4.3 数据处理的方法与模型

在大数据处理时代，大量数据的处理十分重要。本节主要介绍几种数据处理的方法和模型，包括Logistic模型、灰色模型及预测、神经网络方法、模糊综合评价法以及几个具体案例，如水道测量数据问题，电池剩余放电时间预测，葡萄酒的评价问题。通过这些方法或案例的学习，掌握常用的数据处理方法，并学习如何利用软件或编程来完成任务。

目录 CONTENTS

4.3.1 Logistic模型

4.3.2 灰色模型及预测

4.3.3 神经网络方法

4.3.4 模糊综合评判法

4.3.5 水道测量数据问题

4.3.6 电池剩余放电时间预测

4.3.7 葡萄酒的评价问题

4.3.1 Logistic模型

目录

01 马尔萨斯人口模型

02 阻滞型人口模型

03 实例

01 马尔萨斯人口模型

01 马尔萨斯人口模型

设时刻t时人口为$x(t)$，单位时间内人口增长率为r，则 时间内增长的人口为：

$$x(t+\Delta t)-x(t)=x(t).r.\Delta t$$

当$\Delta t \to 0$，得到微分方程：

$$\frac{dx}{dt}=r.x, x(0)=x_0$$

则$x(t)=x_0.e^{r.t}$。待求参数x_0, r。

为便于求解，两边取对数有：$y=a+r.t$

其中 $y=\ln x, a=\ln x_0$

该模型化为线性求解。

02 阻滞型人口模型

02 阻滞型人口模型

设时刻 t 时人口为 $x(t)$，环境允许的最大人口数量为 x_m，人口净增长率随人口数量的增加而线性减少，即

$$r(t) = r.\ (1 - \frac{x}{x_m})$$

由此建立阻滞型人口微分方程：

$$\frac{dx}{dt} = r(1 - \frac{x}{x_m}).x, x(0) = x_0$$

则：$x(t) = \dfrac{x_m}{1 + (\frac{x_m}{x_0} - 1).e^{-r.t}}$，待求参数 x_0, x_m, r。此即为Logistic函数。

当 $x = \dfrac{x_m}{2}$ 时，x 增长最快，即 $\dfrac{dx}{dt}$ 最大。

02 阻滞型人口模型

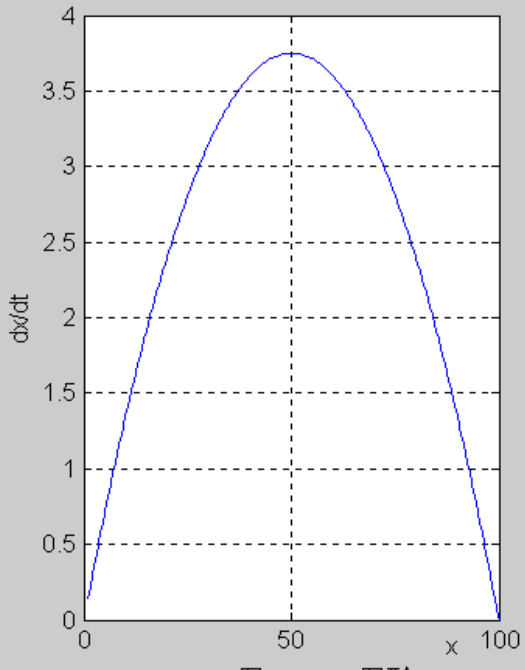


图1 dx/dt~x图形

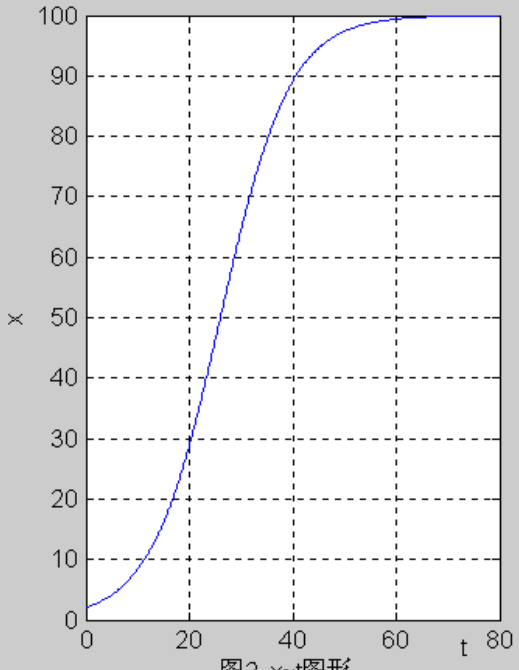


图2 x~t图形

03 实例

03 实例1 美国人口数据处理

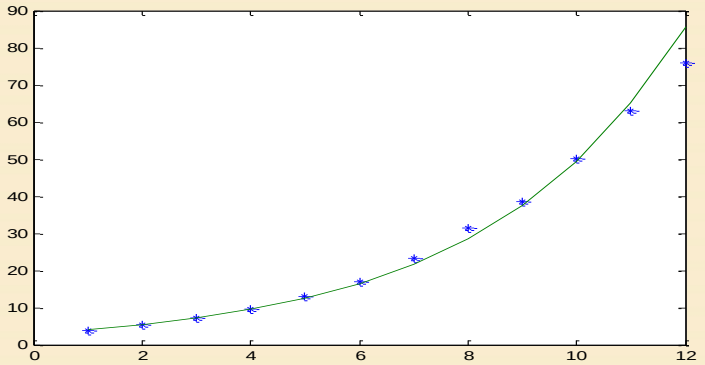
表 美国人口数据表(人口数量单位：百万)

年	1790	1800	1810	1820	1830	1840	1850	1860
实际人口	3.9	5.3	7.2	9.6	12.9	17.1	23.2	31．4
指数模型	4.1884	5.5105	7.2498	9.538	12.549	16.5097	1.7209	28.5769
阻滞模型	8.1699	10.0238	12.2875	15.0464	18.4006	22.4670	27.3796	33.2893
年	1870	1880	1890	1900	1910	1920	1930	1940
实际人口	38.6	50.2	62.9	76.0	92.0	106.5	123.2	131.7
指数模型	37.597	49.464	65.077	85.618				
阻滞模型	40.3625	48.7771	58.7152	70.3529	83.8457	99.3094	116.799	136.2846
年	1950	1960	1970	1980	1990	2000	2010	年
实际人口	150.7	179.3	204.0	226.5	251.4	281.4	309.35	实际人口
指数模型								指数模型
阻滞模型	157.637	180.6116	204.851	229.9025	255.245	280.333	304.645	阻滞模型

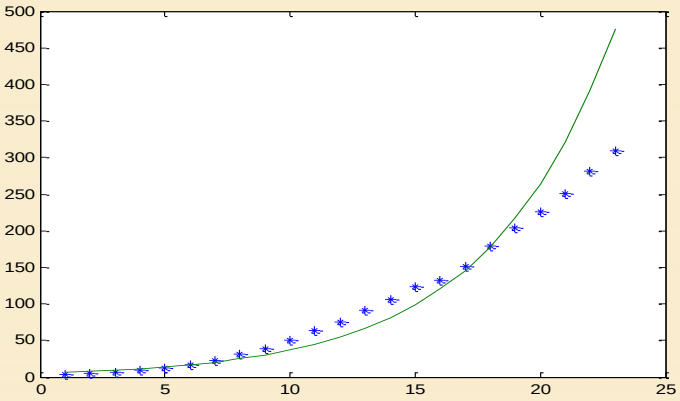
03 实例1 美国人口数据处理

由指数增长模型得到模型为： $y = 3.186e^{0.2743t}$

(1790 ~ 1900年数据)
均方误差根为 RMSE=3.0215
结果图见右图（效果好）



(1790 ~ 2000年数据)
均方误差根为 RMSE=39.8245
结果图见右图（效果不好）



03 实例1 美国人口数据处理

指数模型求解MatLab程序:

```
%美国人口模型，指数增长模型
x=[3.9,5.3,7.2,9.6,12.9,17.1,23.2,31.4,38.6,50.2,62.9,76.0,92.0,...
  106.5,123.2,131.7,150.7,179.3,204.0,226.5,251.4,281.4,309.35]';
n=12;
xx=x(1:n);%1790年到1900年数据
t=[ones(n,1),(1:n)'];
y=log(xx(1:n));
[b,bint,r,rint,stats]=regress(y,t);
RR=stats(1);%复相关系数
F=stats(2);%F统计量值
prob=stats(3); % 概率
x0=exp(b(1)); %参数x0;
r=b(2); %参数r
py=x0*exp(r*t(:,2)); %预测数据
err=xx-py;
rmse=sqrt(sum(err.^2)/n); %均方误差根

plot(1:n,xx,'*',1:n,py); %作对比图
```

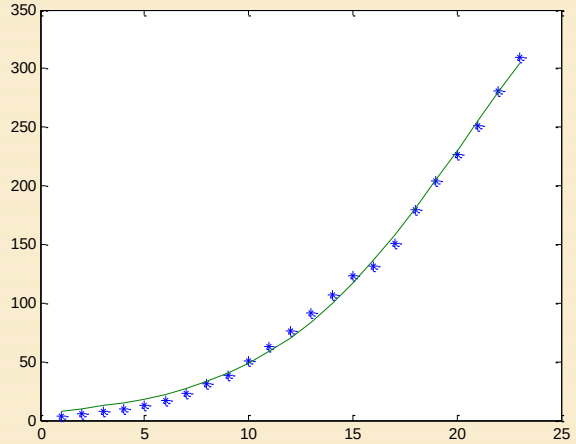
03 实例1　美国人口数据处理

拟合1790年到2000年数据，得到结果为：

$$x_0 = 6.6541, x_m = 486.9046, r = 0.2084$$

$$y = \frac{486.9046}{1 + 72.1733e^{-0.2084t}}$$

均方误差根为　RMSE=4.7141，并预测2020年美国人口为327.7204（百万），结果图见下图。



03 实例1　美国人口数据处理

求解MatLab程序：

```
%美国人口模型，阻滞型增长模型
x=[3.9,5.3,7.2,9.6,12.9,17.1,23.2,31.4,38.6,50.2,62.9,76.0,92.0,...
   106.5,123.2,131.7,150.7,179.3,204.0,226.5,251.4,281.4,309.35]';
n=length(x);
y=x(1:n);%1790年到2010年数据
t=(1:n)';
beta0=[5.3,0.22,400,]; %[x0,r,xm]
[beta,R,J]=nlinfit(t,y,'logisfun',beta0);%R为残差,beta为待求参数

py=beta(3)./(1+(beta(3)/beta(1)-1)*exp(-beta(2)*t));%预测各年人口

p24=beta(3)./(1+(beta(3)/beta(1)-1)*exp(-beta(2)*24));%预测2020年人口

rmse=sqrt(sum(R.^2)/n); %均方误差根

plot(1:n,y,'*',1:n,py); %作对比图

%拟合函数
logisfun.m
function yhat=logisfun(beta,x)
yhat=beta(3)./(1+(beta(3)./beta(1)-1).*exp(-beta(2)*x));
```

03 实例2 根据某省职工历年平均工资统计表，预测未来40年工资数

某省职工33年平均工资如右表所示。

表 山东省职工工资表（单位：元）

年 份	平均工资	年 份	平均工资
1978	566	1995	5145
1979	632	1996	5809
1980	745	1997	6241
1981	755	1998	6854
1982	769	1999	7656
1983	789	2000	8772
1984	985	2001	10007
1985	1110	2002	11374
1986	1313	2003	12567
1987	1428	2004	14332
1988	1782	2005	16614
1989	1920	2006	19228
1990	2150	2007	22844
1991	2292	2008	26404
1992	2601	2009	29688
1993	3149	2010	32074
1994	4338		

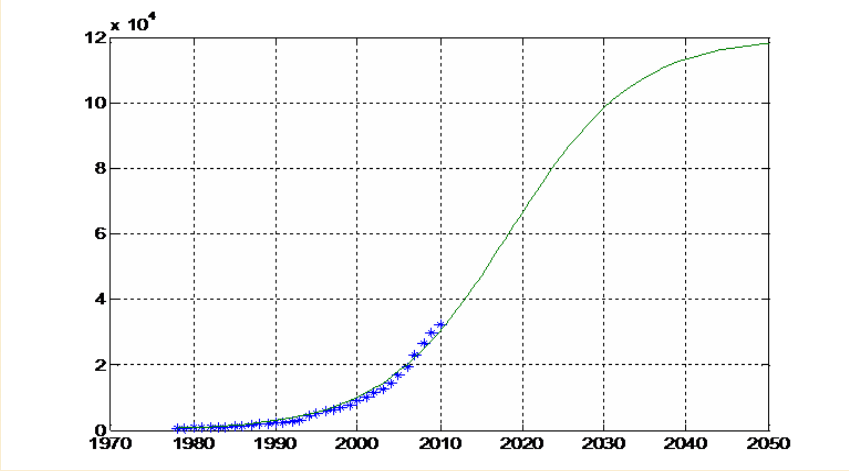
03 实例2 根据某省职工历年平均工资统计表，预测未来40年工资数

采用阻滞型模型

$$x(t) = \frac{x}{1 + (\frac{x_m}{x_0} - 1).e^{-r.t}}$$

将1978年到2010年共33年的年平均工资代入该模型，见图

$$x_0 = 550, r = 0.13, x_m = 120000$$



Logistic拟合结果

03 实例3　2011-ICMC电动汽车问题

汽车的类型分为传统的燃油型（CV）、电动型（EV）和混合型（HEV）三种类型，对比分析了未来50年对环境、社会、经济和健康方面的影响。选定的代表性国家有三个：法国，美国和中国。法国作为欧洲的代表，中国作为亚洲的代表，美国作为美洲的代表。

