实验数据分析

主讲内容

一、Logistic回归分析

二、Probit回归分析

三、方差分析

四、响应面回归分析

主讲内容

Logistic回归分析

一、Logistic回归的理论基础

二、Logistic回归的参数估计

三、Logistic回归的假设检验

四、Logistic回归的预测

五、MATLAB案例分析

第一节　Logistic回归理论基础

“回归”一词的由来

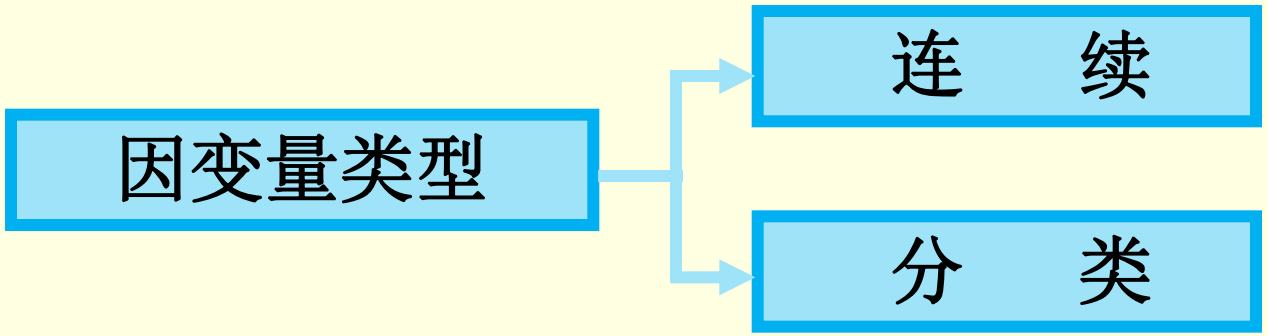
英国著名生物统计学家高尔顿**Galton**和他的学生在研究父代与子代身高之间的关系时发现：

$$y=0.8567+0.516x$$

父母平均身高	成年儿子身高
高一个单位	高半个单位
矮一个单位	矮半个单位

Galton把这种后代的身高向中间靠拢的趋势称为“回归现象”。后来，人们把由一个变量的变化去推测另一个变量的方法称为“回归方法”。

一、 Logistic回归的理论基础



- 研究某病疗效与不同治疗方法之间的关系；
- 研究考研与各科成绩之间的关系；
- 研究美国总统特郎普获胜与选民之间的关系。

一、 Logistic回归的理论基础

模型分类

y \ X	(0,1)	(0,1,2)	连续
(0,1)	例1,2	例3	例4,5
(0,1,2) 有序	例6		
(0,1,2) 无序	例7		

一、 Logistic回归的理论基础

准备：

1、当考研成功；害虫死亡；疾病发作；

产品是次品时，定义反应变量 $y=1$；

2、反之，定义反应变量 $y=0$；

3、记出现考研成功的概率为 $p(y=1)$，

显然 $0 \le p \le 1$。

一、 Logistic回归的理论基础

回忆：线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \qquad E(\varepsilon) = 0$$

故

$$\hat{y} = E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

而此时响应变量是分类变量，取值是0和1。

则

$$E(y) = P\{y = 1\} = p$$

所以

$$p = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

问题：

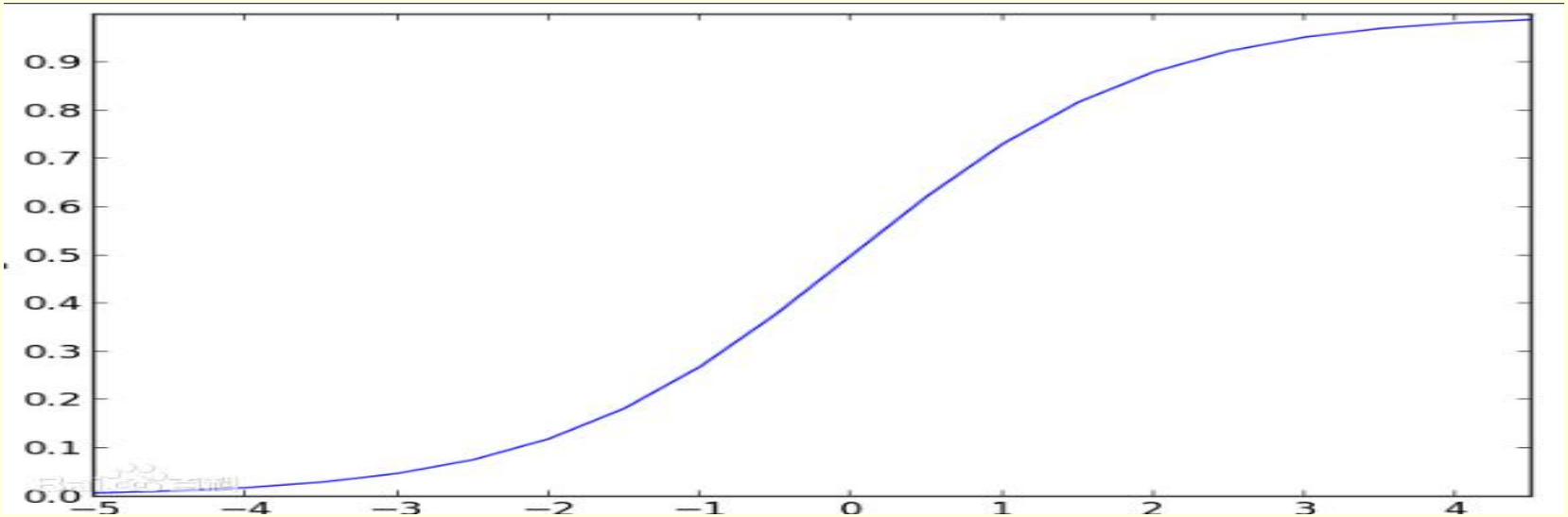
1、取值范围

模型右侧的取值范围是$(-\infty, +\infty)$，但模型左侧的取值范围却是(0, 1)。

一、 Logistic回归的理论基础

2、曲线关联

例：收入和购车的关系。当收入很低或很高时，收入的增加或减少对于买车的概率影响都很小，但当收入在某一临界值附近时，购车的概率会随着收入的增减而迅速增减。



一、 Logistic回归的理论基础

性质：

① $0 \le p \le 1$

②当 $x < c_1$ 和 $x > c_2$ 时$(c_2 > c_1)$，p 不会改变许多；

但当 $c_1 \le x \le c_2$ 时，p 会随着 x 的增大而增大。

一、 Logistic回归的理论基础

综上可知，可对 p 作如下变换：

$$\log it(P) = \ln \frac{p}{1-p}$$

可以验证

当 $0 \le p \le 1$ 时, $-\infty < \log it(P) < +\infty$

且 $\log it(P) = \ln \frac{p}{1-p}$ 是S形。

一、 Logistic回归的理论基础

故

$$\ln\frac{p}{1-p}=\beta_0+\beta_1x_1+\cdots+\beta_rx_r$$

简记为

$$\ln\frac{p}{1-p}=\beta_0+\sum_{k=1}^{r}\beta_kx_k$$

或

$$p=\frac{\exp(\beta_0+\sum_{k=1}^{r}\beta_kx_k)}{1+\exp(\beta_0+\sum_{k=1}^{r}\beta_kx_k)}$$

称上式为线性**Logistic**回归模型，简称**Logistic**模型。

二、 Logistic回归的参数估计

已知 y_i 的分布，即 $P\{y=y_i|X=x_i\}=p_i^{y_i}(1-p_i)^{1-y_i}$

故采用最大似然估计法估计 $\hat{\beta}$ 。

相应的似然函数 $L(\beta)=\prod_{i=1}^{n}P(y_i;\beta)=\prod_{i=1}^{n}p_i^{y_i}(1-p_i)^{1-y_i}$

两边取对数

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

$$=\sum_{i=1}^{n}[y_i\ln p_i+(1-y_i)\ln(1-p_i)]$$

$$=\sum_{i=1}^{n}[y_i\ln(\frac{p_i}{1-p_i})]+\sum_{i=1}^{n}\ln(1-p_i)]$$

二、Logistic回归的参数估计

将Logistic方程代入得

$$\ln L(\beta)=\sum_{i=1}^{n} y_i(\beta_0+\sum_{k=1}^{r}\beta_k x_{ik})-\sum_{i=1}^{n}\ln[1+\exp(\beta_0+\sum_{k=1}^{r}\beta_k x_{ik})]$$

想直接求解，较困难，我们采用Newton-Raphson迭代算法，从而求出参数的最大似然估计，得到。

$$p=\frac{\exp(\hat{\beta}_0+\sum_{k=1}^{r}\hat{\beta}_k x_k)}{1+\exp(\hat{\beta}_0+\sum_{k=1}^{r}\hat{\beta}_k x_k)}$$

二、Logistic回归的参数估计

参数的意义

1、优势（比数） $odds = \dfrac{p}{1-p}$

在一定条件下，发生的概率与不发生的概率的比值。

2、优势比（比数比）

$$OR = \frac{p_1 / (1-p_1)}{p_0 / (1-p_0)}$$

3、相对危险度

$$RR = \frac{p_1}{p_0}$$

当发生的概率很低时，OR≈RR

4、回归系数

$$\beta_0 = \ln \frac{p(y=1|X=0)}{1-p(y=1|X=0)}$$

$$\beta_1 = \ln \frac{p(y=1|X=1)}{1-p(y=1|X=1)} - \ln \frac{p(y=1|X=0)}{1-p(y=1|X=0)} = \ln OR$$

二、Logistic回归的参数估计

例：研究不同治疗方法对某病疗效的影响

治疗组别	有效（effect=1）	无效（effect=0）	合计
传统组（treat=1）	16	48	64
新法组（treat=2）	40	20	60
合计	56	68	124

$$odd_1 = \frac{p_1}{1-p_1} = \frac{16/64}{1-16/64} = \frac{16}{48} = \frac{1}{3} \qquad odd_2 = \frac{p_2}{1-p_2} = \frac{40/60}{1-40/60} = \frac{40}{20} = 2$$

$$OR = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{2}{1/3} = 6$$

表示新法组有效的治疗是传统组的6倍

二、Logistic回归的参数估计

例：研究性别对某病发作的影响

治疗组别	发病（happy=1）	健康（happy=0）	合计
男（sex=0）	16	48	64
女（sex=1）	40	20	60
合计	56	68	124

$$odd_1 = \frac{p_1}{1-p_1} = \frac{16/64}{1-16/64} = \frac{16}{48} = \frac{1}{3} \qquad odd_2 = \frac{p_2}{1-p_2} = \frac{40/60}{1-40/60} = \frac{40}{20} = 2$$

$$OR = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}} = \frac{2}{1/3} = 6$$

