第 11 讲　预测方法

司守奎
烟台市，海军航空工程学院数学教研室
Email：sishoukui@163.com

预测，是指根据事物发展变化或历史实际数据和资料，运用现代科学理论方法，以及相关经验、判断知识，对事物在未来一定时期内的可能变化情况进行预测、估计和分析。预测方法根据统计资料和目前信息，运用一定程序、方法和模型，分析预测对象与相关因素间的联系，以便科学地揭示预测对象的特性和变化规律，但是由于受到多方面随机因素的影响，预测的结果往往是近似的，与将来实际发生的结果存在一定的偏差。同时，也可能因为掌握的信息不够全面准确，或建模过程的简化，导致预测方法存在一定的局限性。因此，在对实际问题进行预测的过程中，要根据问题的需求选择合适的预测模型，同时要对模型的精度进行分析。

11.1 微分方程

例 11.1　认识人口数量的变化规律，建立人口模型，做出较准确的预报，是有效控制人口增长的前提。利用表 11.1 给出的近两个世纪的美国人口统计数据（以百万为单位），建立人口预测模型，最后用它预报 2010 年美国的人口。

表 11.1　美国人口统计数据

年	1790	1800	1810	1820	1830	1840	1850	1860
人口	3.9	5.3	7.2	9.6	12.9	17.1	23.2	31.4
年	1870	1880	1890	1900	1910	1920	1930	1940
人口	38.6	50.2	62.9	76.0	92.0	106.5	123.2	131.7
年	1950	1960	1970	1980	1990	2000		
人口	150.7	179.3	204.0	226.5	251.4	281.4		

1. 建模与求解

记 $x(t)$ 为第 t 年的人口数量，设人口年增长率 $r(x)$ 为 x 的线性函数，$r(x)=r-sx$。自然资源与环境条件所能容纳的最大人口数为 x_m，即当 $x=x_m$ 时，增长率 $r(x_m)=0$，可得 $r(x)=r(1-\dfrac{x}{x_m})$，建立 Logistic 人口模型

$$\begin{cases}\dfrac{dx}{dt}=r(1-\dfrac{x}{x_m})x,\\ x(t_0)=x_0,\end{cases}$$

其解为

$$x(t)=\frac{x_m}{1+(\dfrac{x_m}{x_0}-1)e^{-r(t-t_0)}}. \tag{1}$$

2. 参数估计

由于在模型(1)中参数 x_m 和 r 未知，需要先利用已知数据对参数进行估计，下面将采用四种方法对未知参数进行求解。把表 11.1 中的全部数据，包括后面的四个空格，全部保存到纯文本文件 data111.txt 中。

1）fit 函数的非线性最小二乘拟合

```
clc, clear
a=textread('data111.txt');%把原始数据保存在纯文本文件 data111.txt 中
x=a([2:2:6],:);%提出人口数据
x=x'; x=nonzeros(x);  %去掉后面的零，并变成列向量
t=[1790:10:2000]';
t0=t(1); x0=x(1);
```

```
ft=fittype('xm/(1+(xm/x0-1)*exp(-r*(t-t0)))','problem',{'t0','x0'},'independent','t')
ft=fit(t(2:end),x(2:end),ft,'problem',{1790,x0},'Start',[0.01*randn,randi([300,1000])])
xhat=ft(2010)   %预测 2010 年的人口
```

求解结果不稳定。

注：非线性拟合，由于初值随机给定，初值不合适其结果不会理想，多运行几次，找到合理的解。这里我们无论赋什么初值，都拟合不出合理的参数值。

2）lsqcurvefit 的非线性最小二乘拟合

```
clc, clear
a=textread('data111.txt'); %把原始数据保存在纯文本文件 data111.txt 中
x=a([2:2:6],:)';    %提出人口数据
x=nonzeros(x);  %去掉后面的零，并变成列向量
t=[1790:10:2000]';
t0=t(1); x0=x(1);
fun=@(cs,td)cs(1)./(1+(cs(1)/x0-1)*exp(-cs(2)*(td-t0))); %cs(1)=xm,cs(2)=r
cs=lsqcurvefit(fun,rand(2,1),t(2:end),x(2:end),zeros(2,1))
xhat=fun(cs,[t;2010])   %预测已知年代和 2010 年的人口
```