在该部分中，首先预测了未来50年汽车总量，估计出了未来50年CV、EV和HEV的变化。预测了未来50年三个国家汽车的增长。采用了阻滞型的Logistic模型。

建立的微分方程为：

$$\begin{cases} \dfrac{dx}{dt} = r.x\left(1 - \dfrac{x}{M}\right) \\ x(0) = x_0 \end{cases}$$

由该方程得到的解为：

$$x(t) = \frac{M}{1 + (\frac{M}{x_0} - 1).e^{-rt}}$$

其中r 为增长率，M为饱和量，也就是汽车最大容量，x_0 为初始值，取2010年的汽车总量。

03 实例3　2011-ICMC电动汽车问题

在该模型中，首先需要估计模型的参数：汽车最大容量M和年增长率r 。根据2005年到2010年三个国家的历史数据进行估计。这三个国家历史数据见下表。

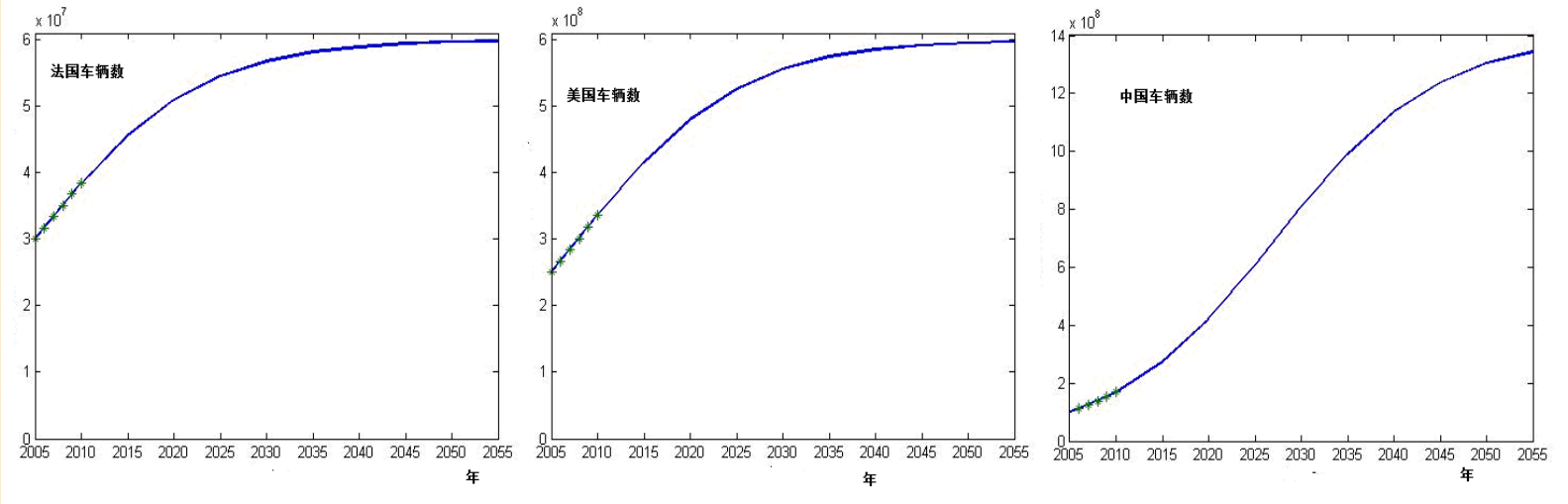
国家	2005	2006	2007	2008	2009	2010
法国(10^7)	3	3.17	3.34	3.51	3.68	3.8
美国(10^8)	2.4	2.5	2.9	3.0	3.1	3.2
中国(10^8)	1	1.11	1.24	1.37	1.52	1.68

估计得到的三个国家的模型参数见下表

参　数	法　国	美　国	中　国
M	60,000,000	600,000,000	1,400,000,000
r	0.115	0.115	0.115

03 实例3　2011-ICMC电动汽车问题

以2010年的数据作为初始值，利用估计得到参数值M和r，对未来50年汽车拥有量进行预测。得到的法国、美国和中国的预测结果曲线见下图



结果显示，法国汽车拥有量在2030年左右保持稳定，饱和量是6,000万辆；美国的汽车拥有量在2030年也变化很小，其饱和量也是60,000万辆；中国在2015年迅速增长，一直增长到 2050年，其饱和量为140,000万辆。

4.3.2 灰色模型及预测

灰色系统理论建模要求原始数据必须**等时间间距**。首先对原始数据进行累加生成，目的是弱化原始时间序列数据的随机因素，然后建立生成数的微分方程。

GM(1,1)模型是灰色系统理论中的单序列一阶灰色微分方程，它所需信息较少，方法简便。

GM(1,1)模型

设已知序列为 $x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n)$,做一次累加AGO（Acumulated Generating Operation）生成新序列：

$$x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n)$$

其中$x^{(1)}(1)=x^{(0)}(1),x^{(1)}(2)=x^{(1)}(1)+x^{(0)}(2),\cdots,\ x^{(1)}(n)=x^{(1)}(n-1)+x^{(0)}(n)$

也即 $x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i)$ $k=1,2,...,n$

生成均值序列：$z^{(1)}(k)=\alpha x^{(1)}(k)+(1-\alpha)x^{(1)}(k-1)$ $k=2,3,...,n$

其中 $0\le\alpha\le1$ 通常可取 $\alpha=0.5$

建立灰微分方程：$x^{(0)}(k)+az^{(1)}(k)=b \qquad k=2,3,...,n$

相应的GM (1,1)白化微分方程为：

$$\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b$$

将方程变形为：$-az^{(1)}(k)+b=x^{(0)}(k) \qquad k=2,3,...,n$

其中a，b为待定模型参数。

将方程组 $-az^{(1)}(k)+b=x^{(0)}(k) \quad k=2,3,...,n$ 采用矩阵形式表达为：

$$\begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \cdots & \cdots \\ -z^{(1)}(n) & 1 \end{bmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \cdots \\ x^{(0)}(n) \end{pmatrix}$$

即：$X\beta = Y$

其中 $X = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \cdots & \cdots \\ -z^{(1)}(n) & 1 \end{bmatrix}$ $\beta = \begin{pmatrix} a \\ b \end{pmatrix}$ $Y = \begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \cdots \\ x^{(0)}(n) \end{pmatrix}$

解方程 $-az^{(1)}(k)+b=x^{(0)}(k) \quad k=2,3,...,n$

得到最小二乘解为： $\beta=(a,b)^T=(X^TX)^{-1}X^T.Y$

求解微分方程 $\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b$ 得到 **GM(1,1)模型离散解**

$$\hat{x}^{(1)}(k)=[x^{(0)}(1)-\frac{b}{a}]e^{-\alpha(k-1)}+\frac{b}{a} \qquad k=2,3,\cdots,n$$

还原为原始数列预测模型为：

$$\hat{x}^{(0)}(k)=\hat{x}^{(1)}(k)-\hat{x}^{(1)}(k-1) \quad k=2,3,4,\cdots,n$$

将 $\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b$ 代入 $-az^{(1)}(k)+b=x^{(0)}(k)$ 得：

$$\hat{x}^{(0)}(k)=[x^{(0)}(1)-\frac{b}{a}]e^{-a(k-1)}(1-e^{a}) \qquad k=2,3,4,\cdots,n$$

GM(1,1)模型与统计模型相比，具有两个显著优点：一是灰色模型即使在少量数据情况下建立的模型，精度也会很高；而统计模型在少量数据情况下，精度会相对差一些。二是灰色模型从其机理上讲，越靠近当前时间点精度会越高，因此灰色模型的预测功能优于统计模型。灰色系统建模实际上是一种以数找数的方法，从系统的一个或几个离散数列中找出系统的变化关系，试图建立系统的连续变化模型。

实例1：

2003年的SARS疫情对中国的经济发展产生了一定的影响，特别是对部分疫情严重的省市的相关行业所造成的影响是明显的。现就某市SARS疫情对商品零售业的影响进行定量的评估分析。现有某市商品零售业统计表。

商品零售额（单位：亿元）

年代	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
1997	83.0	79.8	78.1	85.1	86.6	88.2	90.3	86.7	93.3	92.5	90.9	96.9
1998	101.7	85.1	87.8	91.6	93.4	94.5	97.4	99.5	104.2	102.3	101.0	123.5
1999	92.2	114.0	93.3	101.0	103.5	105.2	109.5	109.2	109.6	111.2	121.7	131.3
2000	105.0	125.7	106.6	116.0	117.6	118.0	121.7	118.7	120.2	127.8	121.8	121.9
2001	139.3	129.5	122.5	124.5	135.7	130.8	138.7	133.7	136.8	138.9	129.6	133.7
2002	137.5	135.3	133.0	133.4	142.8	141.6	142.9	147.3	159.6	162.1	153.5	155.9
2003	163.2	159.7	158.4	145.2	124	144.1	157.0	162.6	171.8	180.7	173.5	176.5

解答：

SARS发生在2003年4月。因此我们可根据1997年到2002年的数据，预测2003年的各月的零售额，并与实际的零售额进行比对。从而判断2003年哪几个月受到SARS影响，并给出影响大小的评估。

将1997—2002年的数据记作矩阵 $A_{6\times12}$代表6年72个数据。

计算各年平均值 $x^{(0)}(i)=\frac{1}{12}\sum_{j=1}^{12}a_{ij} \quad i=1,2,\cdots,6$

得到 $x^{(0)}$ = (87.6167, 98.5000, 108.4750, 118.4167, 132.8083,145.4083)

计算累加序列 $x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i) \qquad k=1,2,...,6$

得到：

$x^{(1)}$ = (87.6167, 186.1167, 294.5917, 413.0083, 545.8167, 691.2250)

生成均值序列： $z^{(1)}(k)=\alpha x^{(1)}(k)+(1-\alpha)x^{(1)}(k-1) \quad k=2,3,\cdots,n$

这里取 $\alpha=0.4$

$z^{(1)}=$ (0, 127.0167, 229.5067, 341.9583, 466.1317, 603.9800)

建立灰微分方程： $x^{(0)}(k)+az^{(1)}(k)=b \quad k=2,3,\cdots,6$

相应的GM(1,1)白化微分方程为：

$$\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b$$

求解微分方程得到 $a=-0.0993 \quad b=35.5985$

GM(1.1)模型的离散解：$\hat{x}^{(1)}(k)=[x^{(0)}(1)-\frac{b}{a}]e^{-\alpha(k-1)}+\frac{b}{a} \quad k=2,3,\cdots,6$

还原为原始数列预测模型为：

$$\hat{x}^{(0)}(k)=\hat{x}^{(1)}(k)-\hat{x}^{(1)}(k-1) \qquad k=2,3,4,5,6$$

将 $\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b$ 代入 $-az^{(1)}(k)+b=x^{(0)}(k)$ 得：

$$\hat{x}^{(0)}(k)=[x^{(0)}(1)-\frac{b}{a}]e^{-a(k-1)}(1-e^{a}) \qquad k=2,3,4,\cdots,6$$

取 $k=7$ 得到2003年销售额平均值的预测值为：

$\hat{x}^{(0)}(7)=162.8793$ 则全年总销售额为 $T=12\times\hat{x}^{(0)}(7)=1954.55$

下面估计2003年各月的销售额。

根据前6年数据估计各月销售额的比例 $r_1,r_2,\cdots,r_{12}$

其中，$r_j=\dfrac{\sum_{i=1}^{6}a_{ij}}{\sum_{i=1}^{6}\sum_{j=1}^{12}a_{ij}}$ 计算得到

$r=(0.0794,0.0807,0.0749,0.0786,0.0819,0.0818,0.0845,0.0838,0.0872,0.0886,0.0866,0.0920)$

从而2003年各月销售额预测为：

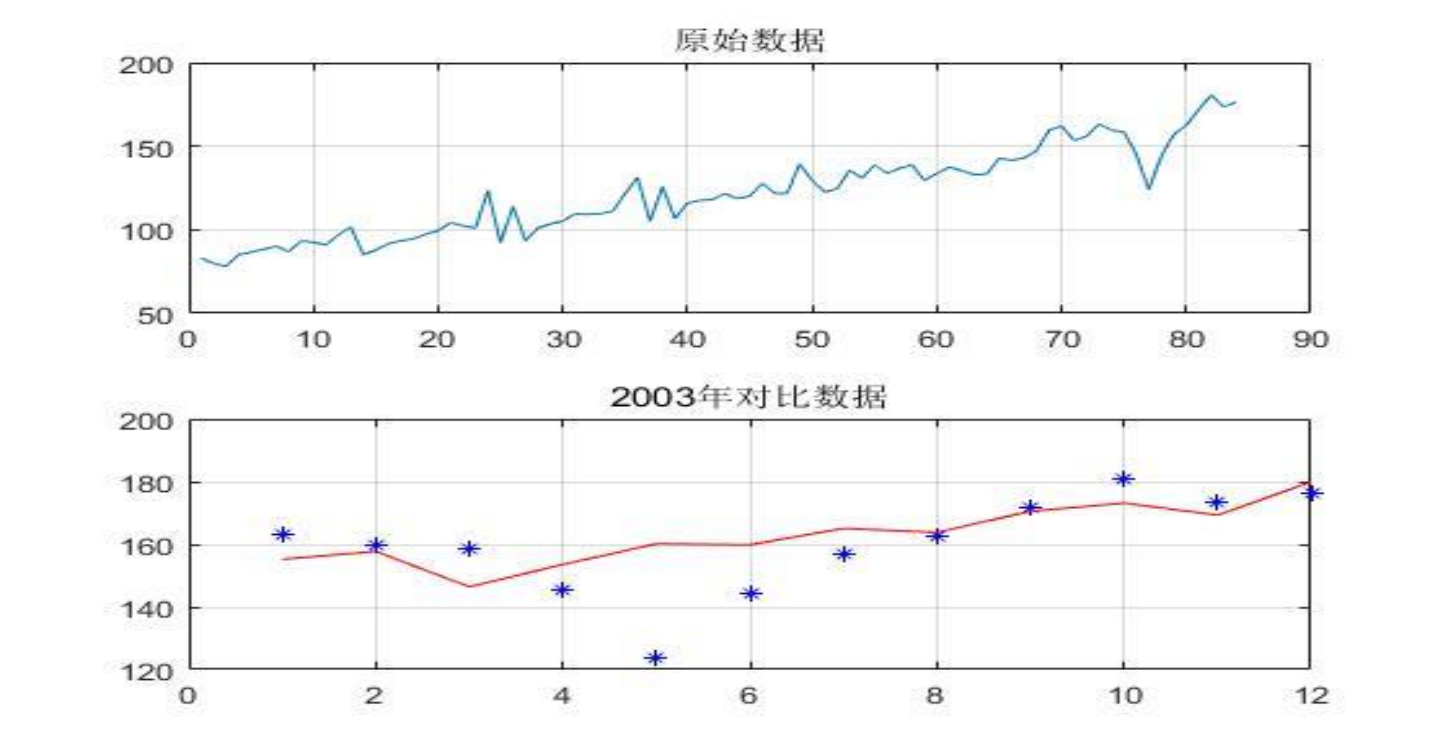
155.2,157.7,146.4,153.5,160.1,159.8,165.1,163.8,170.5,173.1,169.3,179.8