表示女性某病的发作是男性的6倍

三、 Logistic回归的假设检验

(1) 模型成立与否的统计检验

$H_0:\beta_1=\beta_2=\cdots=\beta_r=0 \leftrightarrow H_1:\beta_k$ 至少有一个不为0

检验统计量

$$AIC=-2\ln L(\hat{\beta})+2r$$

$$SC=-2\ln L(\hat{\beta})+r\ln(n)=SIC$$

$$-2LOG\ L=-2\ln L(\hat{\beta})$$

说明：上述三个值越小，则模型拟合的越好。

三、 Logistic回归的假设检验

$$Likelihood\ Ratio = \frac{L(\hat{\beta})}{L(0)} \underset{a}{\overset{H_0}{\sim}} \chi^2(r)$$

$$Score = 2\ln\frac{L(\hat{\beta})}{L(0)} \underset{a}{\overset{H_0}{\sim}} \chi^2(r)$$

$$Wald = \sum_{k=1}^{r} \frac{\hat{\beta}_k^2}{SE_{\hat{\beta}_k}^2} \overset{H_0}{\sim} \chi^2(r)$$

说明： 上述三个值越大，则模型拟合的越好。

三、 Logistic回归的假设检验

(2)某个自变量是否显著的检验

$$H_0 : \beta_k = 0 \leftrightarrow H_1 : \beta_k \neq 0$$

检验统计量

$$Wald\,\chi^2 = \left(\frac{\hat{\beta}_k - 0}{SE_{\hat{\beta}_k}}\right)^2 \overset{H_0}{\sim} \chi^2(1)$$

类似于线性回归系数的t检验

$$U = \frac{\hat{\beta}_k}{SE_{\hat{\beta}_k}} \overset{H_0}{\sim} N(0,1)$$

为了输出结果时OR的置信区间

给定显著水平，拒绝域为 $W = \{Wald\,\chi^2 > \chi^2_\alpha(1)\}$

给定显著水平，拒绝域为 $W = \{|U| > z_{\alpha/2}\}$

四、 Logistic回归的预测

给出预测点 $(x_{01}, x_{02}, \cdots, x_{0r})$

根据

$$p = \frac{\exp(\hat{\beta}_0 + \sum_{k=1}^{r} \hat{\beta}_k x_k)}{1 + \exp(\hat{\beta}_0 + \sum_{k=1}^{r} \hat{\beta}_k x_k)}$$

计算出 p，若 $p > 0.5$，我们认为 $\hat{y} = 1$；

若 $p < 0.5$，我们认为是 $\hat{y} = 0$。

五、MATLAB案例分析

5.1、 MATLAB案例分析一

5.2、 MATLAB案例分析二

5.3、 MATLAB案例分析三

5.4、 MATLAB案例分析四

MATLAB实现logistic回归

Matlab 实现：实现logistic回归不需要编制函数程序，它自身提供了内部的功能函数

一元logistic回归——fitglm

多元logistic回归——mnrfit

MATLAB一元logistic回归——fitglm函数

```
b0 =fitglm(X,Y,'constant','Distribution','binomial','link','logit')
```

返回的回归模型，包含各种参数

输入的自变量x

x的响应变量y

模型类型

'constant'：只包含一个常数项；
'linear'　：常数项和线性项；
'interactions'：常数项、线性项、乘积项
'purequadratic'：常数项、线性项、平方项
'quadratic'：常数项、线性项、平方项、乘积项

MATLAB一元logistic回归——fitglm函数

```
b0 =fitglm(X,Y,'constant','Distribution','binomial','link','logit')
```

响应分布类型的文本名称

分布类型

'normal'：正态分布（默认）
'binomial'：二项分布
'poisson'：泊松分布
'gamma'：伽马分布
'inverse gaussian'：反高斯分布

MATLAB一元logistic回归——fitglm函数

名称-值对

```
b0 =fitglm(X,Y,'constant','Distribution','binomial','link','logit'
)
```

link函数定义了平均响应与预测因子x*b的线性组合之间的关系f(mu)=x*b

link函数的文本名称

Link函数类型

```
'identity': f(mu) = mu
    'log'   : f(mu) = log(mu)
   'logit'  : f(mu) = log(mu/(1-mu))
 'probit'   : f(mu) = norminv(mu)
'comploglog': f(mu) = log(-log(1-mu))
```

MATLAB多元logistic回归——mnrfit函数

```
[B,dev,stats]=mnrfit(X,Y,'model','ordinal','link','logit');
```

返回的系数矩阵

返回的回归模型，包含模型统计信息,如自由度、系数估计的标准误差和残差

输入的自变量X

响应变量Y

MATLAB多元logistic回归——mnrfit函数

```
[B,dev,stats]=mnrfit(X,Y,'model','ordinal','link','logit');
```

文本字符串
“模型”

link函数
的文本名称

Link函数
类型

要匹配的模型类型:
'ordinal'（默认）
'hierarchical'

```
'identity': f(mu) = mu
   'log'  : f(mu) = log(mu)
  'logit' : f(mu) = log(mu/(1-mu))
 'probit' : f(mu) = norminv(mu)
'comploglog': f(mu) = log(-log(1-mu))
```

MATLAB实现logistic回归的基本流程和思路

1、分别建立只含常数项和线性项的logistic回归模型
2、模型的检验
   2.1 信息测量指标
      a、AIC
      b、BIC(SC)
      c、-2倍对数似然
   2.2 回归系数的显著性检验
      a、wald检验
      b、似然比检验
   2.3 模型的拟合优度检验
      a、Pearson检验
      b、deviance检验
3、模型的评价结果
4、模型的结果和结论

5.1、 MATLAB案例分析一

例1：研究不同治疗方法对某病疗效的影响

治疗组别	有效（effect=1）	无效（effect=0）	合计
传统组（treat=1）	16	48	64
新法组（treat=2）	40	20	60
合计	56	68	124

1、分别建立只含常数项与线性项的logistic回归模型

```
clc
clear
treat=[1 2]';
effect=[16 40]’;
n=[64 60]’;
%%只含常数项的logsitc模型拟合
b0 = fitglm(treat,[effect n],'constant','Distri
bution','binomial','link','logit’)
%%含线性项的logsitc模型拟合
b1= fitglm(treat,[effect n],'linear','Distribut
ion','binomial','link','logit')
```

2、模型的检验1——信息测量指标

```
%只含常数项的信息测量指标
AIC0=b0.ModelCriterion.AIC;
BIC0=b0.ModelCriterion.BIC;
LL0=-2*b0.LogLikelihood

%含线性项的信息测量指标
AIC1=b1.ModelCriterion.AIC;
BIC1=b1.ModelCriterion.BIC;
LL1=-2*b1.LogLikelihood;
```

3、模型的检验2——回归系数的显著性检验

```
%%似然比检验
LR=LL0-LL1;
p2=1-chi2cdf(LR, 1);

%Wald检验
wald=b1.Coefficients.tStat(2)^2;
p3=1-chi2cdf(wald, 1);
```

4、模型的评价结果1

```
%根据信息测量指标评价模型 （打印模型拟合检验表）
fprintf('\n------------模型拟合检验表----------\n')
;
fprintf('%3s%10s%10s\n','准则','只含常数项','含治疗
方法');
fprintf('%5s%14.4f%12.4f\n','AIC',AIC0,AIC1);
fprintf('%5s%14.4f%12.4f\n','BIC',BIC0,BIC1);
fprintf('%5s%13.4f%12.4f\n','-2logL',LL0,LL1);
```

4、模型的评价结果2

```
%根据回归系数的显著性检验评价模型(打印显著性检验表)
fprintf('\n--------------似然比和Wald检验表----------\n');
fprintf('%3s%10s%10s%9s\n','检验','卡方','自由度','p值');
fprintf('%3s%14.4f%8.0f%13.4f\n','似然比',LR,1,p2);
fprintf('%5s%15.4f%8.0f%13.4f\n','Wald',wald,1,p3);
```

5、模型的结果

```
%% 参数估计结果（参数的值）
fprintf('\n-参数估计结果-\n');
fprintf('%3s%10s%10s%10s%10s%13s\n','参数','自由度',
'估计值','标准误差','t值','p值');
fprintf('%3s%10.0f%14.4f%15.4f%13.4f%12.4f\n','常数
项',1,b1.Coefficients.Estimate(1),b1.Coefficients.
SE(1),b1.Coefficients.tStat(1),b1.Coefficients.pVa
lue(1));
fprintf('%3s%11.0f%14.4f%15.4f%13.4f%12.4f\n','tre
at',1,b1.Coefficients.Estimate(2),b1.Coefficients.
SE(2),b1.Coefficients.tStat(2),b1.Coefficients.pVa
lue(2));
```

6、模型的结论

```
%% 问题的结论（新法组是传统组疗效的多少倍）
fprintf('\n----OR点估计和区间估计----\n');
fprintf('%3s%10s%17s\n','效应','点估计','置信区间')
;
fprintf('%3s%13.4f%10s%1.4f%1s%1.4f%1s\n','treat
',yf1,'[ ',lower1,' , ',upper1,' ]');
```

5.2、 MATLAB案例分析二

例2：研究性别及疾病严重程度对某病疗效的影响

性别	疾病程度	有效（effect=1）	无效（effect=0）	合计
女 sex=0	不严重 degree=0	21	6	27
	严重 degree=1	9	9	18
男 sex=1	不严重 degree=0	8	10	18
	严重 degree=1	4	11	15

1、分别建立只含常数项与线性项的logistic回归模型

```
clc,clear
sex=[0 0 1 1 ]';
degree=[0 1 0 1 ]';
effect=[21 9 8 4]';
n=[27 18 18 15]';
%%只含常数项的logsitc模型拟合
b0 = fitglm([sex degree],[effect n],'constant',
'Distribution','binomial','link','logit');
%%含线性项的logsitc模型拟合
b1= fitglm(treat,[effect n],'linear','Distribut
ion','binomial','link','logit')
```

