.597610-1.496084x_1-0.001595x_2+0.315865x_3)}$$

即 $Logit(p)=0597610-1.496084x_1-0.001595x_2+0.315865x_3$

（2）逐步筛选变量logistic回归模型：

上述模型中，由于β_2、β_3未通过检验，可类似于线性模型，用step()作变量筛选。

```
logit.step<-step(logit.glm,direction="both") #逐步筛选法变量选择
```

```
Start:  AIC=65.03
y ~ x1 + x2 + x3

       Df Deviance    AIC
- x2    1   57.035 63.035
- x3    1   57.232 63.232
<none>      57.026 65.026
- x1    1   61.936 67.936

Step:  AIC=63.03
y ~ x1 + x3

       Df Deviance    AIC
- x3    1   57.241 61.241
<none>      57.035 63.035
+ x2    1   57.026 65.026
- x1    1   61.991 65.991
```

```
Step:  AIC=61.24
y ~ x1

       Df Deviance    AIC
<none>      57.241 61.241
+ x3    1   57.035 63.035
+ x2    1   57.232 63.232
- x1    1   62.183 64.183
```

summary(logit.step) #逐步筛选法变量选择结果

```
Call:
glm(formula = y ~ x1, family = binomial, data = d5.1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4490  -0.8782  -0.8782   0.9282   1.5096

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6190     0.4688   1.320   0.1867
x1           -1.3728     0.6353  -2.161   0.0307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.183  on 44  degrees of freedom
Residual deviance: 57.241  on 43  degrees of freedom
AIC: 61.241

Number of Fisher Scoring iterations: 4
```

可以看出新的回归方程为：$p = \frac{\exp(0.6190-1.3728x_1)}{1+\exp(0.6190-1.3728x_1}$

（3）：预测发生交通事故的概率

对视力正常和视力有问题的司机分别做预测，即预测发生交通事故的概率。

```
pre1<-predict(logit.step,data.frame(x1=1)) #预测视力正常司机Logistic回归结果
p1<-exp(pre1)/(1+exp(pre1)) #预测视力正常司机发生事故概率
pre2<-predict(logit.step,data.frame(x1=0)) #预测视力有问题的司机Logistic回归结果
p2<-exp(pre2)/(1+exp(pre2)) #预测视力有问题的司机发生事故概率
c(p1,p2) #结果显示
```

可见，p_1 =0.32，p_2 =0.65，即视力有问题的司机发生交通事故的概率是视力正常的司机的两倍以上。

（二）对数线性模型

Poisson分布族模型和拟Poisson分布族模型的使用方法：

```
fm<- glm(formula, family=poisson(link=log),data=data.frame)
fm<- glm(formula, family=quasipoisson(link=log),data=data.frame)
```

$$\ln(E(y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$$即\ E(y) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$$

Poisson分布族模型和拟Poisson分布族模型的唯一差别：
Poisson分布族模型要求响应变量y是整数，
拟Poisson分布族模型无要求。

●5.2 顾客对产品的满意度分析

某企业想了解顾客对其产品是否满意，同时还想了解不同收入的人群对其产品的满意程度是否相同。

	满意	不满意	合计
高	53	38	91
中	434	108	542
低	111	48	159
合计	598	194	792

y	x1	x2
53	1	1
434	2	1
111	3	1
38	1	2
108	2	2
48	3	2

数据形式变为：y表示频数，x1表示收入人群，x2表示满意程度

在R数据中，y表示频数，x1表示收入人群，x2表示满意程度

```
> x=read.table("clipboard",header=T)
> x
   y   x1  x2
1  53   1  1
2 434  2  1
3 111  3  1
4  38   1  2
5 108  2  2
6   48   3  2
>log.glm <-
    glm(y~x1+x2,family=poisson(link=l
    og),data=x)
> summary(log.glm)
```

```
Call:
glm(formula = y ~ x1 + x2, family = poisson(link = log)

Deviance Residuals:
     1          2         3         4         5         6
-10.784   14.444   -8.468   -2.620    4.960   -3.142

Coefficients:
             Estimate    Std. Error   z value   Pr(>|z|)
(Intercept)  6.15687     0.14196      43.371    < 2e-16 ***
x1           0.12915     0.04370      2.955     0.00312 **
x2           -1.12573    0.08262      -13.625   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be
1)

    Null deviance: 662.84  on 5  degrees of freedom
Residual deviance: 437.97  on 3  degrees of freedom
AIC: 481.96

Number of Fisher Scoring iterations: 5
```