比较2003年实际销售额和预测值得到下表。

2003年商品实际销售额和预测（亿元）

月份	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
预测	155.2	157.7	146.4	153.5	160.1	159.8	165.1	163.8	170.5	173.1	169.3	179.8
实际	163.2	159.7	158.4	145.2	124	144.1	157.0	162.6	171.8	180.7	173.5	176.5

结果分析：2003年4、5、6月实际销售额为145.2、124、144.1亿元，统计结果这三个月受SRAS影响最严重，损失估计为62亿元。我们从数据的分析来看，这三个月预测值都高于实际销售额，这也与统计相符合。这三个月我们的预测值总和与实际值总和之差为60.22亿元。与统计也吻合，说明我们所建模型合理，图为对比直观图。(具体程序见书P155)

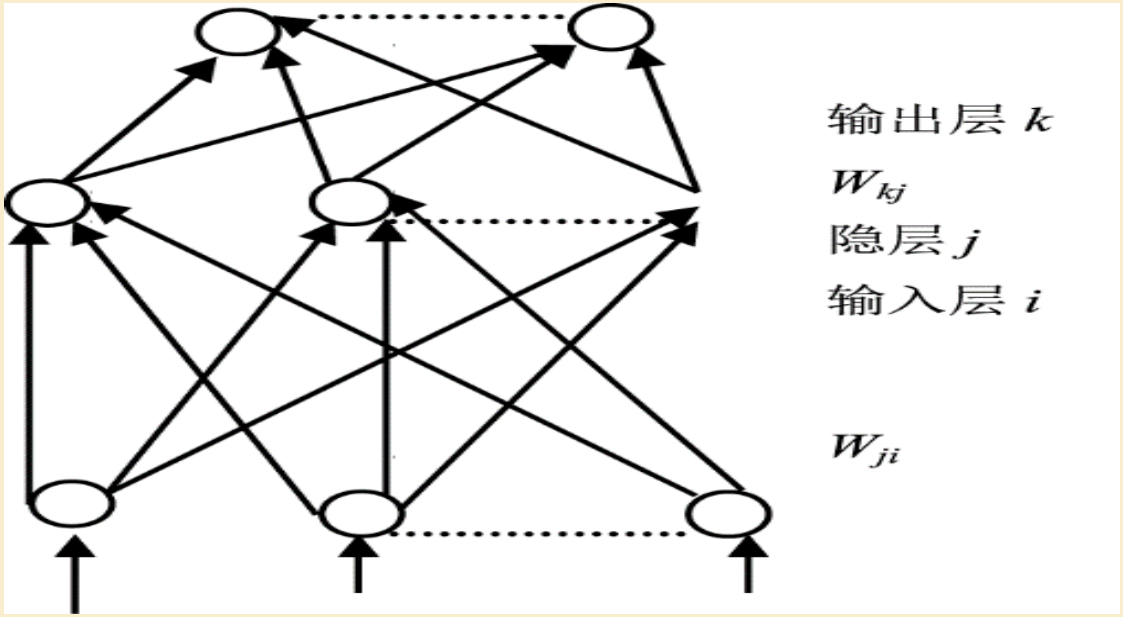


4.3.3 神经网络方法

1. 多层前向神经网络原理介绍

多层前向神经网络(MLP)是神经网络中的一种，它由一些最基本的神经元即节点组成，图4-27就是这样一个网络。这种网络的结构如下：网络由分为不同层次的节点集合组成，每一层的节点输出到下一层节点，这些输出值由于连接不同而被放大、衰减或抑制。除了输入层外，每一节点的输入为前一层所有节点输出值的和。每一节点的激励输出值由节点输入、激励函数及偏置量决定。

下图中，输入模式的各分量作为第*i*层各节点的输入，这一节点的输出，或者完全等于它们的输入值，或由该层进行归一化处理，使该层的输出值都在+1或-1之间。



多层前向神经网络图

在第j层，节点的输入值为：

$$net_j = \sum w_{ji} o_i + \theta_j$$

式中θ_j为阈值，正阀值的作用将激励函数沿x轴向左平移.

节点的输出值为：$o_j = f(net_j)$

式中 f为节点的激励函数，通常选择如下Sigmoid函数:

$$f(x) = \frac{1}{1+\exp(-x)}$$

在第 k 层的网络节点输入为：

$$net_k = \sum w_{kj} o_j + \theta_k$$

而输出为：$o_k = f(net_k)$

在网络学习阶段，网络输入为模式样本 $x_p = \{x_{pi}\}$

网络要修正自己的权值及各节点的阀值,使网络输出不断接近期望值 t_{pk} 每做一次调整后，换一对输入与期望输出，再做一次调整，直到满足所有样本的输入与输出间的对应。一般说来，系统输出值 $\{o_{pk}\}$ 与期望输出值 $\{t_{pk}\}$ 是不相等的。

对每一个输入的模式样本，平方误差 E_p 为：

$$E_p = \frac{1}{2}\sum_k (t_{pk} - o_{pk})^2$$

而对于全部学习样本，系统的总误差为：

$$E = \frac{1}{2p}\sum_p \sum_k (t_{pk} - o_{pk})^2$$

在学习过程中，系统将调整连接权和阀值，使E_p尽可能快地下降。

2. MatLab相关函数介绍

（1）网络初始化函数

net=newff（ $[x_m, x_M]$,[$h_1, h_2, \cdots, h_k$],{ $f_1, f_2, \cdots, f_k$}）

其中，x_m和x_M分别为列向量,存储各个样本输入数据的最小值和最大值；第2个输入变量是一个行向量，输入各层节点数（从隐层开始）；第3个输入变量是字符串，代表该层的传输函数（从隐层开始）。

常用tansig和logsig函数。其中

$$\text{tansig(x)}=\frac{1-e^{-2x}}{1+e^{2x}} \qquad \text{logsig(x)}=\frac{1}{1+e^{-x}}$$

除了上面方法给网络赋值外，还可以用下面格式设 定参数。

Net.trainParam.epochs=1000 设定迭代次数

Net.trainFcn=’traingm’ 设定带动量的梯度下降算法

(2) 网络训练函数

[net,tr,Y1,E]=train(net,X,Y)

其中X为$n\times M$矩阵,n为输入变量的个数,M为样本数，Y为矩阵，m为输出变量的个数。X，Y分别存储样本的输入输出数据。net为返回后的神经网络对象，tr为训练跟踪数据，tr.perf为各步目标函数值。$Y1$为网络的最后输出，$E1$为训练误差向量。

(3) 网络泛化函数

$$Y2=\mathrm{sim}(net,X1)$$

其中$X1$为输入数据矩阵，各列为样本数据，$Y2$为对应输出值。

3. 神经网络实验

（1）函数仿真实验

产生下列函数在[0,10]区间上间隔为0.5的数据，然后用神经网络进行学习。

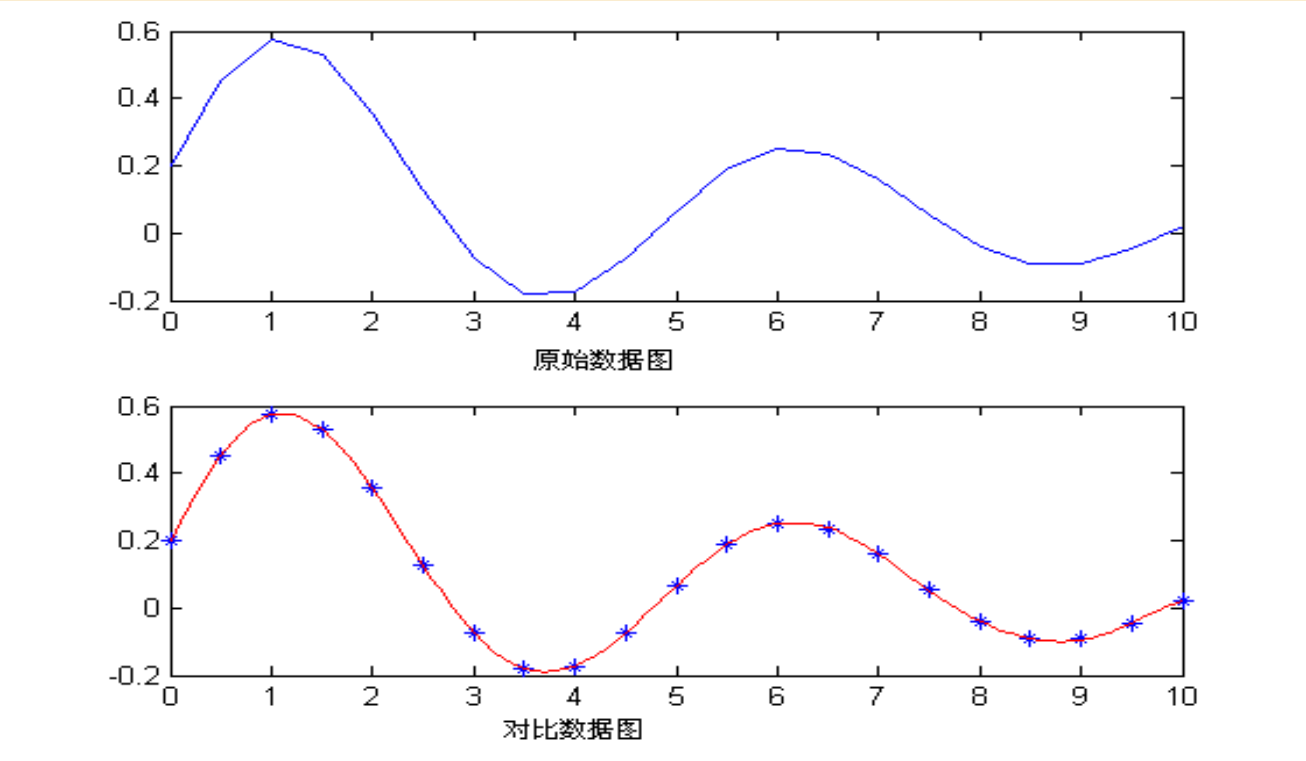
并推广到[0,10]上间隔为0.1上各点的函数值。并分别做出图形。

$$y = 0.2e^{-0.2x} + 0.5 * e^{-0.15x}.\sin(1.25x) \qquad 0 \le x \le 10$$

MATLAB程序：

```
x=0:0.5:10;
y=0.2*exp(-0.2*x)+0.5*exp(-0.15*x).*sin(1.25*x);

plot(x,y)   %画原始数据图

net.trainParam.epochs=5000;  %设定迭代步数
net=newff([0,10],[6,1],{'tansig','tansig'}); %初始化网络

net=train(net,x,y);  %进行网络训练

x1=0:0.1:10;
y1=sim(net,x1);  %数据泛化

plot(x,y,'*',x1,y1,'r');  %作对比图
```

从图形上看，神经网络输出的值比原始数据的曲线光滑。说明神经网络对该函数的学习效果很好。



原始与对比数据图

(2) MCM89A蠓的分类

有两种蠓Af和Apf。根据它们的触角(mm)和翼长(mm)进行区分。现有9只Af和6只Apf。样本数据如表所示。

9只Af 的触角和翼长

触角	1.24	1.36	1.38	1.38	1.38	1.40	1.48	1.54	1.56
翼长	1.72	1.74	1.64	1.82	1.90	1.70	1.82	1.82	2.08