2、模型的检验1——信息测量指标

%只含常数项的信息测量指标

```
AIC0=b0.ModelCriterion.AIC;
BIC0=b0.ModelCriterion.BIC;
LL0=-2*b0.LogLikelihood
```

%含线性项的信息测量指标

```
AIC1=b1.ModelCriterion.AIC;
BIC1=b1.ModelCriterion.BIC;
LL1=-2*b1.LogLikelihood;
```

3、模型的检验2——回归系数的显著性

```
%%似然比检验
LR=LL0-LL1;
p2=1-chi2cdf(LR, 1);


%Wald检验
wald=b1.Coefficients.tStat(2)^2;
p3=1-chi2cdf(wald, 1);
```

3、模型的检验3—模型的拟合优度

```
%deviance检验
deviance=sum((b1.Residuals.Deviance).^2);
p1=1-chi2cdf(deviance, b1.DFE);

%pearson检验
pearson=sum((b1.Residuals.Pearson).^2);
p=1-chi2cdf(pearson, b1.DFE);
```

4、模型的评价结果1—信息测量指标

```
%根据信息测量指标评价模型 （打印模型拟合检验表）
fprintf('\n----------模型拟合检验表----------\n');
fprintf('%3s%10s%10s\n','准则','只含常数项','含性别和程度');
fprintf('%5s%14.4f%12.4f\n','AIC',AIC0,AIC1);
fprintf('%5s%14.4f%12.4f\n','BIC',BIC0,BIC1);
fprintf('%5s%13.4f%12.4f\n','-2logL',LL0,LL1);
```

4、模型的评价结果2—回归系数的显著性

```
%根据回归系数的显著性检验评价模型(打印显著性检验表)
fprintf('\n-----似然比和Wald检验表------\n');
fprintf('%3s%10s%10s%9s\n','检验','卡方','自由度','p值');
fprintf('%3s%14.4f%8.0f%13.4f\n','似然比',LR,2,p2);
fprintf('%5s%15.4f%8.0f%13.4f\n','Wald',wald,2,p3);
```

4、模型的评价结果3—模型的拟合优度

```
%根据模型的拟合优度评价模型(打印显著性检验表)
fprintf('\n--模型拟合优度检验表--\n');
fprintf('%3s%10s%10s%10s%8s\n','准则','值','自由度','值/自由度','p值');
fprintf('%3s%13.4f%8.0f%15.4f%12.4f\n','偏差',deviance,b1.DFE,deviance/b1.DFE,p1);
fprintf('%3s%11.4f%8.0f%15.4f%12.4f\n','Pearson',pearson,b1.DFE,pearson/b1.DFE,p);
```

5、模型的结果

%% 参数估计结果（参数的值）

同实验一

6、模型的结论

%% 问题的结论（通过OR点估计）

```
fprintf('\n---OR点估计和区间估计---\n');
fprintf('%3s%10s%17s\n','效应','点估计','置信区间'
);
fprintf('%3s%13.4f%10s%1.4f%1s%1.4f%1s\n','性别'
,yf1,'[ ',lower1,' , ',upper1,' ]');
fprintf('%3s%13.4f%10s%1.4f%1s%1.4f%1s\n','程度'
,yf2,'[ ',lower2,' , ',upper2,' ]');
```

5.3、MATLAB案例分析三

例3：研究性别和不同疗法对某病治愈的影响

性别	治疗方法（treat）	有效（response=cured）	无效（response=not）	合计
男性	A	78	28	106
	B	101	11	112
Sex=m	C	68	46	114
女性	A	40	5	54
	B	54	5	59
Sex=f	C	34	6	40

1、分别建立只含常数项与线性项的logistic回归模型

```
sex=[1 1 1 0 0 0]';
treata=[1 0 0 1 0 0]’;
treatb=[0 1 0 0 1 0]’;
cured=[78 101 68 40 54 34]’;
n=[106 112 114 45 59 40]’;
%%只含常数项的logsitc模型拟合
b0 = fitglm([sex treata treatb],[cured n],'cons
tant','Distribution','binomial','link','logit’)
;
%%含线性项的logsitc模型拟合
b1= fitglm([sex treata treatb],[cured n],'linea
r','Distribution','binomial','link','logit');
```

2、模型的检验1——信息测量指标

%只含常数项的信息测量指标

```
AIC0=b0.ModelCriterion.AIC;
BIC0=b0.ModelCriterion.BIC;
LL0=-2*b0.LogLikelihood
```

%含线性项的信息测量指标

```
AIC1=b1.ModelCriterion.AIC;
BIC1=b1.ModelCriterion.BIC;
LL1=-2*b1.LogLikelihood;
```

3、模型的检验2——回归系数的显著性

```
%%似然比检验
LR=LL0-LL1;
p2=1-chi2cdf(LR, 1);


%Wald检验
wald=b1.Coefficients.tStat(2)^2;
p3=1-chi2cdf(wald, 1);
```

3、模型的检验3—模型的拟合优度

```
%deviance检验
deviance=sum((b1.Residuals.Deviance).^2);
p1=1-chi2cdf(deviance, b1.DFE);

%pearson检验
pearson=sum((b1.Residuals.Pearson).^2);
p=1-chi2cdf(pearson, b1.DFE);
```

4、模型的评价结果1—信息测量指标

```
%根据信息测量指标评价模型 （打印模型拟合检验表）
fprintf('\n----------模型拟合检验表----------\n');
fprintf('%3s%10s%10s\n','准则','只含常数项','含性别和程度');
fprintf('%5s%14.4f%12.4f\n','AIC',AIC0,AIC1);
fprintf('%5s%14.4f%12.4f\n','BIC',BIC0,BIC1);
fprintf('%5s%13.4f%12.4f\n','-2logL',LL0,LL1);
```

4、模型的评价结果2—回归系数的显著性

```
%根据回归系数的显著性检验评价模型(打印显著性检验表)
fprintf('\n-----似然比和Wald检验表------\n');
fprintf('%3s%10s%10s%9s\n','检验','卡方','自由度','p值');
fprintf('%3s%14.4f%8.0f%13.4f\n','似然比',LR,2,p2);
fprintf('%5s%15.4f%8.0f%13.4f\n','Wald',wald,2,p3);
```

4、模型的评价结果3—模型的拟合优度

```
%根据模型的拟合优度评价模型(打印显著性检验表)
fprintf('\n--模型拟合优度检验表---\n');
fprintf('%3s%10s%10s%10s%8s\n','准则','值','自由度','值/自由度','p值');
fprintf('%3s%13.4f%8.0f%15.4f%12.4f\n','偏差',deviance,b1.DFE,deviance/b1.DFE,p1);
fprintf('%3s%11.4f%8.0f%15.4f%12.4f\n','Pearson',pearson,b1.DFE,pearson/b1.DFE,p);
```

5、模型的结果

%% 参数估计结果（参数的值）

同实验一

6、模型的结论

%% 问题的结论（通过OR点估计得出结论）

```
fprintf('\n---OR点估计和区间估计---\n');
fprintf('\n--OR点估计和区间估计--\n');
fprintf('%3s%10s%17s\n','效应','点估计','置信区间'
);
fprintf('%3s%13.4f%10s%1.4f%1s%1.4f%1s\n','性别'
,yf1,'[ ',lower1,' , ',upper1,' ]');
fprintf('%3s%13.4f%10s%1.4f%1s%1.4f%1s\n','程度'
,yf2,'[ ',lower2,' , ',upper2,' ]');
```

5.4、 MATLAB案例分析四

1、多分类有序因变量的logistic回归

例6：研究性别及疾病严重程度对某病疗效的影响

性别	治疗方法	显效 marked	有效 some	无效 none	合计
女 sex=1	新药疗法 treat=1	16	5	6	27
	传统疗法 treat=0	6	7	19	32
男 sex=0	新药疗法 treat=1	5	2	7	14
	传统疗法 treat=0	1	0	10	11

其他类型的Logistic回归

回归方程

$$\ln(p_1)=\ln\frac{p_1}{1-p_1}=\alpha_1+\beta_1x_1+\cdots+\beta_px_p$$

$$\ln(p_1+p_2)=\ln\frac{p_1+p_2}{1-(p_1+p_2)}=\alpha_2+\beta_1x_1+\cdots+\beta_px_p$$

$$\ln(p_1+p_2+p_3)=\ln\frac{p_1+p_2+p_3}{1-(p_1+p_2+p_3)}=\alpha_3+\beta_1x_1+\cdots+\beta_px_p$$

$$\ln(\sum_{i=1}^{k}p_i)=\ln\frac{\sum_{i=1}^{k}p_i}{\sum_{i=k+1}^{r}p_i}=\alpha_k+\beta_1x_1+\cdots+\beta_px_p$$

说明：传统的二分类反应变量相比，多分类有序因变量有序得到的是取值水平的累积概率。

1、建立多元logistic回归模型

```
clc,clear
sex=[1 1 0 0]';
treat=[1 0 1 0]';
X=[sex treat];
marked=[16 6 5 1]';
some=[5 7 2 0]';
n=[6 19 7 10]';
Y=[marked some n];
[B,dev,stats]=mnrfit(X,Y,'model','ordinal',
'link','logit');
```

2、模型的结果和检验

```
%% 参数估计结果，结果中的p值进行检验
fprintf('\n--参数估计结果--\n');
fprintf('%3s%12s%9s%13s%10s%11s\n','参数','自由度','估计值','
标准误差','t值','p值');
fprintf('%3s%10.0f%15.4f%15.4f%13.4f%12.4f\n','常数项
1',1,stats.beta(1),stats.se(1),stats.t(1),stats.p(1));

fprintf('%3s%10.0f%15.4f%15.4f%13.4f%12.4f\n','常数项
2',1,stats.beta(2),stats.se(2),stats.t(2),stats.p(2));

fprintf('%6s%11.0f%15.4f%15.4f%13.4f%12.4f\n','sex',1,stat
s.beta(3),stats.se(3),stats.t(2),stats.p(3));

fprintf('%6s%11.0f%15.4f%15.4f%13.4f%12.4f\n','treat',1,st
ats.beta(4),stats.se(4),stats.t(3),stats.p(4));
```

3、模型的结论

```
%结论（通过OR点估计）
yf1=exp(B(3,:));   %sex
lower1=exp(B(3,:)-norminv(0.975,0,1)*stats.se(3
,:));
upper1=exp(B(3,:)+norminv(0.975,0,1)*stats.se(3
,:));
yf2=exp(B(4,:));   %degree
lower2=exp(B(4,:)-norminv(0.975,0,1)*stats.se(4
,:));
upper2=exp(B(4,:)+norminv(0.975,0,1)*stats.se(4
,:));
```