从右边的结果来看：

p1=0.0031<0.01

p2<0.01

说明收入和满意程度对产品有重要影响。

四、一般线性模型

这里讲的一般线性模型主要是针对几种典型的试验设计模型。

1.完全随机设计模型
2.随机单位设计模型
3.析因设计模型
4.正交实验设计模型

单因素试验方差分析表

方差来源	平方和	自由度	均方	F 比
因素 A	S_A	$s-1$	$\overline{S}_A = \dfrac{S_A}{s-1}$	$F = \dfrac{\overline{S}_A}{\overline{S}_E}$
误差	S_E	$n-s$	$\overline{S}_E = \dfrac{S_E}{n-s}$	
总和	S_T	$n-1$		

S_A：因素A的效应平方和

S_E：误差平方和

1.完全随机设计模型

处理因素A有G个水平，实验结果是 $y_{ij}, j = 1,2,\ldots,n_i, i = 1,2,\ldots,G$

此时模型为：$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, j = 1,2,\ldots,n_i, i = 1,2,\ldots,G$

$E(\varepsilon) = 0$

$\mathrm{cov}(\varepsilon) = \sigma^2 I$

μ是总体的均值，α_i是哑变量系数

ε_{ij}是误差项

用矩阵表示：$Y = X\beta + \varepsilon$

X为设计阵，元素为0或1

e是误差向量

Y为观察结果向量

$\beta = (\mu, \alpha_1, \alpha_2, \ldots, \alpha_G)'$

机器1	2.36	2.38	2.48	2.45	2.47	2.43
机器2	2.57	2.53	2.55	2.54	2.56	2.61
机器3	2.58	2.64	2.59	2.67	2.66	2.62

例：分析各机器生产的薄板厚度有无显著性差异

机器1	2.36	2.38	2.48	2.45	2.47	2.43
机器2	2.57	2.53	2.55	2.54	2.56	2.61
机器3	2.58	2.64	2.59	2.67	2.66	2.62

带入模型得：

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{16} \\ y_{21} \\ \vdots \\ y_{26} \\ y_{31} \\ \vdots \\ y_{36} \end{bmatrix} = \begin{bmatrix} 2.36 \\ \vdots \\ 2.43 \\ 2.57 \\ \vdots \\ 2.61 \\ 2.58 \\ \vdots \\ 2.26 \end{bmatrix} = \begin{bmatrix} \mu + 1 \cdot \alpha_1 + 0 \cdot \alpha_2 + 0 \cdot \alpha_3 \\ \vdots \\ \mu + 1 \cdot \alpha_1 + 0 \cdot \alpha_2 + 0 \cdot \alpha_3 \\ \mu + 0 \cdot \alpha_1 + 1 \cdot \alpha_2 + 0 \cdot \alpha_3 \\ \vdots \\ \mu + 0 \cdot \alpha_1 + 1 \cdot \alpha_2 + 0 \cdot \alpha_3 \\ \mu + 0 \cdot \alpha_1 + 0 \cdot \alpha_2 + 1 \cdot \alpha_3 \\ \vdots \\ \mu + 0 \cdot \alpha_1 + 0 \cdot \alpha_2 + 1 \cdot \alpha_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{16} \\ e_{21} \\ \vdots \\ e_{26} \\ e_{31} \\ \vdots \\ e_{36} \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Y　　Xβ　　e

$$\beta = (\mu, \alpha_1, \alpha_2, \alpha_3)'$$

用于软件的数据格式：

Y	A
2.36	1
2.38	1
2.48	1
2.45	1
2.47	1
2.43	1
2.57	2
2.53	2
2.55	2
2.54	2
2.56	2
2.61	2
2.58	3
2.64	3
2.59	3
2.67	3
2.66	3
2.62	3

```
> x=read.table("clipboard",header=T)
> anova(lm(Y~factor(A),data=x))
Analysis of Variance Table

Response: Y
           Df  Sum Sq   Mean Sq   F value   Pr(>F)
factor(A)   2   0.122     0.061    40.534  8.94e-07 ***
Residuals 15   0.023     0.0015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P<0.05，说明各机器生产的薄板厚度有显著性差异。

2.随机单位设计模型

处理因素A有G个水平，单位组B有n个看做n个水平
分别产生A的G个哑变量和单位组的n个哑变量
实验结果y_{ij}表示成：

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, j = 1,2,\ldots,n_i, i = 1,2,\ldots,G$$

其中：
μ为总均数；
α_i为处理因素A的第i个水平的效应；
β_j为第j个单位组的效应；
e_{ij}为误差项

	A1	A2	A3	A4
B1	582	491	601	758
B2	562	541	709	582
B3	653	516	392	487

例：分析各种燃料A与各种推进器B对火箭射程有无显著影响

	A1	A2	A3	A4
B1	582	491	601	758
B2	562	541	709	582
B3	653	516	392	487

表中处理因素是燃料A，单位组是推进器B，将实验结果代入

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{41} \\ y_{42} \\ y_{43} \end{bmatrix} = \begin{bmatrix} 582 \\ 562 \\ 653 \\ 491 \\ 541 \\ 516 \\ 601 \\ 709 \\ 392 \\ 758 \\ 582 \\ 487 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \\ e_{33} \\ e_{41} \\ e_{42} \\ e_{43} \end{bmatrix}$$

相应的数据格式为：

Y	A	B
582	1	1
491	2	1
601	3	1
758	4	1
562	1	2
541	2	2
709	3	2
582	4	2
653	1	3
516	2	3
392	3	3
487	4	3