6只Apf的触角和翼长

触角	1.14	1.18	1.20	1.26	1.28	1.30
翼长	1.78	1.96	1.86	2.0	2.0	1.96

另有3只待判的蠓,触角和翼长数据为（1.24,1.80）（1.28,1.84）（1.40,2.04）试对它们进行判断。

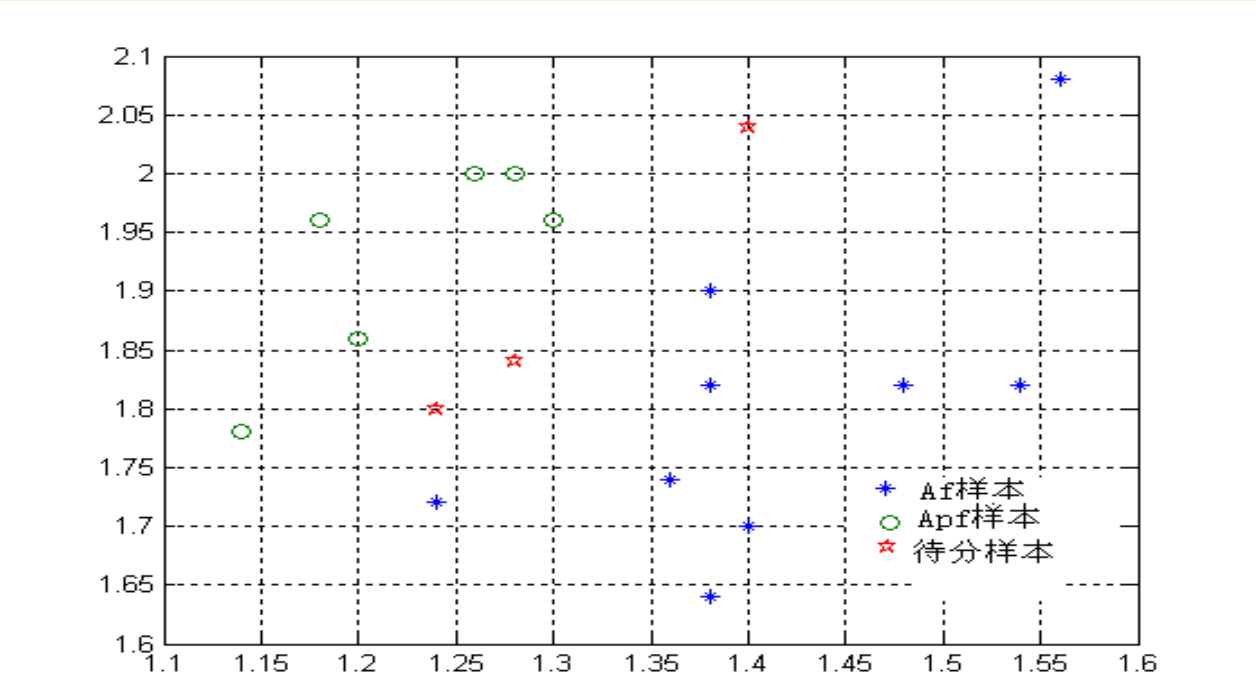
愚蠢的人类，你们能判断出我和我的两个兄弟是哪一类吗？

这里，我们可用三层神经网络进行判别。

输入为15个二维向量，输出也为15个二维向量。其中Af对应的目标向量为(1,0)，Apf对应的目标向量为(0,1)。

MATLAB程序：

```
x=[1.24,1.36,1.38,1.38,1.38,1.40,1.48,1.54,1.56,1.14,1.18,1.20,1.26,1.28,1.30;
   1.72,1.74,1.64,1.82,1.90,1.70,1.82,1.82,2.08,1.78,1.96,1.86,2.0, 2.0,1.96];

y=[1,1,1,1,1,1,1,1,1,0,0,0,0,0,0;
   0,0,0,0,0,0,0,0,0,1,1,1,1,1,1];

xmin1=min(x(1,:)); %求最小与最大值
xmax1=max(x(1,:));
xmin2=min(x(2,:));
xmax2=max(x(2,:));
net.trainParam.epochs=2500;  %设定迭代步数
net=newff([xmin1,xmax1;xmin2,xmax2],[5,2],{'logsig','logsig'}); %初始化网络
net=train(net,x,y);  %进行网络训练
x1=[1.24,1.28,1.40;
    1.80,1.84,2.04];%待分样本
y1=sim(net,x1);  %数据泛化

plot(x(1,1:9),x(2,1:9),'*',x(1,10:15),x(2,10:15),'o',x1(1,:),x1(2,:),'p')  %画原始数据图（图4-28）。
```



Af、Apf及待分样本数据图

三个样本输出值为：

```
y1=0.1235     0.8995     0.0037
   0.8785     0.0951     0.9986
```

以两个分量越靠近就判断为哪一类。从该结果看，第二个样本分为Af；而第一和第三个样本分为Apf。

但由于每次训练初始参数的随机性，而待判的3个样本在两类的临界区，导致不同的训练结果会有差异，这也正常。

4.3.4 模糊综合评判法

1. 模糊综合评判理论方法

所谓模糊综合评判是在模糊环境下，考虑了多因素的影响，为了某种目的对一事物做出综合决策的方法。
设有两个有限论域：

$$U=\{x_1,x_2,\cdots,x_n\} \qquad V=\{y_1,y_2,\cdots,y_m\}$$

其中，U 代表综合评判的多种因素组成的集合，称为因素集；V 为多种决断构成的集合，称为评判集或评语集。

一般地，因素集中各因素对被评判事物的影响是不一致的，所以因素的权重分配是 U 上的一个模糊向量，记为：

$$A=(a_1,a_2,\cdots,a_n)\in F(U)$$

$$A=(a_1,a_2,\cdots,a_n)\in F(U)$$

其中，a_i 表示U 中第 i 个因素的权重(a_i 通常采用德尔斐法和层次分析法确定)，且满足

$$\sum_{i=1}^{n}a_i=1$$

此外，m 个评语也并非绝对肯定或否定。

因此，综合后的评判可看作是V 上的模糊集，记为：

$$B=(b_1,b_2,\cdots,b_m)\in F(V)$$

其中，b_j 表示第 j 种评语在评判总体V 中所占的地位。

如果有一个从U 到V 的模糊关系，那么利用R 就可以得到一个模糊变换TR。

因此，便有如下结构的模糊综合评判数学模型：

① 因素集；　② 评判集；　③ 构造模糊变换

$$\mathrm{TR}：F(U)\to F(V)$$

其中，R 为U 到V 的模糊关系矩阵，$R=(r_{ij})_{n\times m}$ 。

实际应用中，r_{ij}可以通过德尔斐法或随机调查法得到。这样，由(U,V,R)三元体构成了一个模糊综合评判数学模型。此时，若输入一个权重分配 $W=(w_1,w_2,\cdots,w_n)\in F(U)$，就可以得到一个综合评判 $B=(b_1,b_2,\cdots,b_m)\in F(V)$ 。

即

$$(b_1,b_2,\cdots,b_m)=(w_1,w_2,\cdots,w_n)\circ\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{pmatrix}$$

其中，$b_j=\vee(w_i\wedge r_{ij})$，j=1,2,…,m。这里算子“∨”“∧”分别表示“取大”“取小”的含义。也可以采用通常的乘法和加法计算。

如果$b_k=\max\{b_1,b_2,\cdots,b_m\}$，则综合评判结果为对该事物做出决断$b_k$ 。

综合评判的核心在于“综合”。众所周知，对于由单因素确定的事物进行评判是容易的，但是，当事物涉及多个因素时，就要综合诸因素对事物的影响，做出一个接近于实际的评判，以避免仅从一个因素就做出评判而带来的片面性，这就是综合评判的特点。

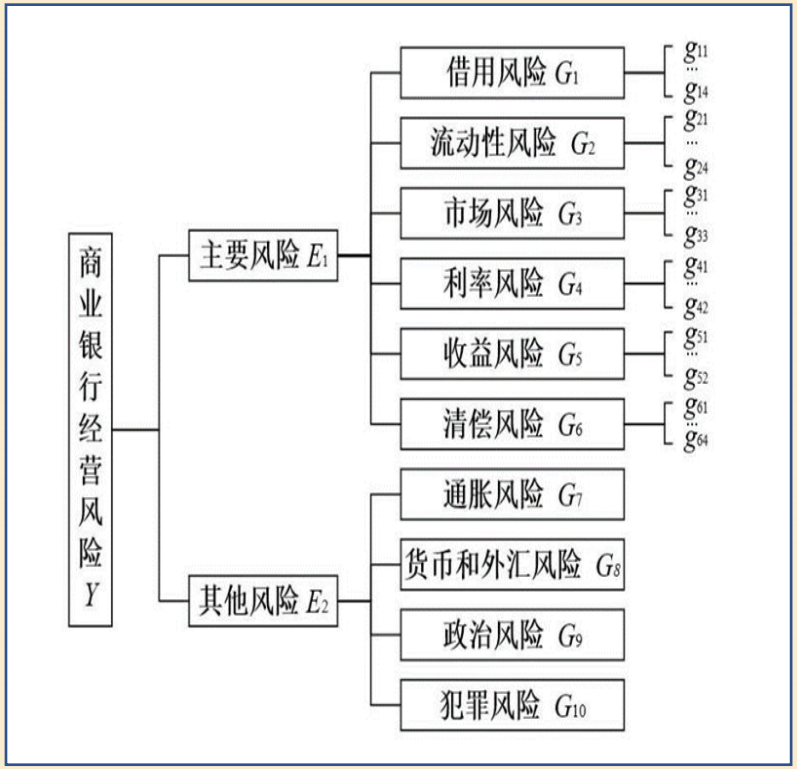
2. 实例计算

商业银行经营风险的模糊评价

资产质量是商业银行的生命线，控制风险是确保业务稳健发展的前提。
随着我国开放程度的加深及世贸组织的加入，作为市场经济主体的商业银行在谋求利润最大化的同时，经营也将处于更多的国际、国内的不确定因素之中，承受更多的风险。所以要准确地判断和评估经营风险，加强静态、动态分析，完善控制和化解风险的手段，从而确保资产质量，增强参与国际竞争的能力。

对风险的评估首先要给出评估指标体系，然后给出评估方法。下面是具体过程。

（1）商业银行经营风险的评估原则上应是商业银行经营风险程度的真实内涵。经考察，影响商业银行经营风险的因素如下图所示：



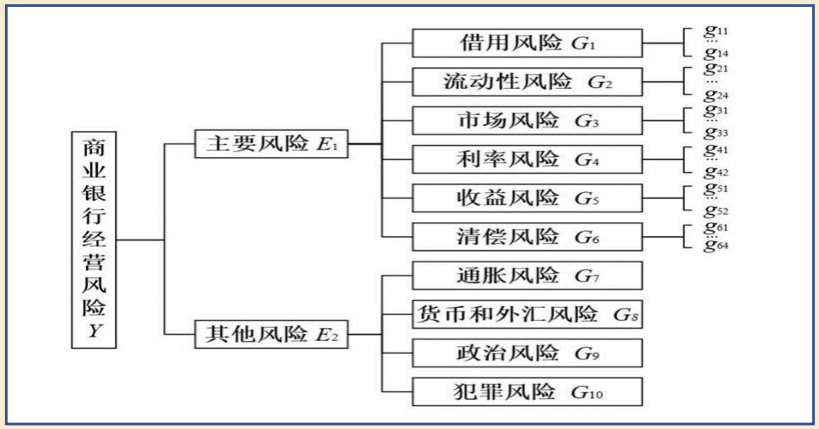
在图中所示的三层次综合评价指标体系，其中$G_1, G_2, \cdots, G_{10}$ 分别表示不同的指标子集，具体含义如下：

G1(信用风险)= $\{g_{11}, g_{12}, g_{13}, g_{14}\}$ ={不良资产与贷款和租赁合同总额之比；净贷款冲消额与贷款和租赁合约总额之比；每年贷款损失准备金(PLL)提取额与贷款和租赁合约总额比或股本总额之比；贷款损失准备金(ALL)与贷款和租赁合约总额之比或与股本总额之比}。

G2(流动性风险)= $\{g_{21}, g_{22}, g_{23}, g_{24}\}$ ={借入资金与资产总额之比；净贷款与资产总额之比；现金和同业存款与资产总额之比；现金资产加政府债券与资产总额之比}。

G3(市场风险)= $\{g_{31}, g_{32}, g_{33}\}$ ={银行资产账面价值与估计的市值之比；固定利率贷款和证券与浮动贷款和证券之比(固定利率负债与浮动利率负债之比)；银行股本的账面价值与市场价值之比}。

G4(利率风险)= $\{g_{41}, g_{42}\}$ ={利率敏感性资产与利率敏感性负债之比；未投保的存款与存款总额之比)。



G5(收益风险)=$\{g_{51}, g_{52}\}$ ={税后净收入的标准差或方差,银行股本收益率(ROE)和资产收益率(ROA)的标准差或方差}。

G6(清偿风险)=$\{g_{61}, g_{62}, g_{63}, g_{64}\}$ ={银行发行的债券的市场收益率与同期限政府证券的市场收益率的利差；银行股价与其年度每股收益之比；股本(净值)与银行资产总额之比；借入资金与负债总额之比}。

G7(通胀风险); G8(货币和外汇风险); G9(政治风险);G10(犯罪风险)。

(2) 确定各指标层的权重。采用德尔斐法和层次分析法(AHP)确定各指标层的权重。

(3) 确定评价商业银行经营风险的向量评语集。这里取评判评语集V={$v1$, $v2$, $v3$, $v4$, $v5$}，

其中$v1$、$v2$、$v3$、$v4$、$v5$分别表示指标的评语为“优” “良” “中” “可” “差”。
对应的商业银行经营风险程度为“低” “较低” “中等” “较高” “高”。

（4）对每个G_i(i=1,2,…,10)分别进行模糊综合评判。若单独考虑G_i(i=1,2,…,10)下的指标 g_{ij} ,可以通过德尔斐法或随机调查法得到 g_{ij} 隶属于第 k 个评语 v_t 的程度 r_{ijk} ， 得到G_i的模糊评价矩阵:

$$R_i=\begin{pmatrix} r_{i11} & r_{i12} & \cdots & r_{i1m} \\ r_{i21} & r_{i22} & \cdots & r_{i2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{in1} & r_{in2} & \cdots & r_{inm} \end{pmatrix}$$

其中 n 为G_i中评价的指标数目, m 为向量评语集中评语数目。

$$G_i=W_i\circ R_i=(w_1,w_2,\cdots,w_n)\circ\begin{pmatrix} r_{i11} & r_{i12} & \cdots & r_{i1m} \\ r_{i21} & r_{i22} & \cdots & r_{i2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{in1} & r_{in2} & \cdots & r_{inm} \end{pmatrix}=(a_{i1},a_{i2},\cdots,a_{im})$$