四、 其他类型的Logistic回归

2、多分类无序因变量的logistic回归

例7、 研究多个因素对于获得健康知识是途径的影响

社区	性别	获取健康知识的途径		
		大众媒体	网络	社区教育
1	男性male 女性female	20 10	35 27	26 57
2	男性male 女性female	42 16	17 12	26 26
3	男性male 女性female	15 11	15 12	16 20

四、其他类型的Logistic回归

回归方程

$$\ln\frac{p_1}{p_4}=\alpha_1+\beta_{11}x_1+\cdots+\beta_{1p}x_p$$

$$\ln\frac{p_2}{p_4}=\alpha_2+\beta_{21}x_1+\cdots+\beta_{2p}x_p$$

$$\ln\frac{p_3}{p_4}=\alpha_3+\beta_{31}x_1+\cdots+\beta_{3p}x_p$$

说明：模型会定义因变量的某一个水平为参照水平，其他水平与其相比。

四、 其他类型的Logistic回归

3、 配对logistic回归

适用于配对方法收集到的资料，又称条件 Logistic 回归模型。例如临床药物试验中，让一个病人选择药物A，另一个病人选择药物B.....，然后考察病人的病情的好转情况时，从而形成一个匹配。有1:1匹配、1:m匹配和n:m匹配。

第二节　Probit回归分析简介

一、Probit回归模型

二、Probit回归模型的参数估计

三、Probit回归模型的检验

四、logistic模型与Probit模型的对比

五、案例分析

一、Probit回归模型

- 统计学主要回归模型是最小二乘模型（线性模型）。最小二乘模型假设因变量与自变量之间具有线性关系，并通过最小化残差项对参数进行估计，有广泛应用，但也有局限。
- **Probit**模型是一个处理分类因变量的概率模型，是一个非线性回归模型，是现代统计学中应用最为广泛的模型之一。

一、Probit回归模型

假设Y是一个二值的因变量，取值为0或1，X为自变量，它可以是一个标量或向量。令

$$p_i = P\{y_i = 1 \mid X = x_i\}$$

定义一个连续的被解释变量Y*，使得

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,1)$$

且Y与Y*之间存在以下对应关系，

$$Y = \begin{cases} 1, & Y^* \geq 0, \\ 0, & Y^* < 0, \end{cases}$$

一、Probit回归模型

则有 $$p_i = P\{y_i = 1 \mid X = x_i\} = P\{\beta_0 + \beta_1 x_i + \varepsilon_i > 0\}$$

$$= P\{\varepsilon_i > -\beta_0 - \beta_1 x_i\} = P\{\varepsilon_i < \beta_0 + \beta_1 x_i\}$$

$$= \Phi(\beta_0 + \beta_1 x_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_0 + \beta_1 x_i} e^{-\frac{t^2}{2}} dt,$$

上式转化为线性模型，得到**Probit**回归模型

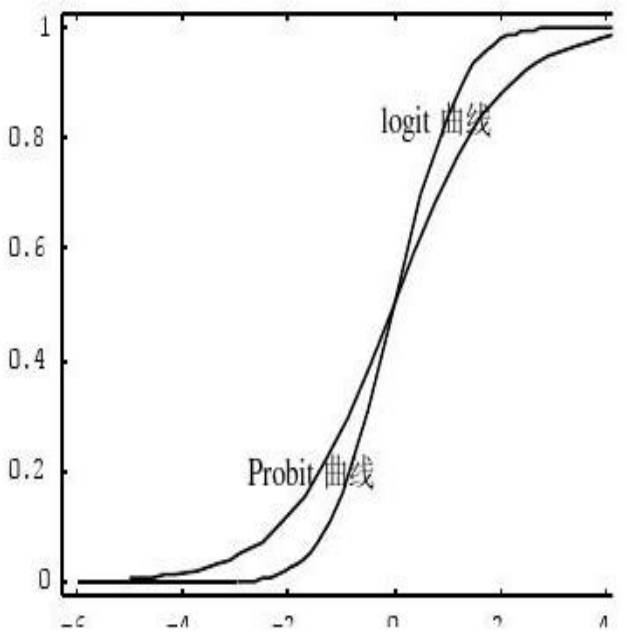
$$\Pr obit(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_i$$

多元情形：

$$\Pr obit(p_i) = \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

两个回归模型的对比

y_i	$p_i = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y_i} e^{-\frac{t^2}{2}} dt$	$p_i = \frac{1}{1+e^{-y_i}}$
-3.0	0.0013	0.0474
-2.0	0.0228	0.1192
-1.5	0.0668	0.1824
-1.0	0.1587	0.2689
-0.5	0.3085	0.3775
0.0	0.5000	0.5000
0.5	0.6915	0.6225
1.0	0.8413	0.7311
1.5	0.9332	0.8176
2.0	0.9772	0.8808
3.0	0.9987	0.9526



Logit转换与Probit转换的对比图

二、Probit回归模型的参数估计

- 最大似然法估计，其具体过程与**logistic**回归参数估计；
- 非线性模型，从符号上判断自变量的增加导致因变量出现与否的概率增减。

三、Probit回归模型的检验

1.皮尔逊检验

假设

H_0 协变量类型中的实际观测值的与预测值没有差异

H_1 协变量类型中的实际观测值的与预测值有显著差异

检验统计量

$$\chi^2 = \sum_{i=1}^{n} \frac{(residuals_i)^2}{n\hat{p}_i(1-\hat{p}_i)}$$

注：$p > 0.05$ 时，接受原假设。

三、Probit回归模型的检验

2.似然比检验

假设

$$H_0:\beta_1=\beta_2=\cdots=\beta_r=0 \leftrightarrow H_1:\beta_1,\beta_2,\cdots,\beta_r\text{不全为0}$$

定义似然比统计量

$$LR=-2[\ln L(0)-\ln L(\hat{\beta})]\overset{H_0}{\sim}\chi^2(r)$$

判别规则：若 $LR\le\chi_\alpha^2(r)$，则接受原假设，认为约束条件成立；否则，拒绝原假设，认为约束条件成立。

四、logistic与Probit模型的对比

- 当自变量中分类变量较多，用logistic回归，回归的系数可以得到很好解释；
- 当自变量中连续性变量较多且服从正态分布是，用Probit回归，但回归系数解释较麻烦；
- 实际应用中，两个回归结果接近，均用于解释事件发生的概率，且logistic模型系数约为Probit模型系数的1.6倍。

五、案例分析

例2.1.2：研究性别及疾病严重程度对某病疗效的影响

性别	疾病程度	有效（effect=1）	无效（effect=0）	合计
女 sex=0	不严重 degree=0	21	6	27
	严重 degree=1	9	9	18
男 sex=1	不严重 degree=0	8	10	18
	严重 degree=1	4	11	15

五、案例分析

1.拟合优度检验

表 2.2.1 模型拟合优度检验

准则	值	自由度	值/自由度	p 值
偏差	0.2150	1	0.2150	0.6429
Person	0.2162	1	0.2162	0.6419

表 2.2.2 模型拟合检验表

准则	只含常数项	含性别和程度
AIC	27.0784	19.3099
BIC	26.4647	17.4688
$-2\log L$	25.0784	13.3099

表 2.2.3 似然比和 Wald 检验表

检验	χ^2	自由度	p 值
似然比	11.7685	2	0.0028
Wald	11.3259	2	0.0035

表2.2.1 P值均大于0.05，通过了拟合优度检验

表2.2.3 似然比检验值小于0.01，整体拟合好。

五、案例分析

2.参数估计

表 2.2.4 基于最大似然法的参数估计结果

参数	自由度	估计值	标准误差	t 值	p 值
常数项	1	0.7102	0.2384	2.9790	0.0029
sex	1	-0.7830	0.3012	-2.5998	0.0093
degree	1	-0.6436	0.3012	-2.1371	0.0326

在参数检验中，各参数对应的p值均小于0.05，且Probit回归模型表达式为：

$$\Pr obit(p) = \Phi^{-1}(p) = 0.7102 - 0.7830 * sex - 0.6436 * degree$$

疾病治疗有效的概率的计算公式：

$$p = P\{y = 1\} = \Phi(0.7102 - 0.7830 * sex - 0.6436 * degree)$$

第三节 方差分析

实际中，某个试验指标的取值，往往与多个因素有关。

- 农作物的产量可能与作物品种、施肥量、土壤条件、苗间距等因素有关；

- 化工产品的转化率可能与原料配方、催化剂用量、反应温度、加热时间等因素有关。

第三节 方差分析

例：为比较温度对着色度的影响，选了4种不同温度在其他外界条件都一样，每种温度各在四块布料上进行试验，得着色度如下：

温度A	着色度
A_1	0.981，0.964，0.917，0.669
A_2	0.607，0.693，0.506，0.358
A_3	0.791，0.642，0.810，0.705
A_4	0.901，0.703，0.792，0.883

实验指标

实验因素

因素水平

单因素
方差分析

问温度对着色度有无显著性影响。

第三节 方差分析

例：为比较温度和硫酸对着色度的影响，选了4种不同温度，3种浓度的硫酸，在其他外界条件都一样，每种组合各在两块布料上进行试验，得着色度如下。

温度 \ 硫酸	B_1	B_2	B_3
A_1	0.693，0.506	0.607，0.358	0.810，0.705
A_2	0.810，0.705	0.981，0.964	0.792，0.883
A_3	0.791，0.642	0.810，0.705	0.843，0.766
A_4	0.917，0.669	0.657，0.703	0.901，0.703

实验指标

实验因素

因素水平

双因素方差分析

问温度、硫酸种类及交互作用对着色度有无影响。

第三节 方差分析

问题: 1) 影响试验指标的诸多因素中，哪些因素的影响显著，哪些影响不显著？

2) 有显著影响的因素在何种水平时对指标的影响效果最佳？

方差分析：解决上述问题的有效的统计方法。

方法：利用数据自身的差异。-----方差

第三节 方差分析

● 单因素方差分析

问题： 已知因子 A，其有 r 种水平 $A_1, A_2, \cdots, A_r$，在每一种水平下做了 $n_i (i = 1,2,\cdots,r)$ 次试验，且 $\sum_{i=1}^{r} n_i = n$，设在水平 A_i 下的试验值 $X_i \sim N(\mu_i, \sigma^2)$ 且所有 X_i 之间相互独立。