```
> x=read.table("clipboard",header=T)
> anova(lm(Y~factor(A)+factor(B),data=x))
Analysis of Variance Table

Response: Y
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
factor(A)     3  15759     5253    0.4306  0.7387
factor(B)     2  22385    11192    0.9174  0.4491
Residuals     6  73198    12200
```

P_A>0.05，说明各种燃料A对火箭射程无显著性影响；
P_B>0.05，说明各种推进器B对火箭射程也无显著影响。

3.析因设计模型

先考虑两因素析因分析：

假定A因素有I个水平，B因素有J个水平；

实验中共有I*J个处理，每个处理重复r次。

两因素析因分析模型为：

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk} \quad j = 1,2,\ldots,J, i = 1,2,\ldots,I, k = 1,2,\ldots,r$$

αβ不是表示α×β，仅是一个符号，表示A、B因素间的交互作用。

例：研究两种方法提取甲、乙两种化合物的回收效果

采用2*2析因设计实验，各个处理重复4次。
实验结果（回收率）列于下表：

方法A	新法		旧法	
化合物B	甲化合物	乙化合物	甲化合物	乙化合物
数据	52	84	52	47
	48	88	44	64
	44	90	40	52
	44	80	26	45
合计	188	342	162	208

$$
\begin{bmatrix} y_{111} \\ y_{112} \\ y_{113} \\ y_{114} \\ y_{121} \\ y_{122} \\ y_{123} \\ y_{124} \\ y_{211} \\ y_{212} \\ y_{213} \\ y_{214} \\ y_{221} \\ y_{222} \\ y_{223} \\ y_{224} \end{bmatrix} = \begin{bmatrix} 52 \\ 48 \\ 44 \\ 44 \\ 84 \\ 88 \\ 90 \\ 80 \\ 52 \\ 44 \\ 40 \\ 26 \\ 47 \\ 64 \\ 52 \\ 45 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{113} \\ e_{114} \\ e_{121} \\ e_{122} \\ e_{123} \\ e_{124} \\ e_{211} \\ e_{212} \\ e_{213} \\ e_{214} \\ e_{221} \\ e_{222} \\ e_{223} \\ e_{224} \end{bmatrix}
$$

方法A	新法		旧法	
化合物B	甲	乙	甲	乙
数据	52	84	52	47
	48	88	44	64
	44	90	40	52
	44	80	26	45
合计	188	342	162	208

数据格式为：

Y	A	B
52	1	1
48	1	1
44	1	1
44	1	1
84	1	2
88	1	2
90	1	2
80	1	2
52	2	1
44	2	1
40	2	1
26	2	1
47	2	2
64	2	2
52	2	2
45	2	2

```
> x=read.table("clipboard",header=T)
> anova(lm(Y~A+B+A:B,data=x))
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq    F value    Pr(>F)
A          1   1600    1600.00    28.402    0.0001795 ***
B          1   2500    2500.00    44.379    2.321e-05 ***
A:B        1    729     729.00     12.941    0.0036638 **
Residuals 12    676   56.33
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P_A<0.05，说明不同方法对回收率有显著影响；

P_B<0.05，说明不同化合物对回收率有显著影响；

P_{AB}<0.05，说明方法和化合物之间交互作用对回收率有显著影响。

3.析因设计模型

三因素以上的析因分析略显复杂。

现以三因素为例：

假定A因素有I个水平，B因素有J个水平，C因素有M个水平。

实验中共有I*J*M个处理，每个处理重复r次。

模型为：

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_m + (\alpha\beta)_{ij} + (\alpha\gamma)_{im} + (\beta\gamma)_{jm} + (\alpha\beta\gamma)_{ijm} + e_{ijmk}$$

$$j = 1,2,\ldots,J, \qquad i = 1,2,\ldots,I, \qquad m = 1,2,\ldots,M, \qquad k = 1,2,\ldots,r$$