得到 G 层各指标的模糊综合评判集合，$A_i=(a_{i1},a_{i2},\cdots,a_{im})$ ， 其中 w_i 为每个G_i中评价指标权重向量，a_{ij}采用F（·,）算子（有界和“⊕”与普通乘法“”算子）求得（这里的合成运算没有采用“∨”“∧”算子）。这里，有界和“⊕”的含义是$\alpha\oplus\beta=\min\{\alpha+\beta,1\}$。

同理可得：

$$E_1=(p_1,p_2,p_3,p_4,p_5,p_6)\circ\begin{pmatrix}G_1\\G_2\\\vdots\\G_6\end{pmatrix}=(e_{11},e_{12},\cdots,e_{1m})\qquad E_2=(q_1,q_2,q_3,q_4)\circ\begin{pmatrix}G_7\\G_8\\G_9\\G_{10}\end{pmatrix}=(e_{21},e_{22},\cdots,e_{2m})$$

p_i 为对应每指标 G_i (i=1,2,…,6)的权重，q_i (i=1,2,…,4)为对应每个G_i (i=7,8,9,10)的权重。

（5）确定评价商业银行经营风险向量元素集。

$$Y=K\circ E=(k_1,k_2)\circ\begin{pmatrix}E_1\\E_2\end{pmatrix}=(y_1,y_2,\cdots,y_m)$$

其中，k_i 为对应每个 (i=1,2)的权重向量。

(6) 评判结果的处理用加权平均法。

将评判集 V 中各元素量化后，最终评判结果 $V = B.Y^T$ 。其中$B=(b_1,b_2,\cdots,b_m)$ 为 m 个评语的量化值。

具体计算如下：

商业银行经营风险指标评价体系的权重系数和评价等级在研究过程中可以是虚拟的，但在实际的评价过程中，这些数据则应由专门的机构和评估专家根据实际情况用特定的方法来确定。

现假定通过德尔斐法或随机调查法得到如下有关某商业银行经营风险指标体系的具体数据：

K=(k1,k2) = (0.6,0.4)

P=(p1,p2,p3,p4,p5,p6) =(0.2,0.15,0.15,0.2,0.15,0.15) Q=(q1,q2,q3,q4) =(0.2,0.3,0.2,0.3)

W1=(0.35,0.25,0.3,0.1) W2=(0.4,0.3,0.2,0.1) W3=(0.5,0.25,0.25) W4=(0.6,0.4) W5=(0.3.0.7)
W6=(0.3,0.2,0.2,0.3)

R1=[0.3 0.3 0.25 0.1 0.05
0.4 0.3 0.15 0.1 0.05
0.35 0.4 0.15 0.05 0.05
0.35 0.3 0.15 0.15 0.05]

R2=[0.3 0.3 0.25 0.1 0.05
0.35 0.4 0.15 0.05 0.05
0.4 0.3 0.15 0.1 0.05
0.3 0.3 0.2 0.2 0]

R3=[0.35 0.4 0.15 0.05 0.05
0.35 0.3 0.15 0.15 0.05
0.3 0.3 0.25 0.1 0.05]

R4=[0.4 0.3 0.15 0.1 0.05
0.3 0.3 0.25 0.1 0.05]

R5=[0.3 0.3 0.25 0.1 0.05
0.35 0.3 0.15 0.15 0.05]

R6=[0.35 0.4 0.15 0.05 0.05
0.2 0.3 0.2 0.15 0.15
0.4 0.3 0.15 0.1 0.05
0.3 0.3 0.25 0.1 0.05]

G1=W1○ R1=(0.345 0.33 0.185 0.09 0.05)　　G2=W2○ R2=(0.335 0.33 0.195 0.095 0.045)
G3=W3○ R3=(0.3375 0.35 0.175 0.0875 0.05)　　G4=W4○ R4=(0.36 0.3 0.19 0.1 0.05)
G5=W5○ R5=(0.335 0.3 0.18 0.135 0.05)　　G6=W6○ R6=(0.315 0.33 0.19 0.095 0.07)

E1=(p1,p2,p3,p4,p5,p6)○ (G1;G2;G3;G4;G5;G6)=
(0.2　0.15　0.15　0.2　0.15　0.15)
[0.345　0.33　0.185　0.09　0.05
0.335　0.33　0.195　0.095　0.045
0.3375　0.35　0.175　0.0875　0.05
0.36　0.3　0.19　0.1　0.05
0.335　0.3　0.18　0.135　0.05
0.315　0.33　0.19　0.095　0.07]

=(0.3394，0.3225，0.1860，0.0999，0.0523)

E2=(q1,q2,q3,q4) ○(G7;G8;G9;G10)=
(0.2　0.3　0.2　0.3)
[0.3 0.3 0.2 0.1 0.1
0.6 0.3 0.1 0 0
0.4 0.3 0.2 0.1 0
0.5 0.2 0.2 0.1 0]

=(0.47，0.27，0.17，0.07，0.02)

$$Y=(k1\ k2)\circ(E1;E2)=(0.6\quad 0.4)\begin{bmatrix}0.3394,0.3225,0.1860,0.0999,0.0523\\0.47,\ 0.27,\ 0.17,\ 0.07,\ 0.02\end{bmatrix}=(0.3916,0.3015,0.1796,0.0879,0.0394)$$

若规定评价集V中各元素的量化值为v1=100，v2=85，v3=70，v4=55，v5=40，则最终评判结果V的值介于100到40之间，通常越接近100，经营风险越小，越接近40，风险越高。

示例中$V=BY'=$ (100,85,70,55,40)×(0.3916,0.3015,0.1796,0.0879,0.0394)'=83.77。

故该商业银行经营风险较低

[MatLab程序代码见书P166]

4.3.5 水道测量数据问题

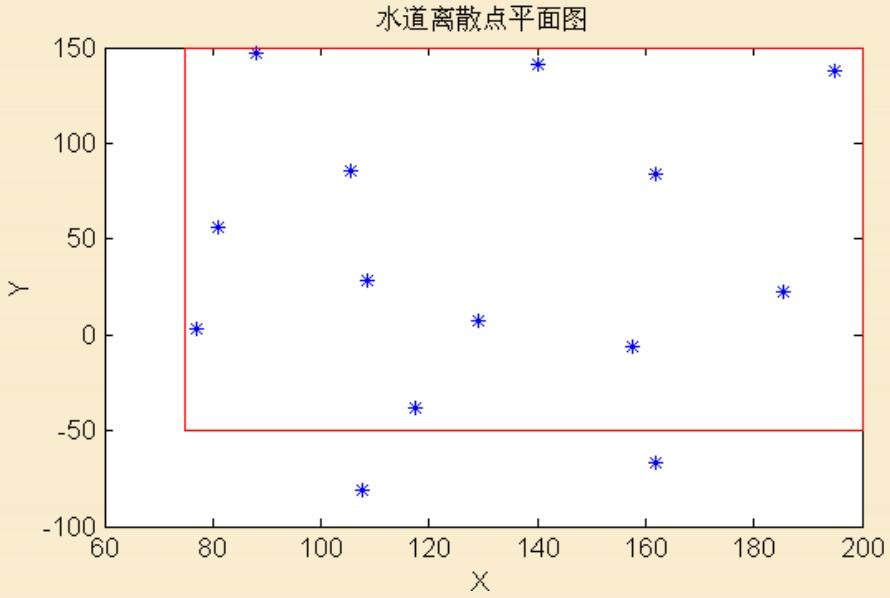
下表给出了在以码（1码=0.914米）为单位的直角坐标为X，Y的水面一点处以英尺（1英尺=0.3048米）计的水深Z。水深数据是在低潮时测得的。

X（码）	Y（码）	Z（英尺）	X（码）	Y（码）	Z（英尺）
129.0	7.5	4	157.5	-6.5	9
140.0	141.5	8	107.5	-81.0	9
108.5	28.0	6	77.0	3.0	8
88.0	147.0	8	81.0	56.5	8
185.5	22.5	6	162.0	84.0	4
195.0	137.5	8	117.5	-38.5	9
105.5	85.5	8	162.0	-66.5	9

船的吃水深度为5英尺。在矩形区域(75,200)×(-50,150)里哪些地方船要避免进入。

解答：

所给14个点的平面散点如下图，其中有两点不在所给区域：(75,200)×(-50,150)。



本问题采用地球科学上的反距离权重法(IDW)。

首先将所给区域按(75,200) × (-50,150)较细的网格进行剖分，然后利用所给14个点的水深值Z，按照IDW方法求出所有剖分点的水深值Z，并找出水深低于5英尺的点。
然后做出水底曲面图，等值线图，标出水深低于5m的区域。

IDW算法：

设有 n 个点(x_i, y_i, z_i)，计算平面上任意点(x, y) 的 z 值。

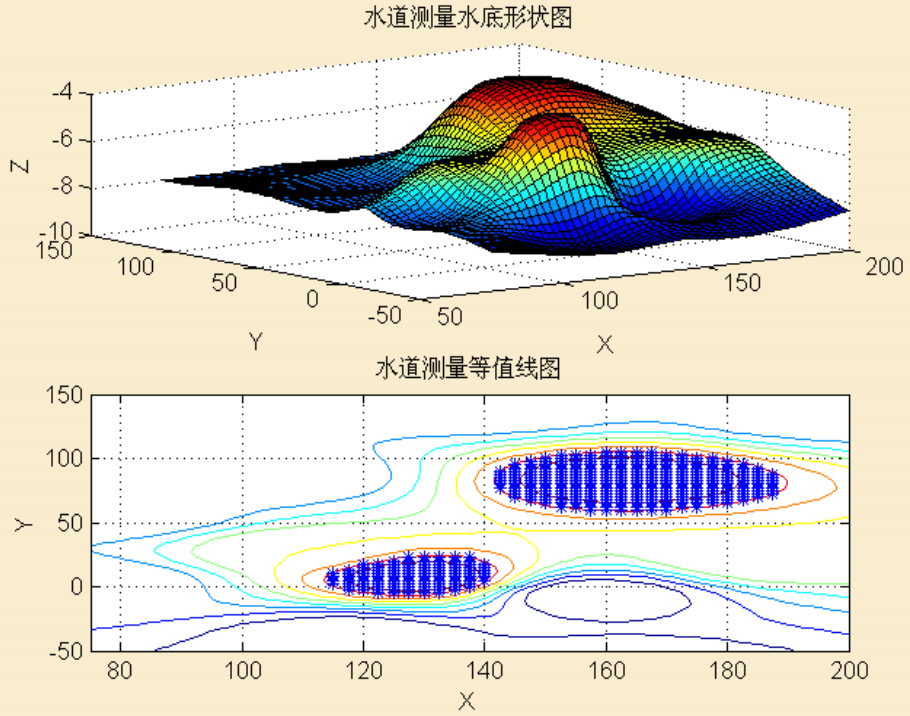
$$z = \sum_{i=1}^{n} w_i . z_i$$

其中权重：$w_i = \dfrac{1/d_i^p}{\sum_{i=1}^{n} 1/d_i^p}$，$d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$

即(x, y) 处的 z 值由各已知点加权得到，其权重为(x, y) 到各点距离的 p 次方成反比。

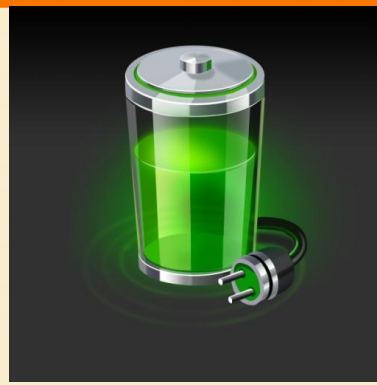
p 决定距离(x, y) 近的 (x_i, y_i) 作用的相对大小。
当 p 越大，则当 (x_i, y_i) 距离(x, y) 越近，其相对作用越大，越远相对作用越小。
本题取 $p = 3$ 。

按照IDW方法， x 方向剖分50个区间，区间间隔 dx=2.5；y 方向剖分80个区间，区间间隔 dy=2.5，得到的水底河床图及等值线如下图所示:



水道测量水底形状图和等值线图（其中蓝色竖线为水深不足5英尺的点）

[MatLab程序代码见书P169]



4.3.6电池剩余放电时间预测

问题导入

铅酸电池作为电源被广泛用于工业、军事、日常生活中。在铅酸电池以恒定电流强度放电过程中，电压随放电时间单调下降，直到额定的最低保护电压（*Um*，本题中为9V）。从充满电开始放电，电压随时间变化的关系称为放电曲线。电池在当前负荷下还能供电多长时间（即以当前电流强度放电到*Um*的剩余放电时间）是使用中必须关注的问题。电池通过较长时间使用或放置。

问题提出

问题1：附件1（参考CUMCM2016C附件1）是同一生产批次电池出厂时以不同电流强度放电测试的完整放电曲线的采样数据。请根据附件1用初等函数表示各放电曲线，并分别给出各放电曲线的平均相对误差（MRE，定义见附件1）。如果在新电池使用中，分别以30A、40A、50A、60A和70A电流强度放电，测得电压都为9.8V时，根据你获得的模型，电池的剩余放电时间分别是多少？