因素	观测值				组均值
A_1	x_{11}	x_{12}	$\cdots$	x_{1n_1}	$\bar{x}_{1\cdot}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2n_2}	$\bar{x}_{2\cdot}$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	x_{r1}	x_{r2}	$\cdots$	x_{rn_r}	$\bar{x}_{r\cdot}$
总体均值	μ_1	μ_2	$\cdots$	μ_r	$\bar{x}$

试分析因子 A 对试验指标有无显著影响？

第三节 方差分析

1 数学模型

$$\begin{cases} x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \sum_{i=1}^{r} \alpha_i = 0 \\ \varepsilon_{ij} \sim N(0, \sigma^2) \quad i = 1, \cdots r; j = 1, \cdots n_i \end{cases}$$

假设 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r \leftrightarrow H_1 : \mu_i$ 不全相等；

即 $H_0 : \delta_1 = \delta_2 = \cdots = \delta_r = 0 \leftrightarrow H_1 :$ 至少有一个 $\delta_i \neq 0$

第三节 方差分析

2 统计检验

(1) 平方和分解

$$S_T = S_A + S_E$$

(2) 检验统计量

$$F = \frac{\frac{Q_A}{\sigma^2}/(r-1)}{\frac{Q_E}{\sigma^2}/(n-r)} = \frac{Q_A/(r-1)}{Q_E/(n-r)} \overset{\Delta}{=} \frac{S_A^2}{S_E^2} \overset{H_0}{\sim} F(r-1, n-r)$$

第三节 方差分析

3 单因素方差分析表

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

第三节 方差分析

● 双因素方差分析

在实际应用中，一个试验结果（试验指标）往往受多个因素的影响。不仅这些因素会影响试验结果，而且这些因素的不同水平的搭配也会影响试验结果。统计学上把多因素不同水平搭配对试验指标的影响称为交互作用。

例如：某些合金，当单独加入元素A或元素B时，性能变化不大，但当同时加入元素A和B时，合金性能的变化就特别显著。

任务：分别影响、交互影响

第三节 方差分析

设因素A有r个水平A_1，A_2，…，A_r，B有s个水平B_1，B_2，….B_s，则A与B的不同水平组合A_iB_j（i=1，2，…，r；j=1，2，…，s）共有rs个在水平组合，每个水平组合称为一个处理，每个处理下进行t次独立试验，得到rst个样本观测值X_{ijk}（k=1，2，…，t）：

因素	B_1	B_2		B_s
A_1	$x_{111}\cdots x_{11t}$	$x_{121}\cdots x_{12t}$	…	$x_{1s1}\cdots x_{1st}$
A_2	$x_{211}\cdots x_{21t}$	$x_{221}\cdots x_{22t}$	…	$x_{2s1}\cdots x_{2st}$
	⋮	⋮		⋮
A_r	$x_{r11}\cdots x_{r1t}$	$x_{r21}\cdots x_{r2t}$	…	$x_{rs1}\cdots x_{rst}$

第三节 方差分析

(1) 数学模型

$$\begin{cases} x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ \sum_{i=1}^{r} \alpha_i = 0,\ \sum_{j=1}^{s} \beta_j = 0,\ \sum_{i=1}^{r} \gamma_{ij} = \sum_{j=1}^{s} \gamma_{ij} = 0 \\ \varepsilon_{ijk} \sim N(0, \sigma^2) \quad i = 1, \cdots r;\ j = 1, \cdots s;\ k = 1, \cdots, t \end{cases}$$

(2) 统计检验

原假设和备择假设为以下三种形式：

$$H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0; \quad H_{11}: \alpha_1, \alpha_2, \cdots, \alpha_r \text{ 不全为零},$$

$$H_{02}: \beta_1 = \beta_2 = \cdots = \beta_s = 0; \quad H_{12}: \beta_1, \beta_2, \cdots, \beta_s \text{ 不全为零},$$

$$H_{03}: \gamma_{11} = \gamma_{12} = \cdots = \gamma_{rs} = 0; \quad H_{13}: \gamma_{11}, \gamma_{12}, \cdots, \gamma_{rs} \text{ 不全为零}。$$

第三节 方差分析

(3) 双因素方差分析表

$$S_T = S_E + S_A + S_B + S_{AB}$$

方差来源	平方和	自由度	均方	F 比
因素 A	S_A	$r-1$	$\bar{S}_A = \frac{S_A}{r-1}$	$F=\frac{\bar{S}_A}{\bar{S}_E}$
因素 B	S_B	$s-1$	$\bar{S}_B = \frac{S_B}{s-1}$	$F=\frac{\bar{S}_B}{\bar{S}_E}$
交互作用	S_{AB}	$(r-1)(s-1)$	$\bar{S}_{AB} = \frac{S_{AB}}{(r-1)(s-1)}$	$F=\frac{\bar{S}_{AB}}{\bar{S}_E}$
误差	S_E	$(r-1)(s-1)$	$\bar{S} = \frac{S_E}{rs(r-1)}$	
总和	S_T	$rs-1$		

第三节 方差分析

例2.3.1 现有一学校的6个学院69个班级共2077名学生的数学成绩表格，现要分析不同学院的学生的成绩是否有显著性差别。

	A	B	C	D	E	F
1	学号	姓名	班级	学院	学院编号	成绩
2	05010101	郭强	050101	机械	1	87
3	05010102	张旭鹏	050101	机械	1	71
4	05010103	李桂艳	050101	机械	1	75
5	05010104	杨功	050101	机械	1	78
6	05010105	禹善强	050101	机械	1	76
7	05010106	刘达	050101	机械	1	66
8	05010107	刘中晗	050101	机械	1	61
9	05010108	王振波	050101	机械	1	67
10	05010109	赵长亮	050101	机械	1	82
11	05010110	石增辉	050101	机械	1	74
12	05010111	过建奇	050101	机械	1	72

第三节 方差分析

(1) 正态性检验

调用lillietest( )函数检验6个学院的学生的考试成绩是否服从正态分布

```
result =
    0.0734
    0.1783
    0.1588
    0.1494
    0.4541
    0.0727
```

对6个学院的学生的考试成绩进行的正态检验的p值均大于0.05，说明在显著性水平0.05下均接受原假设，认为6个学院的学院的考试成绩服从正态分布

第三节 方差分析

(2)方差齐次性检验

调用vartestn( )函数检验6个学院的学生的考试成绩是否服从方差相同的正态分布

[p,stats]=vartestn(score,college)

输出结果：p = 0.7138

Group Summary Table

Group	Count	Mean	Std Dev
机械	510	72.5608	9.09237
电信	404	74.4703	8.6516
化工	349	79.8968	8.53766
环境	206	73.1068	8.50184
经管	303	69.4323	8.77352
计算机	305	67.9508	8.48494
Pooled	2077	73.0857	8.72236

Bartlett's statistic	2.91
Degrees of freedom	5
p-value	0.714

检验的p值>0.05，说明在显著性水平0.05下接受原假设，6个学院的学院的考试成绩服从方差相同的正态分布

第三节 方差分析

(3)方差分析

调用anoval（ ）函数进行单因素一元方差分析，检验不同学院的学生的考试成绩有无显著差别

[p,table,stats]=anova1(score,college)

输出结果：p = 5.6876e-74

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	29191.9	5	5838.38	76.74	5.68764e-74
Error	157560.8	2071	76.08		
Total	186752.7	2076			

anova1函数返回的p值<0.05，故拒绝原假设，认为不同学院的学生的考试成绩有非常显著的差别

第三节 方差分析

(3)多重比较

不同学院的学生的考试成绩有非常显著的差别，但这并不意味着任意两个学院学生的成绩都有显著性差别。多重比较，为找出考试成绩存在显著性差别的学院。

调用multcompare( )函数

第三节 方差分析

'组序号'	'组序号'	'置信下限'	'置信上限'	'组均值差'	'置信上限'
[1]	[2]	[-3.5650]	[-1.9095]	[-0.2540]	[0.0130]
[1]	[3]	[-9.0628]	[-7.3361]	[-5.6093]	[2.0676e-08]
[1]	[4]	[-2.5980]	[-0.5460]	[1.5060]	[0.9743]
[1]	[5]	[1.3255]	[3.1284]	[4.9313]	[1.1298e-05]
[1]	[6]	[2.8108]	[4.6100]	[6.4092]	[2.0679e-08]
[2]	[3]	[-7.2430]	[-5.4266]	[-3.6101]	[2.0676e-08]
[2]	[4]	[-0.7645]	[1.3635]	[3.4915]	[0.4489]
[2]	[5]	[3.1490]	[5.0380]	[6.9270]	[2.0676e-08]
[2]	[6]	[4.6340]	[6.5195]	[8.4049]	[2.0676e-08]
[3]	[4]	[4.6061]	[6.7901]	[8.9740]	[2.0676e-08]
[3]	[5]	[8.5128]	[10.4645]	[12.4163]	[2.0676e-08]
[3]	[6]	[9.9977]	[11.9460]	[13.8943]	[2.0676e-08]
[4]	[5]	[1.4299]	[3.6745]	[5.9190]	[4.5367e-05]
[4]	[6]	[2.9144]	[5.1560]	[7.3976]	[2.1451e-08]
[5]	[6]	[-0.5346]	[1.4815]	[3.4976]	[0.2902]

第三节 方差分析

[p,table,stats]=anova1( ) 单因素方差分析

[p,table,stats]=anova2(X,reps)： 双因素方差分析

[p,table,stats]=anovan ( ) 多因素方差分析

(1) 样本观测值矩阵X：每一列对应因素A的一个水平，每行对应因素B的一个水平
(2) reps：因素A和B下的每一个水平组合下重复实验的次数
(3) p：若reps取值等于1，则p是一个包含2个元素的行向量；若reps取值大于1，则p是一个包含3个元素的行向量，其元素分别是与H_{01}，H_{02}，H_{03}对应的检验的p值
(4) table：方差分析表
(5) stats：结构体变量，用于进行后续的多重比较。

第三节 方差分析

例 2.3.2 为了研究肥料使用量对水稻产量的影响，某研究所做了氮（因素A）、磷（因素B）两种肥料施用量的二因素试验。氮肥用量设低、中、高三个水平，分布使用N1，N2和N3表示；磷肥用量设低、高2个水平，分别用P1，P2表示。供3x2=6个处理，重复4次，随机区组设计，测得水稻产量如下表：

处理	区组			
	1	2	3	4
N1P1	38	29	36	40
N1P2	45	42	37	43
N2P1	58	46	52	51
N2P2	67	70	65	71
N3P1	62	64	61	70
N3P2	58	63	71	69

根据上表中的数据，不考虑区组因素，分析氮、磷两种肥料的施用量对水稻产量是否有显著性影响，并分析交互作用是否显著.