其中：$(\alpha\beta\gamma)_{ijm}$表示A的第i个水平与$(\beta\gamma)_{jm}$的交互效应，
或B的第j个水平与$(\alpha\gamma)_{im}$的交互效应，或C的第m个水平与$(\alpha\beta)_{ij}$的交互效应，
或 A的第i个水平与B的第j个水平以及C的第m个水平之间的二级交互作用。

4.正交实验设计模型

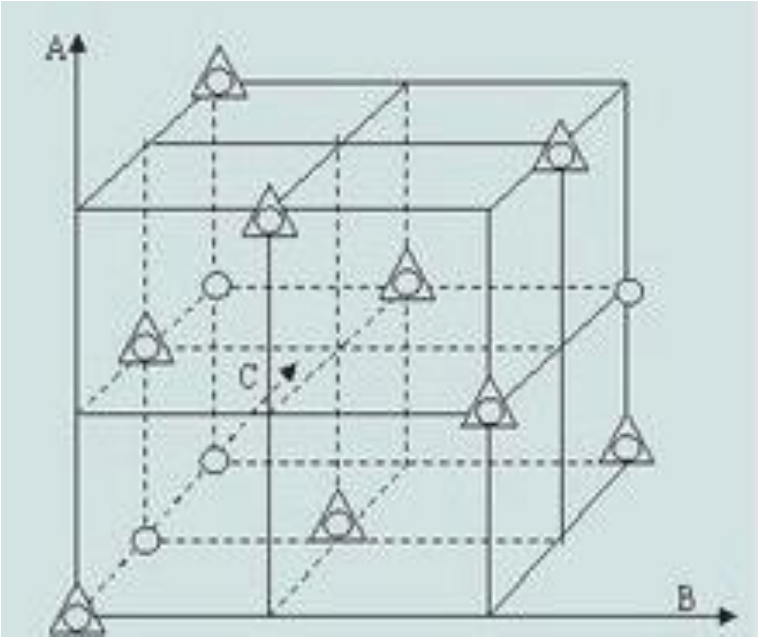
正交试验选择的水平组合列成表格，称为正交表

例如：三因素三水平的实验，3^3 = 27种组合的实验。

若按$L_9(3)$正交表安排实验，只需作9次

表 2　因子安排　试验方案

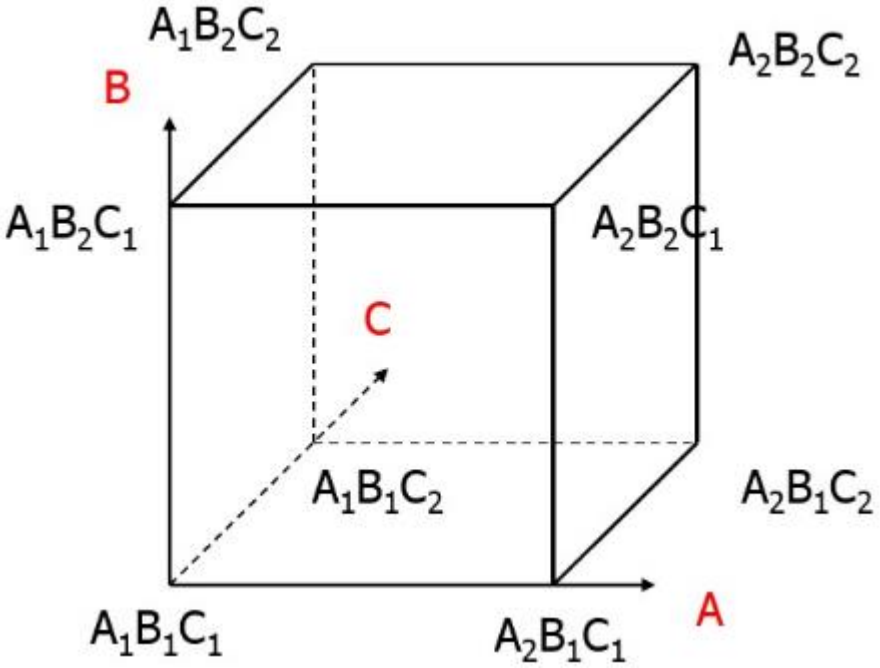
列号 / 行号	A	B	C			试验号	水平组合
	1	2	3	4			
1	1	1	1	1	⇨	1	$A_1B_1C_1$
2	1	2	2	2		2	$A_1B_2C_2$
3	1	3	3	3		3	$A_1B_3C_3$
4	2	1	2	3		4	$A_2B_1C_2$
5	2	2	3	1		5	$A_2B_2C_3$
6	2	3	1	2		6	$A_2B_3C_1$
7	3	1	3	2		7	$A_3B_1C_3$
8	3	2	1	3		8	$A_3B_2C_1$
9	3	3	2	1		9	$A_3B_3C_2$