问题2：试建立以20A到100A之间任一恒定电流强度放电时的放电曲线的数学模型，并用MRE评估模型的精度。用表格和图形给出电流强度为65A时的放电曲线。

问题3：附件2是同一电池在不同衰减状态下以同一电流强度从充满电开始放电的记录数据。试预测附件2中电池衰减状态3的剩余放电时间。

问题一的解决

先绘制出不同电流强度下的放电曲线便于观察，如图1所示

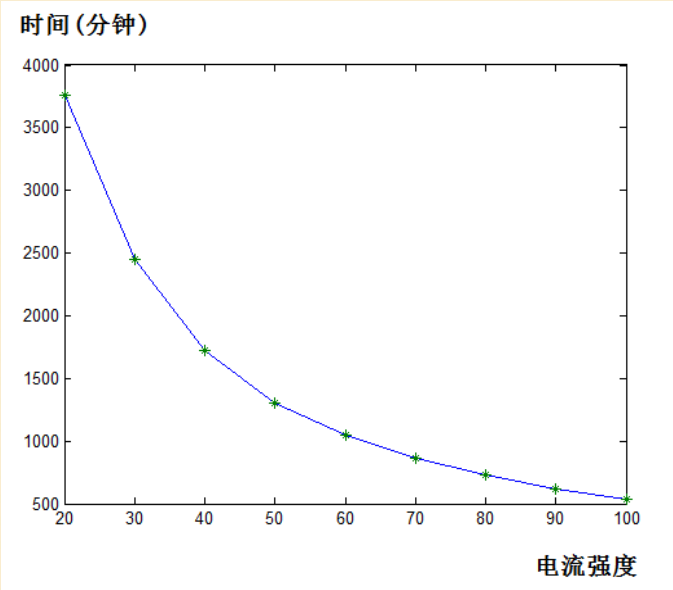


图1 不同电流强度下放电到9.0V所花时间

该图说明，随着电流强度增大，放电到9V所花时间越少，即放电越快，如图2所示。

续之

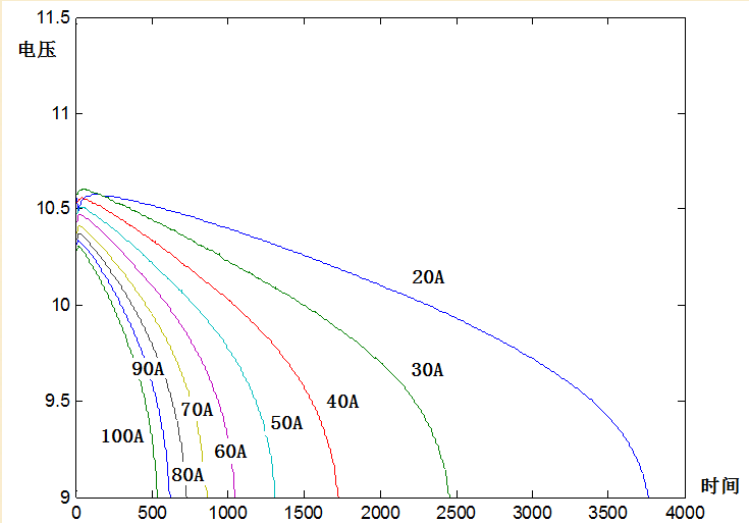


图2 不同电流强度下的放电时间曲线

思路：采用多项式分别拟合不同电流强度下放电的电压与时间的函数关系 V=V(t)

经尝试，采用四次多项式拟合。

$$U = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4$$

将数据导入SAS数据文件d2,d3,d4,d5,d6,d7,d8,d9,d10（数据文件名C2016_2,...,C2016_10），分别计算并观察多项式合适的次数，如图3、图4、图5所示。

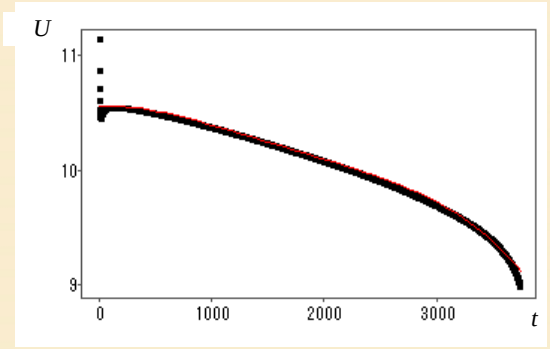


图3 I=20A时的四次函数拟合

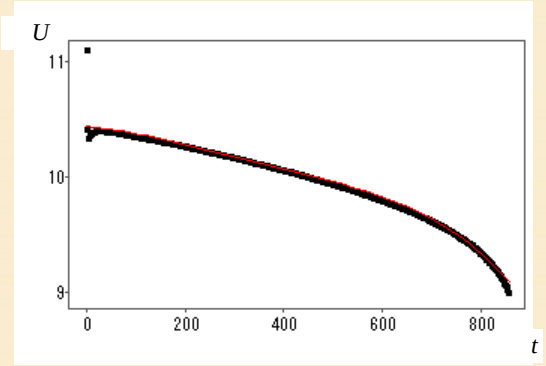


图4 I=70A时的四次函数拟合

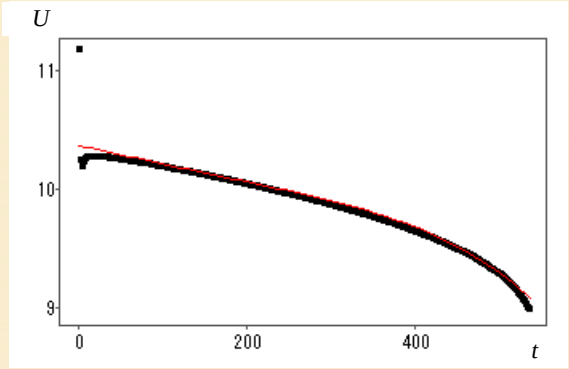


图5 I=100A时的四次函数拟合

如对 I=100A四次函数拟合为：

$$U = 10.368 - 1.4158\times10^{-3}t - 20694\times10^{-6} + a_2t^2 + 11377\times10^{-8}t^3 - 20384\times10^{-11}t^4$$

平均相对误差（MRE）的定义：从Um开始按不超过0.005V的最大间隔提取231个电压样本点。设这些电压值对应采样已放电时间为 t_i ，对应模型已放电时间为 $t_i^{'}$ ，则平均相对误差（MRE）为：

$$MRE = \frac{1}{231}\sum_{i=1}^{231}\frac{\left|t_i - t_i^{'}\right|}{t_i}$$

计算结果见表1

表1 不同电流强度下MRE

电流强度	20A	30A	40A	50A	60A	70A	80A	90A	100A
MRE	0.673%	0.570%	0.646%	0.680%	0.780%	1.051%	1.575%	3.271%	4.047%

对30A、40A、50A、60A和70A电流强度放电，测得电压都为9.8V时，估计电池的剩余放电时间。采用方法：代入拟合函数，计算U=9.8V的时间 t_1，及U=9.0V的时间 t_2，则 t_2-t_1 为电池剩余放电时间。

或估计出 t=t(U) 的函数，然后将U=9.8和U=9.0代入计算出各自放电时间，二者之差则为电池剩余放电时间。

示例:

I=70，A拟合曲线为：

$$U=-6.0802\times10^{-12}t^4+7.8095\times10^{-9}t^3-3.6765\times10^{-6}t^2-3.0913\times10^{-4}t+10.435$$

给定U时求时间t，等价于求函数的根。该函数为：

$$f(t)=-6.0802\times10^{-12}t^4+7.8095\times10^{-9}t^3-3.6765\times10^{-6}t^2-3.0913\times10^{-4}t+10.435-U$$

求导有：

$$f'(t)=-4\times6.0802\times10^{-12}t^3+3\times7.8095\times10^{-9}t^2-2\times3.6765\times10^{-6}t-3.0913\times10^{-4}$$

采用牛顿迭代法求根公式为：

$$t_{n+1} = t_n - \frac{f(t_n)}{f'(t_n)} \quad n = 1, 2, 3, \ldots\ldots$$

当U=9.8时，由于观测值为t=606 ~ 608，因此可取初始值t_0=606，经过两次迭代即得 $t_1 = 615.47$

当U=9.0时，由于观测值为t=862，因此可取初始值t_0=862，经过两次迭代即得 $t_2 = 878.32$

因此根据模型估计电池剩余放电时间为 $t_2 - t_1 = 262.85$分钟

其他电流强度下计算方法完全相同。

结果见表2。

表2 不同电流强度的剩余放电时间

电流强度	U =9.8的放电时间t_1	U=9.0的放电时间t_2	估计剩余放电时间	观测剩余放电时间	相对误差
30A	1875.79	2521.11	645.32	594.00	8.64%
40A	1304.14	1766.10	461.96	430.00	7.43%
50A	989.10	1347.77	358.67	328.00	9.35%
60A	774.61	1076.17	301.56	278.00	8.48%
70A	615.47	878.32	262.85	256.00	2.68%

这里剩余放电时间根据模型估计得到，观测剩余放电时间根据观测值得到。相对误差为

$$\gamma = \frac{|\text{观测剩余放电时间-估计剩余放电时间}|}{\text{观测剩余放电时间}} * 100$$

从结果来看，虽然估计剩余放电时间与观测剩余放电时间有一定误差，但相对误差都不超过10%，估计具有一定可靠性。

问题二的解决

思路：要建立以20A到100A之间任一恒定电流强度放电时的放电曲线的数学模型，可将电压U看作时间t和电流强度的二元函数 U=U(t, I)，然后利用多项式进行拟合，并利用MRE评估模型的精度，将电流强度I=55A代入函数计算出电压U与时间t的放电曲线。

将所有的数据点合成在一起，以(t,I)为自变量，U为因变量，总共数据点合成有6531个。

拟合函数形式采用：

$$U = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + b_1 I + b_2 I^2 + b_3 I^3 + b_4 I^4 + c_1 t \cdot I + c_2 t^2 \cdot I + c_3 t^3 I + d_1 t \cdot I^2 + d_2 t \cdot I^3 + d_4 t^2 I^2$$

生成数据文件导入SAS8，存为C2016_q2，SAS8计算结果为：

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	923.84379	65.98884	45762.8	<.0001
Error	6516	9.39591	0.00144		
Corrected Total	6530	933.23970			

Root MSE	0.03797	R-Square	0.9899
Dependent Mean	10.03532	Adj R-Sq	0.9899
Coeff Var	0.37840		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.22688	0.02797	365.65	<.0001
t	1	0.00166	0.00006012	27.68	<.0001
t2	1	-0.00000167	4.545445E-8	-36.66	<.0001
t3	1	4.09466E-10	1.14563E-11	35.74	<.0001
t4	1	-2.3046E-14	8.84107E-16	-26.07	<.0001
i	1	0.02843	0.00218	13.06	<.0001
i2	1	-0.00064261	0.00005767	-11.14	<.0001
i3	1	0.00000466	6.365985E-7	7.32	<.0001
i4	1	-1.02977E-8	2.515461E-9	-4.09	<.0001
ti	1	-0.00011180	0.00000347	-32.26	<.0001
t2i	1	8.691338E-8	2.028637E-9	42.84	<.0001
t3i	1	-1.3111E-11	3.11552E-13	-42.08	<.0001
ti2	1	0.00000168	5.587181E-8	30.13	<.0001
t2i2	1	-1.03445E-9	1.81711E-11	-56.93	<.0001
ti3	1	-7.42196E-9	2.66795E-10	-27.82	<.0001

得到模型为：

$$U(I,t)=10.22688+0.00166t-0.00000167t^2+4.09466\times10^{-10}t^3-2.3046\times10^{-14}t^4$$
$$+0.028431-0.00064261I^2+0.00000466I^3-1.02977\times10^{-8}I^4$$
$$-0.00011180t.I+8.691338\times10^{-8}t^2.I-1.311\times10^{-11}t^3I$$
$$+0.00000168t.I^2-7.42196\times10^{-9}t.I^3-1.03445\times10^{-9}t^2I^2$$

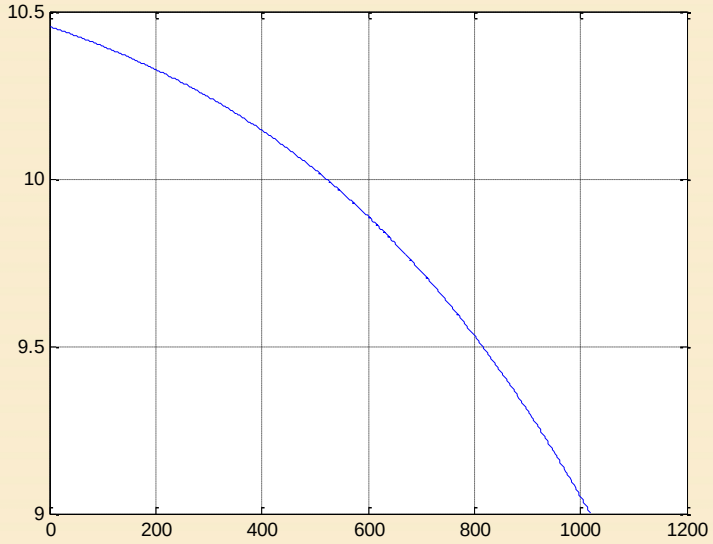


图6 I=65A的放电曲线

$$U(t)=10.45573-0.00054726t-0.00000039118t^2-4.4275\times10^{-10}t^3-2.3046\times10^{-14}t^4$$

采用牛顿迭代法，设：

$$f(t)=10.45573+0.00054726t-0.00000039118t^2-4.4275\times10^{-10}t^3-2.3046\times10^{-14}t^4-U$$

导数为：

$$f'(t)=0.00054726-2\times0.00000039118t-3\times4.4275\times10^{-10}t^2-4\times2.3046\times10^{-14}t^3$$

对U=9.8，取初始值为 $t_0=(606+766)/2=686$

其中606为I=9.8V对应I=70A的放电时间粗略估计值，766为I=9.8V对应I=60A的放电时间粗略估计值，取二者平均值作为I=65A对应9.8V的初始放电时间，是为了选取更好的初始值，使迭代速度更快。