第三节 方差分析

(1) 整理数据矩阵

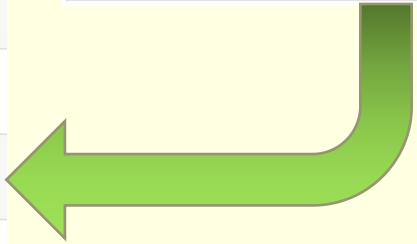
每一列对应一个A因素（氮）水平

每一行对应一个B因素（磷）水平

方法：先转置，再把第2列，第4列，第6列接到第1列，第3列，第5列下面

N1P1	N1P2	N2P1	N2P2	N3P1	N3P2
38	45	58	67	62	58
29	42	46	70	64	63
36	37	52	65	61	71
40	43	51	71	70	69

N1P1	N2P1	N3P1
38	58	62
29	46	64
36	52	61
40	51	70
N1P2	N2P2	N3P2
45	67	58
42	70	63
37	65	71
43	71	69



	N1	N2	N3
P1	38	58	62
P1	29	46	64
P1	36	52	61
P1	40	51	70
P2	45	67	58
P2	42	70	63
P2	37	65	71
P2	43	71	69

第三节 方差分析

(2) 方差分析

```
yield=[38 29 36 40
    45 42 37 43
    58 46 52 51
    67 70 65 71
    62 64 61 70
    58 63 71 69];
yield=yield';
```

%定义一个矩阵，输入原始数据

%矩阵转置

```
yield=[yield(:,[1,3,5]);yield(:,[2,4,6])];
```

%将数据矩阵yield转换成8行3列的矩阵，列对应因素A（氮），行对应因素B（磷）

```
top={'因素','N1','N2','N3'};
left={'P1';'P1';'P1';'P1';'P2';'P2';'P2';'P2'};
```

%定义元胞数组，以元胞数组形式显示转换后的数据

```
[top;left,num2cell(yield)]
```

%显示数据

```
[p,table,stats]=anova2(yield,4)
```

%调用anova2函数作双因素方差分析

第三节 方差分析

运行结果：

p = 0.0000 0.0004 0.0080

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	3067	2	1533.5	78.31	0
Rows	368.167	1	368.17	18.8	0.0004
Interaction	250.333	2	125.17	6.39	0.008
Error	352.5	18	19.58		
Total	4038	23			

结果表明：氮、磷两种肥料的施用量对水稻的产量均有显著的影响。但是，若想知道氮、磷在哪种组合下水稻产量更高，需要进行多重比较。

第三节 方差分析

(3) 多重比较

%调用multcompare对各处理进行多重比较

1.0000	2.0000	-25.9446	-16.0000	-6.0554	0.0009
1.0000	3.0000	-38.4446	-28.5000	-18.5554	0.0000
1.0000	4.0000	-15.9446	-6.0000	3.9446	0.4236
1.0000	5.0000	-42.4446	-32.5000	-22.5554	0.0000
1.0000	6.0000	-39.4446	-29.5000	-19.5554	0.0000
2.0000	3.0000	-22.4446	-12.5000	-2.5554	0.0093
2.0000	4.0000	0.0554	10.0000	19.9446	0.0483
2.0000	5.0000	-26.4446	-16.5000	-6.5554	0.0006
2.0000	6.0000	-23.4446	-13.5000	-3.5554	0.0047
3.0000	4.0000	12.5554	22.5000	32.4446	0.0000
3.0000	5.0000	-13.9446	-4.0000	5.9446	0.7926
3.0000	6.0000	-10.9446	-1.0000	8.9446	0.9995
4.0000	5.0000	-36.4446	-26.5000	-16.5554	0.0000
4.0000	6.0000	-33.4446	-23.5000	-13.5554	0.0000
5.0000	6.0000	-6.9446	3.0000	12.9446	0.9251

第三节 方差分析

```
ans =

    'A=N1,B=P1'    [35.7500]    [2.2127]
    'A=N2,B=P1'    [51.7500]    [2.2127]
    'A=N3,B=P1'    [64.2500]    [2.2127]
    'A=N1,B=P2'    [41.7500]    [2.2127]
    'A=N2,B=P2'    [68.2500]    [2.2127]
    'A=N3,B=P2'    [65.2500]    [2.2127]
```

第5个处理（即N2P2）的平均值最大，而处理5与处理3,6差异不显著，所以可以认为第3个和第6个处理也是可以的，所以，综上，可以在处理3,5,6中做出选择，即N3P1，N2P2，N3P2。

第四节　响应面回归分析简介

一、响应面回归的试验设计

二、响应面回归的理论基础

一、 响应面回归的试验设计

响应面分析(Response Surface Analysis)把因变量和自变量之间的关系用图形表示出来，称为响应面分析。

原因：如果试验有许多因素，它们的重要性通过正交设计识别出少数几个因素，进一步当试验区域接近响应曲面的最优区域或位于最优区域中，采用二阶段设计。

一、响应面回归的试验设计

目的：在最优值附近找到使得因变量达到最优值的因素的水平或区间。

方法：采用多元二次回归方程来拟合因素与因变量之间的函数关系。

原因：一般函数在最值附近可用二次函数很好的近似，计算简单。

一、响应面回归的试验设计

优点：目的性强。

缺点：响应面优化的前提是：设计的试验点应包括最佳的试验条件或在最佳试验条件附近，如果实验点的选取不当，使用响应面优化法是不能得到很好的优化结果的。

一、响应面回归的试验设计

因素水平编码

将所有因素的取值范围都转化为中心在原点的一个“立方体”中，这一变换称为对因素水平的编码。

设计变量初选试验范围，最大值编码为1，最小值编码为-1，中间值编码为0。

常用的试验法BBD：中心组合设计。

一、响应面回归的试验设计

$BBD_{15}(3^3)$

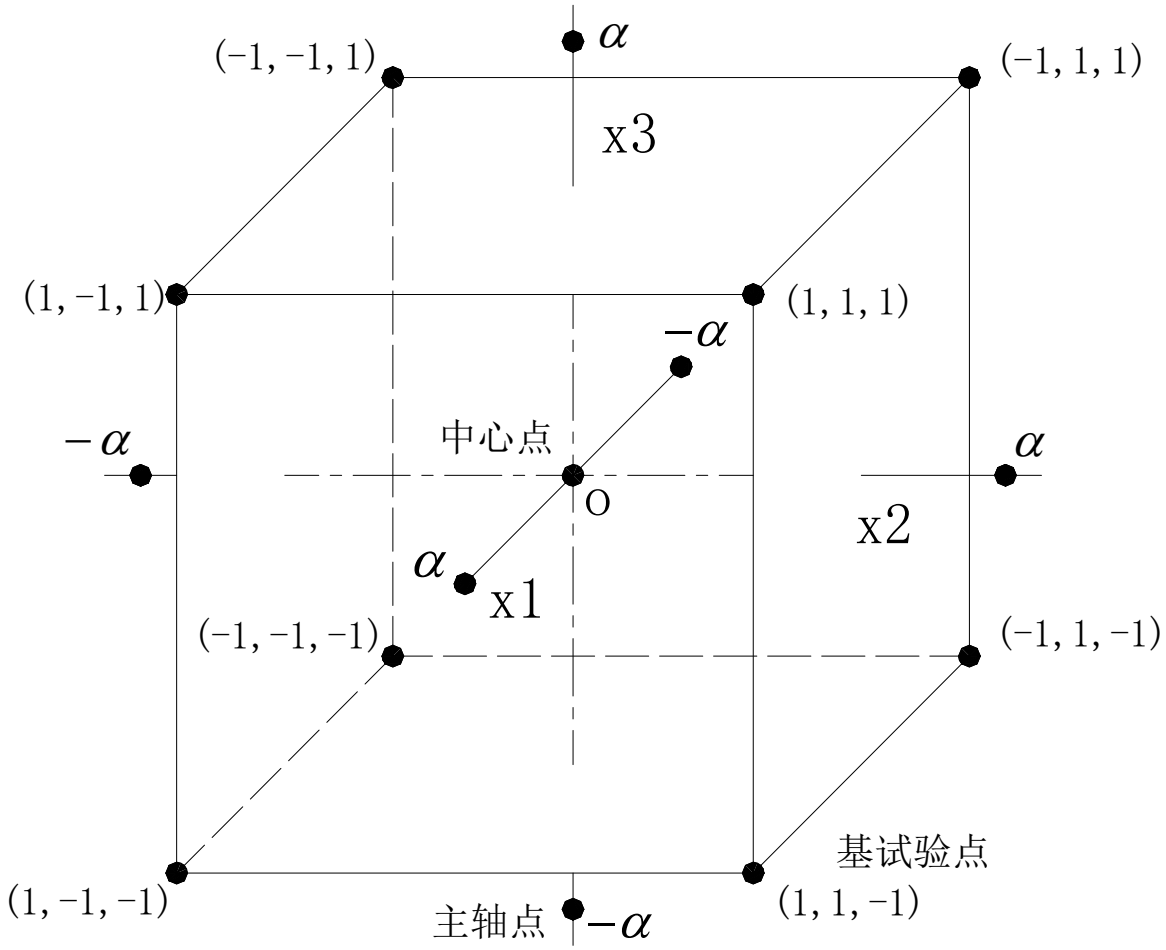
试验号	因素1	因素2	因素3
1	-1	-1	0
2	-1	0	-1
3	-1	0	1
4	-1	1	0
5	0	-1	-1
6	0	-1	1
7	0	1	-1
8	0	1	1
9	1	-1	0
10	1	0	-1
11	1	0	1
12	1	1	0
13	0	0	0
14	0	0	0
15	0	0	0

正交试验设计与中心组合设计的对比

正交试验多在边上，而中心组合设计多重复并且在中心

步骤：广撒网，有重点，即先正交再做中心组合试验

一、 响应面回归的试验设计



三因素响应面设计的试验点的分布图

二、响应面回归的理论基础

1、二次响应面（多元二次多项式）模型

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{i<j}^{p} \beta_{ij} x_i x_j + \sum_{j=1}^{p} \beta_{jj} x_j^2 + \varepsilon$$