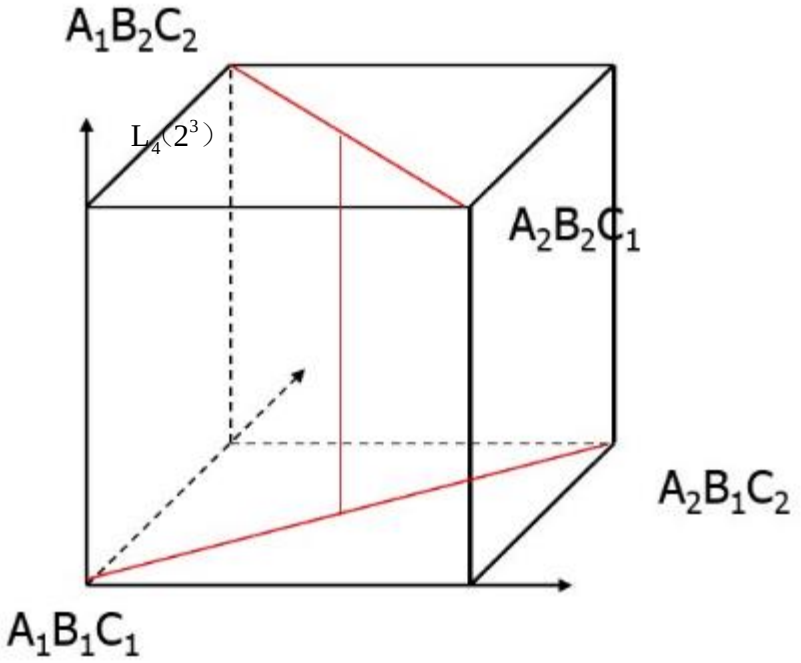


均衡搭配

$L_4(2^3)$

试验次数　　水平数　　因素数





例：对农药收率的因素分析

四个因素：A（反应温度）、B（反应时间）、C（原料配比）、D（真空度）；
每个因素有两个水平。A1、A2；B1、B2；C1、C2；D1、D2；
并考虑A、B的交互作用。选用正交表L_8（2^7）安排试验；
得出的结果为Y。

$$y_{ijmnr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_k + \gamma_m + \theta_n + e_{ijmkr}$$

列号	1	2	3	4	5	6	7	
表头	A	B	A*B	C			D	Y
1	1	1	1	1	1	1	1	86
2	1	1	1	2	2	2	2	95
3	1	2	2	1	1	2	2	91
4	1	2	2	2	2	1	1	94
5	2	1	2	1	2	1	2	91
6	2	1	2	2	1	2	1	96
7	2	2	1	1	2	2	1	83
8	2	2	1	2	1	1	2	88

所用的数据格式为：

A	B	C	D	Y
1	1	1	1	86
1	1	2	2	95
1	2	1	2	91
1	2	2	1	94
2	1	1	2	91
2	1	2	1	96
2	2	1	1	83
2	2	2	2	88

```
> x=read.table("clipboard",header=T)
> attach(x)
> anova(lm(Y~A+B+A*B+C+D),data=x)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq  F value  Pr(>F)
A          1     8.0     8.0      3.2  0.21554
B          1    18.0    18.0      7.2  0.11535
C          1    60.5    60.5     24.2  0.03893 *
D          1     4.5     4.5      1.8  0.31175
A:B        1    50.0    50.0     20.0  0.04654 *
Residuals  2     5.0     2.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P_A>0.05，说明反应温度A对农药的收率无显著影响；
P_B>0.05，说明反应时间B对农药的收率无显著影响；
P_C<0.05，说明原料配比C对农药的收率有显著影响；
P_D>0.05，说明真空度D对农药的收率有无著影响；
P_{AB}<0.05，说明反应温度A和反应时间B之间的交互作用对农药收率显著影响。