采用牛顿迭代法

$$t_{n+1}=t_n-\frac{f(t_n)}{f'(t_n)}\quad n=1,2,3,\ldots$$

迭代2次则满足要求，得到 t_1=655.48分钟。

对U=9.0，取初始值为 $t_0=(862+1042)/2=952$

其中862为I=9.0伏对应I=70A的放电时间粗略估计值，1042为I=9.0V对应I=60A的放电时间粗略估计值。

迭代2次满足要求，得到 $t_2=1018.50$分钟

则放电时间估计为 $dt=t_2-t_1=363.01$分钟

实现程序见C2016_q2.m。

问题三的解决

思路：我们采用的方法是对新电池状态，衰减状态1，衰减状态2，衰减状态3拟合时间t关于电压U的函数 t=t(U),然后对新电池状态，衰减状态1，衰减状态2计算U=9.0时的时间，与观测时间比较，根据其误差的大小来判断该方法的有效性。如果有效，则采用该方法估计衰减状态3在U=9.0时的时间，从而估算衰减状态3的剩余放电时间。

1. 函数拟合 t=t(U)

数据点从20点开始，前面放电时间电压不稳定，舍弃不考虑。SAS8数据文件$p0$（新电池状态），$p1$（衰减状态1），$p2$（衰减状态2），$p3$（衰减状态3）。

作函数拟合 t=t(U)

经尝试，采用三次拟 合 $t = c_0 + c_1 U + c_2 U^2 + c_3 U^3$,拟合结果如图7、图8所示

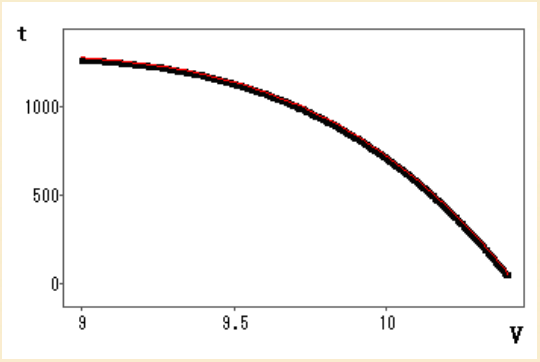


图7 新电池状态3次拟合结果
（衰减状态1和2图形完全类似）

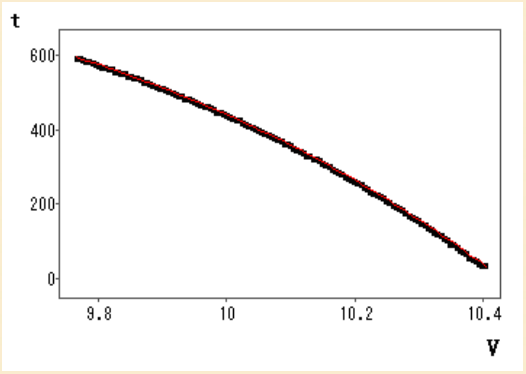


图8 衰减状态3的3次拟合结果

2. 误差估计

对新电池状态，衰减状态1，衰减状态2，比较衰减到9V的实际放电时间和模型估计时间，估计误差

设 t_0 为衰减到9V的观测时间，t_e 为衰减到9V的估计时间，绝对误差定义为：

$$\Delta t = |t_o - t_e|$$

相对误差定义为：

$$\eta = \frac{|t_o - t_e|}{t_o} \times 100\%$$

MatLab程序计算结果：

第k= 1种状态，U=9.0， 观测时间=1281.10，估计时间1282.39，绝对误差1.29，相对误差0.10%。

第k= 2种状态，U=9.0，观测时间=1104.80，估计时间1105.95，绝对误差1.15，相对误差0.10%。

第k= 3种状态，U=9.0，观测时间=979.00，估计时间978.89，绝对误差0.11，相对误差0.01%。

第k= 4种状态，U=9.0，当前时间=596.20，衰减到9.0V估计时间844.58，剩余放电时间248.38分钟。

k/V斜体，下同

计算结果见表3。

表3 放电时间比较

状　态	衰减到9V观测时间（分钟）	衰减到9V估计时间（分钟）	绝对误差	相对误差
新电池状态	1281.10	1282.39	1.29	0.10%
衰减状态1	1104.80	1105.95	1.15	0.10%
衰减状态2	979.00	978.89	0.11	0.01%

从该结果来看，最大绝对误差不超过1.29分钟，最大相对误差不超过0.1 %，说明该方法有效。

由此我们计算出，当U=9.0时，计算衰减状态3所花时间为844.58分钟。

衰减状态3当前时间为596.2分钟，从而得到剩余放电时间为：844.58-596.2=248.38分钟。

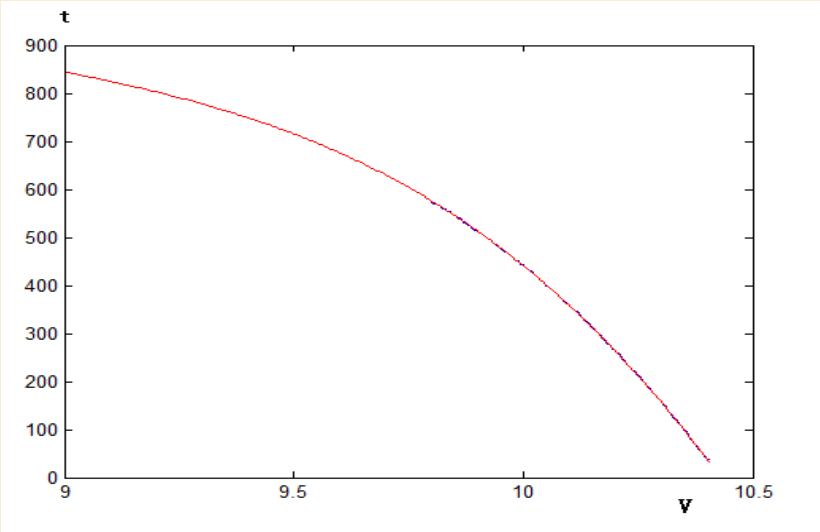


图9 衰减状态3的3次拟合及预测结果

源程序见p179-184

4.3.7 葡萄酒的评价问题

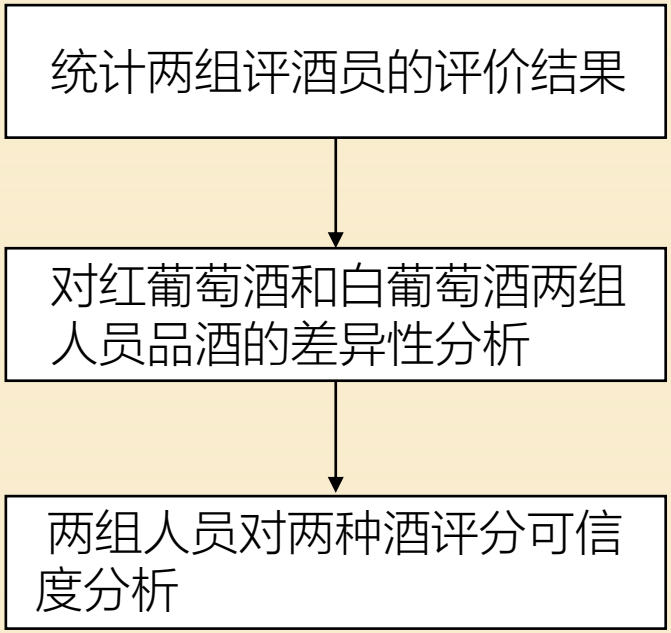
问题简介

确定一款葡萄酒质量时一般是通过聘请一批有资质的评酒员进行品评。每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到其总分，从而确定葡萄酒的质量。附件1给出了某一年份一些葡萄酒的评价结果，分析附件1中两组评酒员的评价结果有无显著性差异，哪一组结果更可信。

附件1：葡萄酒品尝评分表（含4个表格，参考CUMCM2012A附件1）

问题分析

此类问题为分析显著性差异问题，结合问题所给数据，大致求解流程如下：



1. 统计两组评酒员的评价结果

对附件1中数据，采用VBA编程序，统计27种红葡萄酒两组评酒员的得分；统计28种白葡萄酒两组评酒员的得分结果。计算时对评酒员各分类指标得分求和得到总分，从而确定葡萄酒的质量。
操作过程如下：

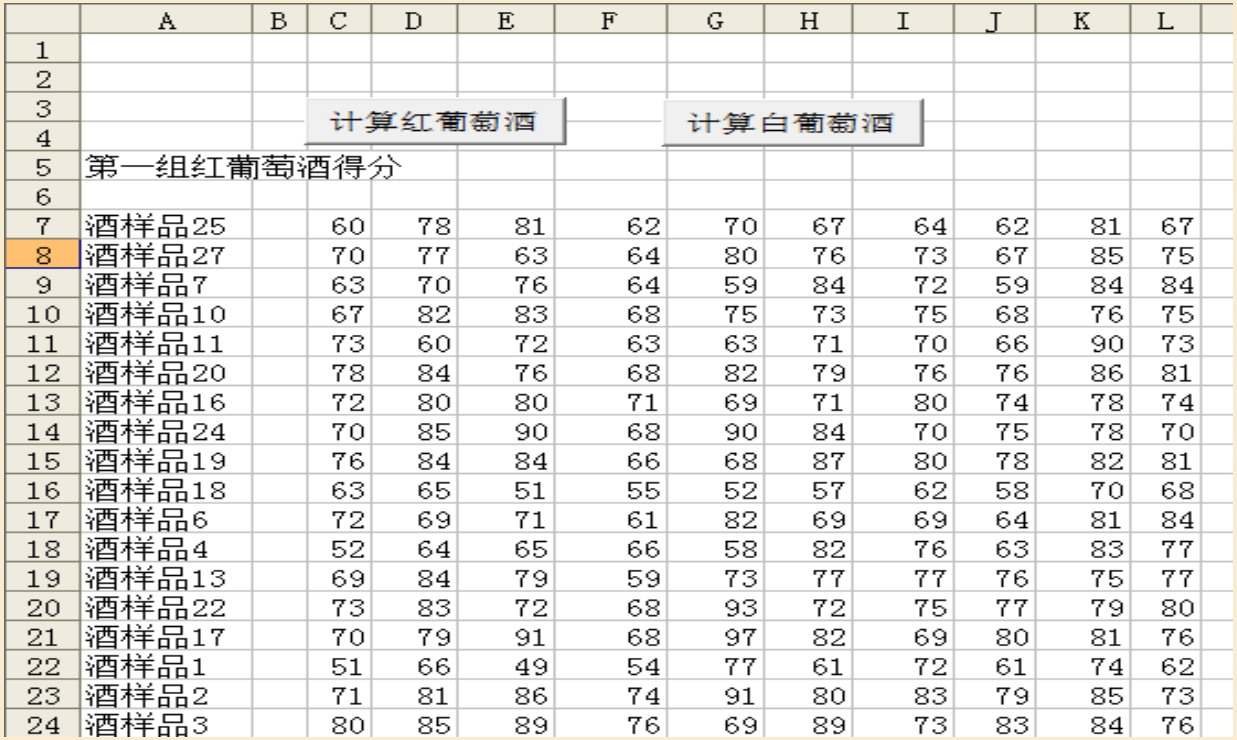
（1）打开附件1所在的xls文件，新增一个表单，命名为“计算结果”，用于存储所有结果。

（2）选中“计算结果”表单。鼠标单击【视图】→【工具栏】→【控件工具箱】 。这样在表单中就会出现控件工具箱如右图。

（3）单击控件工具箱中的“命令按钮”，表示选中该控件，然后在表单中你想放置该控件的位置单击，命令按钮就出现在该位置了。按钮上出现的名称叫“CommandButton1”，可以修改成你希望的名字。方法是将鼠标放在该按钮上用右键单击，弹出一个菜单，在菜单中选“属性”出现一个“属性”框，点Caption，命名为“计算红葡萄酒”。这样该命令按钮上的字就变为“计算红葡萄酒”。

（4）双击“计算红葡萄酒”按钮，出现对应的代码编写窗口，编写VBA代码的函数（函数代码见本书p185）

“计算结果”表单添加命令按钮和执行计算后结果如下：



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2												
3			计算红葡萄酒				计算白葡萄酒					
4												
5	第一组红葡萄酒得分											
6												
7	酒样品25		60	78	81	62	70	67	64	62	81	67
8	酒样品27		70	77	63	64	80	76	73	67	85	75
9	酒样品7		63	70	76	64	59	84	72	59	84	84
10	酒样品10		67	82	83	68	75	73	75	68	76	75
11	酒样品11		73	60	72	63	63	71	70	66	90	73
12	酒样品20		78	84	76	68	82	79	76	76	86	81
13	酒样品16		72	80	80	71	69	71	80	74	78	74
14	酒样品24		70	85	90	68	90	84	70	75	78	70
15	酒样品19		76	84	84	66	68	87	80	78	82	81
16	酒样品18		63	65	51	55	52	57	62	58	70	68
17	酒样品6		72	69	71	61	82	69	69	64	81	84
18	酒样品4		52	64	65	66	58	82	76	63	83	77
19	酒样品13		69	84	79	59	73	77	77	76	75	77
20	酒样品22		73	83	72	68	93	72	75	77	79	80
21	酒样品17		70	79	91	68	97	82	69	80	81	76
22	酒样品1		51	66	49	54	77	61	72	61	74	62
23	酒样品2		71	81	86	74	91	80	83	79	85	73
24	酒样品3		80	85	89	76	69	89	73	83	84	76

将所计算数据按样品排序后的结果见本书表4-28—表4-31

2.2对红白葡萄酒两组人员品酒的差异性分析

对红葡萄酒或白葡萄酒两组合在一起采用三因素方差分析。
A：酒样品；　B：组别；　C：品酒员
三因素方差分析：

$$SS_T = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{p}(x_{ijk}-\bar{x})^2 = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_E$$