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

- 三元二次响应面模型描述：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \varepsilon$$

矩阵描述　　$y = X\beta + \varepsilon$　　$\varepsilon \sim N(0, \sigma^2 I)$

二、响应面回归的理论基础

2、回归系数的最小二乘估计

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

$$X'Y = X'X\beta$$

当$X'X$可逆时，解得

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y$$

二、响应面回归的理论基础

3、回归方程的显著性检验

$$H_0:\beta_1=\beta_2=\cdots=\beta_p=0 \quad \leftrightarrow \quad H_1:\text{至少有一个}\beta_j\neq 0$$

记 $$y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}$$

检验统计量

$$F=\frac{S_R/p}{S_E/(n-p-1)}\overset{H_0}{\sim}F(p,n-p-1)$$

$$W=\{F\geq F_\alpha(p,n-p-1)\}$$

二、响应面回归的理论基础

4、回归系数的检验

$$H_{0j}:\beta_j=0 \quad \leftrightarrow \quad H_{1j}:\beta_j\neq 0$$

检验统计量

$$T=\frac{\hat{\beta}_j/\sqrt{c_{jj}}}{\hat{\sigma}} \overset{H_0}{\sim} t(n-p-1)$$

$$F=\frac{\hat{\beta}_j^2/c_{jj}}{\hat{\sigma}^2} \overset{H_0}{\sim} F(1,n-p-1)$$

$$W=\{|T|\geq t_{\frac{\alpha}{2}}(n-p-1)\}$$

$$W=\{F\geq F_{\alpha}(1,n-p-1)\}$$

二、响应面回归的理论基础

5、失拟检验

安排重复试验的目的：弄清影响因变量的因素除指定的因素外，考虑是否还有不可忽视的其他因素，例如交互作用，这种检验称为失拟检验。

失拟检验是一种用来判断回归模型是否合适的检验。

二、响应面回归的理论基础

模型好坏的判断：残差。

残差是由两部分组成的：一部分是模型拟合得再好，它也消除不了，称为随机误差或纯误差； 另一部分与模型有关，模型不合适，这部分的值就大，称为失拟误差。

二、响应面回归的理论基础

残差平方和分解为误差平方和与失拟平方和。

$$S_E = S_e + S_{Lf}$$

其中 $S_e = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2$ $df = N - n$ $\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$

$S_{Lf} = \sum_{i=1}^{n} m_i(\bar{y}_i - \hat{y}_i)^2$ $df = n - p - 1$

检验统计量

$$F = \frac{S_{Lf}/(n-p-1)}{S_e/(N-n)} \overset{H_0}{\sim} F(n-p-1, N-n)$$

$$W = \{F \geq F_\alpha(n-p-1, N-n)\}$$

二、响应面回归的理论基础

例：在化学反应中，采用中心组合设计，选取对萃取影响较大的三个因素：压力x_1、温度x_2、时间x_3作为考察对象，得到的萃取率如下表，试作二次响应面回归分析。

二、响应面回归的理论基础

萃取率与各因素的中心组合试验方案与结果

试验号	x_1	x_2	x_3	萃取率
1	-1	-1	0	63.46
2	-1	0	-1	62.88
3	-1	0	1	67.24
4	-1	1	0	56.76
5	0	-1	-1	63.15
6	0	-1	1	72.96
7	0	1	-1	68.42
8	0	1	1	69.65
9	1	-1	0	59.22
10	1	0	-1	68.90
11	1	0	1	70.63
12	1	1	0	70.72
13	0	0	0	67.13
14	0	0	0	65.84
15	0	0	0	66.96

二、响应面回归的理论基础

一、响应面回归分析的Matlab函数命名为

xiangying.m

函数的调用方法：

xishu=xiangying(Y, X, n)

In	X 自变量观测数组；	Output	xishu
	Y 因变量观测数组；		
	n 非重复试验次数		

二、响应面回归的理论基础

二、xiangying函数主要模块及其功能

1. 输入参数检验模块
2. 数据处理模块
3. 线性回归分析模块
4. 失拟检验模块模块
5. 数据表格输出模块

注： 1、5主要是实现函数的辅助功能；

2~4是函数的核心部分.

二、响应面回归的理论基础

2. 数据处理模块　[nx,px]=size(X);

```
my=mean(Y);
pingfang=X.^2;
u=0;  %计数
for i=1:px-1     %px=3是X的列数
   for j=i+1:px
       u=u+1;
       jiaocha(:,u)=X(:,i).*X(:,j);
    end
end
zong=[X, jiaocha, pingfang];
if (nx-size(zong,2)-1)<2
    error('非重复观测数据太少，至少为%d',3+size(zong,2));
end
```

X列交叉项的乘积按列保存在jiaocha数组中

X　　Y

x_1	x_2	x_3	萃取率
-1	-1	0	63.46
-1	0	-1	62.88
-1	0	1	67.24
-1	1	0	56.76
0	-1	-1	63.15
0	-1	1	72.96
0	1	-1	68.42
0	1	1	69.65
1	-1	0	59.22
1	0	-1	68.90
1	0	1	70.63
1	1	0	70.72
0	0	0	67.13
0	0	0	65.84
0	0	0	66.96

二、响应面回归的理论基础

3. 线性回归分析模块

MATLAB统计工具箱中提供了regstats函数，也可用来作多重线性或广义线性回归分析，本例调用regstats函数进行线性回归分析，返回结构体变量ST. (统计函数regstats使用方法参考附录)

```
STxian = regstats(Y,X,'linear');
STjiao = regstats(Y, jiaocha,'linear');
STping = regstats(Y, pingfang,'linear');
STzong = regstats(Y, zong,'linear');
```

二、响应面回归的理论基础

3. 线性回归分析模块

```
%计算相对平方项F值及p值
fping=STping.rsquare/size(pingfang,2) / ((1-STzong.rsquare)/(nx-size(zong,2)-1));
if fping>100
    fping=100;
end
jingdu=1;pping=0.99999;
while jingdu>0.00001
    jingdu=jingdu/10;
    while finv(pping,size(pingfang,2),nx-size(zong,2)-1)>fping
        pping=pping-jingdu;
        if pping<=0
            break
        end
    end
    pping=pping+jingdu;
end
```

二、响应面回归的理论基础

3. 线性回归分析模块

```
%计算相对线性项F值及p值
fxian=STxian.rsquare/size(X,2) / ((1-STzong.rsquare)/(nx-size(zong,2)-1));
if fxian>100
    fxian=100;
end
jingdu=1;pxian=0.99999;
while jingdu>0.00001
    jingdu=jingdu/10;
    while finv(pxian,size(X,2),nx-size(zong,2)-1)>fxian
        pxian=pxian-jingdu;
        if pxian<=0
            break
        end
    end
    pxian=pxian+jingdu;
end
```

二、响应面回归的理论基础

3. 线性回归分析模块

```
%计算相对交叉项F值及p值
fjiao=STjiao.rsquare/size(jiaocha,2) / ((1-STzong.rsquare)/(nx-size(zong,2)-1));
if fjiao>90
    fjiao=90;
end
jingdu=1;pjiao=0.99999;
while jingdu>0.00001
    jingdu=jingdu/10;
    while finv(pjiao,size(jiaocha,2),nx-size(zong,2)-1)>fjiao
        pjiao=pjiao-jingdu;
        if pjiao<=0
            break
        end
    end
    pjiao=pjiao+jingdu;
end
if STzong.fstat.f>100
    STzong.fstat.f=100;
end
```

二、响应面回归的理论基础

4. 失拟检验模块模块

```
fcan=nx-size(zong,2)-1;
 suiji=sum((Y(n+1:nx)-
mean(Y(n+1:nx))).^2
fsui=nx-n-1;
shini=0;
for i=1:n
    shini=shini+(Y(i)-
sum(STzong.tstat.beta'.*[1,zong(i,:)]))^2;
end
shini=shini+(
sum(STzong.tstat.beta'.*[1,zong(n+1,:)])-
mean(Y(n+1:nx)))^2;
fshi=fcan-fsui;
F=(shini/fshi) / (suiji/fsui);
jingdu=1;P=0.999;
```

%fcan残差误差自由度

%suiji 随机误差，
%13至15行为重复实验

%fsui 随机误差自由度

%shini失拟误差

%fshi失拟误差自由度

%计算F值

%筛选相应的p值

二、响应面回归的理论基础

4. 失拟检验模块模块

接上页

```
while jingdu>0.0001
    jingdu=jingdu/10;
    while finv(P,fshi,fsui)>F
        P=P-jingdu;
        if P<=0
            break
        end
    end
    P=P+jingdu;
end
```

二、响应面回归的理论基础

三、计算结果分析

表2.4.1 响应面回归分析信息表

响应平均	66.26133
Root MSE	1.0065
判定系数(R^2)	0.9818
变异系数	1.5190

表2.4.2 模型检验表

回归	自由度	平方和	R^2	F值	p值
线性项	3	88.1364	0.3169	28.9990	0.0014
平方项	3	81.9923	0.2948	26.9774	0.0016
交叉项	3	102.9433	0.3701	33.8708	0.0009
总模型	9	273.0721	0.9818	29.9491	0.0008

二、响应面回归的理论基础

由表2.4.1和2.4.2可知，模型中的一次项、二次项、交叉项以及总模型所对应的显著概率均小于0.01，故认为它们在模型中均是极为显著的。

二、响应面回归的理论基础

表2.4.3 模型失拟检验表

残差	自由度	平方和	F值	p值
失拟误差	3	4.0830	2.7706	0.2763
随机误差	2	0.9825		
残差平方和	5	5.0655		

由表2.4.3可以看到，失拟检验的显著性概率为0.2763大于0.05，故可以认为模型没有出现拟合不足，说明模型拟合优良。

二、响应面回归的理论基础

表2.4.4 参数估计结果

变量	估计值	标准误差	t值	p值
常数项	66.6433	0.5811	114.6810	0.0000
x_1	2.3913	0.3559	6.7196	0.0011
x_2	0.8450	0.3559	2.3745	0.0636
x_3	2.1412	0.3559	6.0171	0.0018
$x_1 \cdot x_2$	4.5500	0.5033	9.0410	0.0003
$x_1 \cdot x_3$	-0.6575	0.5033	-1.3065	0.2483
$x_2 \cdot x_3$	-2.1450	0.5033	-4.2622	0.0080
$x_1 \cdot x_1$	-2.6179	0.5238	-4.9978	0.0041
$x_2 \cdot x_2$	-1.4854	0.5238	-2.8358	0.0364
$x_3 \cdot x_3$	3.3871	0.5238	6.4662	0.0013