其中SSA是酒样品因素，SSB是组别因素，SSC是品酒员因素，SSE是误差因素。

这里，我们采用SAS8进行三因素方差分析。
先利用MatLab程序a2012.m，该程序比较简单，但太长，这里没有列出，可参看光盘程序。该程序生成需要的数据文件，然后导入SAS8。其中生成的R.txt，是红葡萄酒两组人员的评价数据，导入SAS8数据名为am2012_r。生成的B.txt，是白葡萄酒两组人员的评价数据，导入SAS8数据名为am2012_b。用于三因素方差分析和单因素方差分析。
生成的 R1.txt,R2.txt分别是红葡萄酒两组人员的评价数据，导入SAS8数据名分别为am2012_r1,am2012_r2，用于双因素方差分析对红葡萄酒两组人员可信度的评价。生成的 B1.txt,B2.txt分别是白葡萄酒两组人员的评价数据，导入SAS8数据名分别为am2012_b1,am2012_b2，用于双因素方差分析对白葡萄酒两组人员可信度的评价。

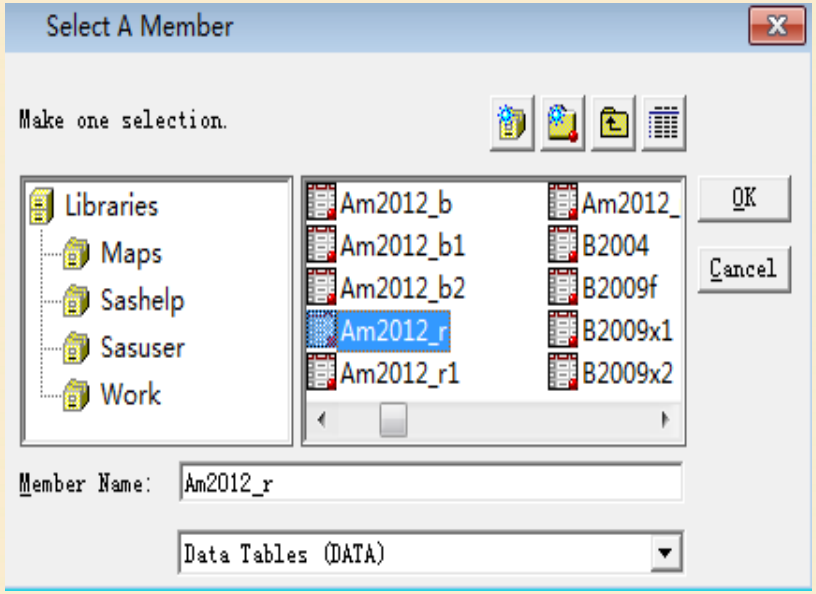
2.2对红白葡萄酒两组人员品酒的差异性分析

采用SAS8对红葡萄酒两组人员采用三因素方差分析和单因素方差分析

（1） 红葡萄酒三因素方差分析操作如下：

① 单击Solutions→Analysis→ Analyst。

② 单击File→Open By SAS Name，选中Sasuser下的数据am2012_r，然后单击OK按钮。

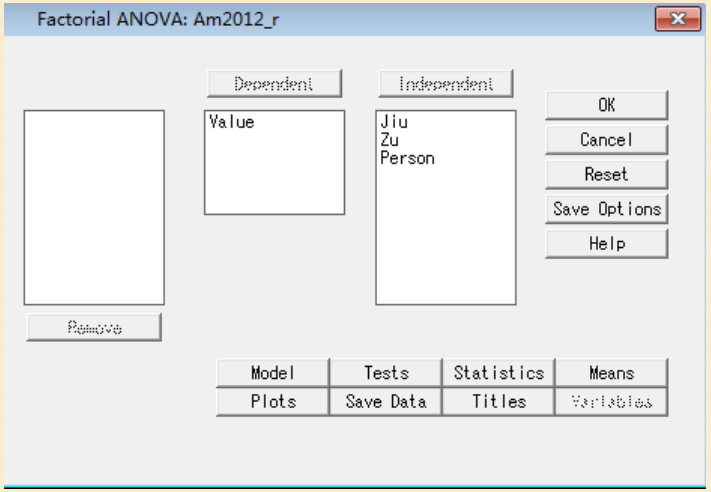


VIEWTABLE: Sasuser.Am2012_r

	Jiu	Zu	Person	Value
1	1	1	1	51
2	1	1	2	66
3	1	1	3	49
4	1	1	4	54
5	1	1	5	77
6	1	1	6	61
7	1	1	7	72
8	1	1	8	61
9	1	1	9	74
10	1	1	10	62
11	1	2	1	68
12	1	2	2	71
13	1	2	3	80
14	1	2	4	52
15	1	2	5	53
16	1	2	6	76
17	1	2	7	71
18	1	2	8	73
19	1	2	9	70
20	1	2	10	67
21	2	1	1	71
22	2	1	2	81
23	2	1	3	86
24	2	1	4	74
25	2	1	5	91
26	2	1	6	80
27	2	1	7	83
28	2	1	8	79

其中Jiu代表因素酒样品，Zu代表组别，Person代表品酒员，Value代表酒样品的得分。如Jiu为2，Zu为1,Person为3，Value为86，代表酒样品2，第一组的第3人评分为86。

③ 单击Statistics->ANOVA->Factorial ANOVA.。打开三因素方差分析对话框。将Value选入Dependent框，将Jiu、Zu、Person选入Independent框。



④ 单击OK按钮，得到结果

The GLM Procedure

Dependent Variable: Value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	36	18711.65556	519.76821	11.54	<.0001
Error	503	22660.27778	45.05025		
Corrected Total	539	41371.93333			

R-Square	Coeff Var	Root MSE	Value Mean
0.452279	9.349565	6.711949	71.78889

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Jiu	26	14344.23333	551.70128	12.25	<.0001
Zu	1	876.56296	876.56296	19.46	<.0001
Person	9	3490.85926	387.87325	8.61	<.0001

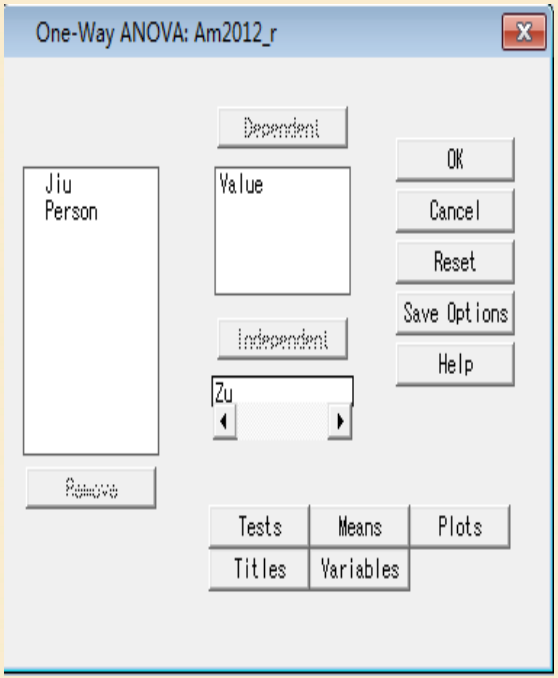
从结果看，两组人员之间F=19.46，p<0.0001，说明对红葡萄酒的评价，两组人员有显著差异。

（2）红葡萄酒两组人员单因素方差分析。

① 单击Solutions->Analysis->Analyst。

② 单击File->Open By SAS Name，选中Sasuser下的数据am2012_r，然后单击OK按钮。

③ 单击Statistics->ANOVA->One Way ANOVA，打开单因素方差分析的对话框，如图4-45所示。将Value选入Dependent框，将Zu选入Independent框。



④ 单击OK按钮，得到结果

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	1	876.56296	876.56296	11.65	0.0007
Error	538	40495.37037	75.27021		
Corrected Total	539	41371.93333			

R-Square	Coeff Var	Root MSE	Value Mean
0.021187	12.08521	8.675840	71.78889

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Zu	1	876.5629630	876.5629630	11.65	0.0007

从结果看，两组人员之间F=11.65，p=0.0007<0.01，说明对红葡萄酒，两组人员有显著差异。

对白葡萄酒也可采用三因素方差分析和单因素方差分析。操作过程与红葡萄酒相同，使用的数据集为Am2012_b。结果见表4-34和表4-35。

表4-34 白葡萄酒三因素方差分析结果

The GLM Procedure

Dependent Variable: Value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	37	19517.43571	527.49826	8.03	<.0001
Error	522	34303.30714	65.71515		
Corrected Total	559	53820.74286			

R-Square	Coeff Var	Root MSE	Value Mean
0.362638	10.76967	8.106488	75.27143

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Jiu	27	5458.34286	202.16085	3.08	<.0001
Zu	1	890.06429	890.06429	13.54	0.0003
Person	9	13169.02857	1463.22540	22.27	<.0001

从结果看，两组人员之间F=22.27，p<0.0001，说明对白葡萄酒，两组人员有显著差异。

表4-35 白葡萄酒单因素方差分析结果

The ANOVA Procedure

Dependent Variable: Value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	890.06429	890.06429	9.38	0.0023
Error	558	52930.67857	94.85785		
Corrected Total	559	53820.74286			

R-Square	Coeff Var	Root MSE	Value Mean
0.016538	12.93917	9.739499	75.27143

Source	DF	Anova SS	Mean Square	F Value	Pr >
Zu	1	890.0642857	890.0642857	9.38	0.0023

从结果看，两组人员之间F=11.65，p=0.0023<0.01，说明对白葡萄酒，两组人员有显著差异。

3. 两组人员对两种酒评分可信度分析

对红葡萄酒或白葡萄酒的每一组数据，采用可信度评价时可采用两因素方差分析。（操作过程与三因素方差分析相同）

A：酒样品；　　*B*：品酒员

两因素方差分析：

$$SS_T = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2 = SS_A + SS_B + SS_{AB} + SS_E$$

其中SS_A是酒样品因素，SS_B是品酒员因素，SS_E是误差因素。

（1）对红葡萄酒两组人员可信度的评价。

红葡萄酒第一组数据名为am2012_r1，部分数据见图4-46。

	Jiu	Person	Value
1	1	1	51
2	1	2	66
3	1	3	49
4	1	4	54
5	1	5	77
6	1	6	61
7	1	7	72
8	1	8	61
9	1	9	74
10	1	10	62
11	2	1	71
12	2	2	81
13	2	3	86
14	2	4	74
15	2	5	91
16	2	6	80
17	2	7	83
18	2	8	79
19	2	9	85
20	2	10	73
21	3	1	80
22	3	2	85
23	3	3	89
24	3	4	76

图4-46 红葡萄酒第一组部分数据

对红葡萄酒，每组采用双因素方差分析。操作过程与三因素方差分析相同，这里不再叙述。

（1）对红葡萄酒两组人员可信度的评价。

红葡萄酒第一组人员的方差分析结果如下：

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	17139.50741	489.70021	10.39	<.0001
Error	234	11030.42222	47.13856		
Corrected Total	269	28169.92963			

R-Square	Coeff Var	Root MSE	Value Mean
0.608433	9.397035	6.865752	73.06296

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Jiu	26	13965.42963	537.13191	11.39	<.0001
Person	9	3174.07778	352.67531	7.48	<.0001

从红葡萄酒第一组人员来看，酒之间的差异$F1=11.39$，$p1<0.0001$，说明红葡萄酒有差异，品酒员品酒$F2=7.48$，$p2<0.0001$，说明品酒员之间也有差异。

红葡萄酒第二组人员的方差分析结果如下：

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	7175.11481	205.00328	9.31	<.0001
Error	234	5150.32593	22.00994		
Corrected Total	269	12325.44074			

R-Square	Coeff Var	Root MSE	Value Mean
0.582139	6.653177	4.691475	70.51481

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Jiu	26	4114.340741	158.243875	7.19	<.0001
Person	9	3060.774074	340.086008	15.45	<.0001

从红葡萄酒第二组人员来看，酒之间的差异$F1=7.19$，$p1<0.0001$，说明红葡萄酒有差异，品酒员品酒$F2=15.45$，$p2<0.0001$，说明品酒员之间有差异。

酒之间的差异反映了酒的区分度，品酒员品酒差异反映了人员之间的一致性程度，采用评价指标：

$$F=\frac{F_{人}}{F_{酒}}$$

当F值越小，则评价结果越可信。对红葡萄酒的第一组人员$F1=0.6567$第二组人员，$F2=2.1488$ 则第一组红葡萄酒品酒员的评价结果更可信。

（2） 对白葡萄酒两组人员可信度的评价

第一组白葡萄酒数据集为am2012_b1

白葡萄酒第一组人员的方差分析结果如下

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	36	23369.20000	649.14444	12.39	<.0001
Error	243	12729.76786	52.38588		
Corrected Total	279	36098.96786			

R-Square	Coeff Var	Root MSE	Value Mean
0.647365	9.779407	7.237809	74.01071

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Jiu	27	6231.26786	230.78770	4.41	<.0001
Person	9	17137.93214	1904.21468	36.35	<.0001

$$F1 = \frac{F_{人}}{F_{酒}} = 36.35/4.41 = 8.2426$$

白葡萄酒第二组人员的方差分析结果如下

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	36	9439.91429	262.21984	8.62	<.0001
Error	243	7391.79643	30.41892		
Corrected Total	279	16831.71071			

R-Square	Coeff Var	Root MSE	Value Mean
0.560841	7.206560	5.515335	76.53214

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Jiu	27	2714.810714	100.548545	3.31	<.0001
Person	9	6725.103571	747.233730	24.56	<.0001

$$F2=\frac{F_{人}}{F_{酒}}=24.56/3.31=7.4199$$

对白葡萄酒，第二组人员评价指标$F2$=7.4199＜第一组人员评价指标$F2$=8.2426，故对白葡萄酒，第二组人员评价结果更可信。

总结

通过该案例的练习，可达到如下目标。

1. 学会利用VBA编程对xls文件中数据进行处理。
2. 练习利用统计软件进行三因素方差分析、双因素方差分析与单因素方差分析。
3. 掌握对多组人员评价可信度的比较。