二、响应面回归的理论基础

从参数估计表**2.4.4**可得出二次响应面回归模型为

$$y = 66.6433 + 2.3913x_1 + 0.8450x_2 + 2.1413x_3$$
$$+4.5500x_1x_2 - 0.6575x_1x_3 - 2.1450x_2x_3$$
$$-2.6179x_1^2 - 1.4854x_2^2 + 3.3871x_3^2$$

二、响应面回归的理论基础

(3) 图形分析

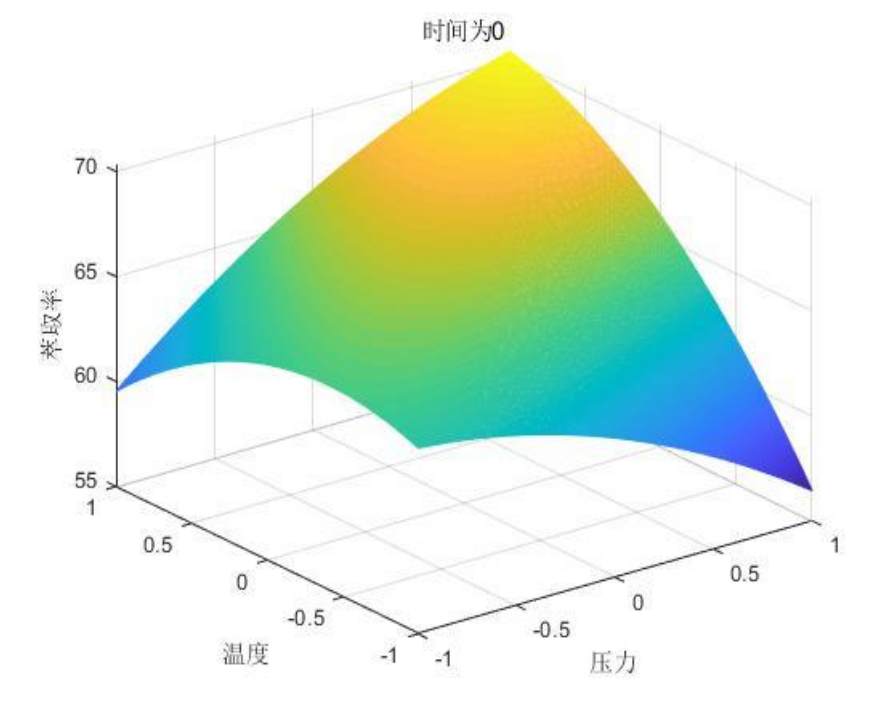


图2.4.2 萃取率对压力x_1与温度x_2的响应面

在本试验水平范围内，当时间处于中心水平时，压力减小、温度增大时，萃取率先增大后减小。当压力和温度编码取值为1时，萃取率取得最大值。

二、响应面回归的理论基础

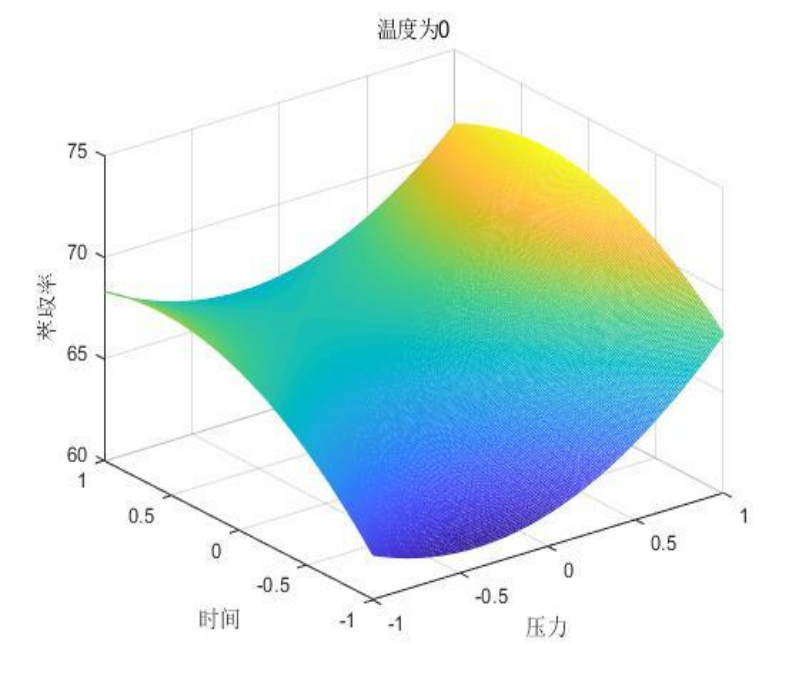


图2.4.3 萃取率对压力x_1与时间x_3的响应面

在本试验水平范围内，温度处于中心水平时，固定温度和时间中某一因素水平不变，萃取率随压力的减小而减小，随时间的增大而增大，压力和时间的编码取值在1、1附近时，萃取率可得最大值。

二、响应面回归的理论基础

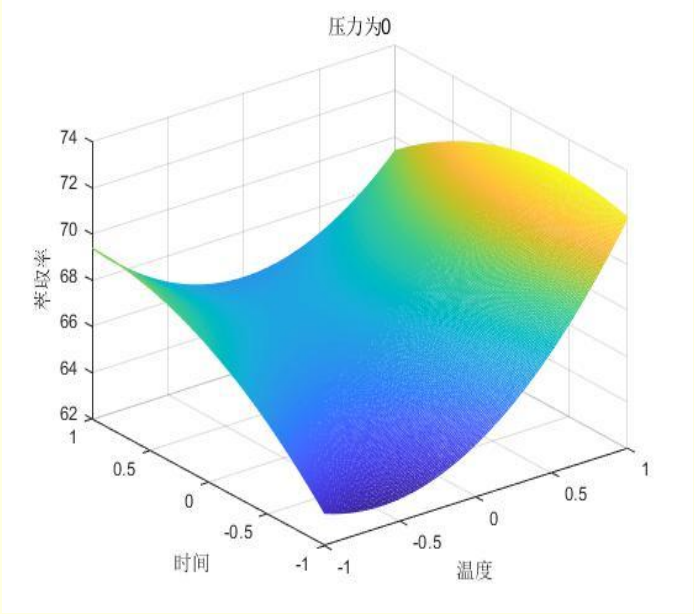


图2.4.4 萃取率对温度x_2与时间x_3的响应面

在本试验水平范围内，当压力处于中心水平时，当时间编码值为-1，温度编码值在-1附近时，萃取率最小。固定温度为某一水平时，随着时间的增加，萃取率取值也逐渐变大。温度和时间的编码取值在-1、1附近时，萃取率可得最大值。

二、响应面回归的理论基础

响应面回归是一种综合试验设计和数学建模的优化方法，可有效减少试验次数，给出直观等高线图和三维立体图，并可考察影响因素之间的交互作用。

响应面分析法不仅建立了预测模型，并对模型适应性、模型和系数显著性和失拟项进行检验，进一步进行方差分析、模型诊断。

二、响应面回归的理论基础

通过响应面法能有效指导工艺参数的优化，有利于提高生产效益。但是构造能够满足实际工程优化设计的响应面近似模型是一个比较复杂的过程，还需要反复进行试验数据的收集、近似模型的拟合及响应面精度。

二、响应面回归的理论基础

当然，响应面回归法也有局限性。响应面优化的前提是：设计的实验点应包括最佳的实验条件，如果实验点的选取不当，使用响应面分析法是不能得到很好的优化结果的。因而在使用响应面回归方法之前，应当确定合理的实验影响因素与水平。

谢 谢

CONTRACTED WIND POWERPOINT TEMPLATE DESIGNS CONTRACTED WIND POWERPOINT TEMPLATE DESIGNS
CONTRACTED WIND POWERPOINT TEMPLATE DESIGNS CONTRACTED WIND POWERPOINT TEMPLATE DESIGNS

附录

MATLAB统计工具箱中提供了regstats函数，也可用来作多重线性或广义线性回归分析，它的调用方式如下：

regstats(y,X,model)

stats = regstats(…)

stats = regstats(y,X,model,whichstats)

（1）regstats(y,X,model)

作多重线性回归分析。输入参数X为自变量观测值矩阵（或设计矩阵），它是的矩阵。默认情况下，regstats函数自动在X第1列元素的左边加入一列1，不需要用户自己添加。输入参数y为因变量的观测值向量，是的列向量。可选的输入参数model是一个字符串，用来控制回归模型的类型，其可用的字符串如表1-2所示。

表1-2 regstats函数支持的model参数

附录

model参数的参数值

说 明

'linear'

带有常数项的线性模型（默认情况）

'interaction'

带有常数项、线性项和交叉项的模型

'quadratic'

带有常数项、线性项、交叉项和平方项的模型

'purequadratic'

带有常数项、线性项和平方项的模型

在这种调用方式下，regstats函数会生成一个交互式图形用户界面（GUI），界面上带有回归诊断统计量列表，包括系数的估计值、因变量的预测值、残差、判定系数、调整的判定系数、F检验和t检验的相关结果等，共23个可选项。通过这个界面，用户可以很方便地将回归分析的各种结果导入MATLAB工作空间。

附录

（2）stats = regstats(…)

返回一个结构体变量stats，它有24个字段，包括了回归分析的所有诊断统计量。这种调用方式不生成图形用户界面，stats的后23个字段分别与图形用户界面上的23个选项相对应。

（3）stats = regstats(y,X,model,whichstats)

仅返回由whichstats参数指定的统计量。whichstats可以是形如 'leverage' 的单个字符串，也可以是形如{'leverage' 'standres' 'studres'}的字符串元胞数组。若whichstats是字符串 'all' ，则返回所有统计量。

注意：当需要计算F统计量的观测值时，模型中应包含常数项。若模型中不包含常数项，regstats函数输出的F统计量的观测值是不正确的。在不考虑常数项的情况下，计算出的判定系数2R的值可能是负的，说明所用模型不适合用户的数据。