求得 $x_m=342.4368$， $r=0.0274$，2010 年人口的预测值为 282.68 百万。

3）利用向后差分，得到差分方程，用线性最小二乘法

为了利用简单的线性最小二乘法估计这个模型的参数 r 和 x_m，把 Logistic 方程表示为

$$\frac{1}{x}\cdot\frac{dx}{dt}=r-sx,\quad s=\frac{r}{x_m},$$

利用向后差分，得到差分方程

$$\frac{x_k-x_{k-1}}{\Delta t}\frac{1}{x_k}=r-sx_k,\quad k=2,3,\cdots,22,$$

其中步长 $\Delta t=10$， k 为已知观测值的编号，$k=1,2,\cdots,22$， x_k 是人口的第 k 个观测值。下面拟合其中的参数 r 和 s。编写如下的 Matlab 程序

```
clc, clear
a=textread('data111.txt'); %把原始数据保存在纯文本文件 data111.txt 中
x=a([2:2:6],:)'; x=nonzeros(x);
t=[1790:10:2000]';
a=[ones(21,1), -x(2:end)];
b=diff(x)./x(2:end)/10;
cs=a\b;
r=cs(1), xm=r/cs(2)
```

求得 $x_m=373.5135$， $r=0.0247$。

4）利用向前差分，得到差分方程，用线性最小二乘法

利用向前差分，得到的差分方程为

$$\frac{x_{k+1}-x_k}{\Delta t}\frac{1}{x_k}=r-sx_k,\quad k=1,2,\cdots,21,$$

拟合的 Matlab 程序如下

```
clc, clear
a=textread('data111.txt'); %把原始数据保存在纯文本文件 data111.txt 中
x=a([2:2:6],:)'; x=nonzeros(x);
t=[1790:10:2000]';
a=[ones(21,1), -x(1:end-1)];
b=diff(x)./x(1:end-1)/10;
cs=a\b;
r=cs(1), xm=r/cs(2)
```

求得 $x_m=294.386$， $r=0.0325$。

从上面的四种拟合方法可以看出，拟合同样的参数，方法不同可能结果相差较大。

11.2 差分方程

11.2.1 养老保险模型

例 11.2 某保险公司的一份材料指出，在每月交费 200 元至 59 岁年底，60 岁开始领取养老金的约定下，男子若 25 岁起投保，届时月养老金 2282 元；假定人的寿命为 75 岁，试求出保险公司为了兑现保险责任，每月至少应有多少投资收益率？

解 设 r 表示保险金的投资收益率，缴费期间月缴费额为 p 元，领养老金期间月领取额为 q 元，缴费的月数为 N，缴费月数和领取养老金月数的总月数为 M，投保人在投保后第 k 个月所交保险费及收益的累计总额为 F_k，那么容易得到数学模型为分段表示的差分方程

$$F_{k+1}=F_k(1+r)+p,\quad k=0,1,\cdots,N-1,$$

$$F_{k+1}=F_k(1+r)-q,\quad k=N,N+1,\cdots,M-1,$$

这里 $p=200$，$q=2282$，$N=420$，$M=600$。

可推出差分方程的解（这里 $F_0=F_M=0$）

$$F_k=[(1+r)^k-1]\frac{p}{r},\quad k=0,1,2,\cdots,N, \tag{2}$$

$$F_k=\frac{q}{r}[1-(1+r)^{k-M}],\quad k=N+1,\cdots,M. \tag{3}$$

由式(2)和式(3)得

$$F_N=[(1+r)^N-1]\frac{p}{r},$$

$$F_{N+1}=\frac{q}{r}[1-(1+r)^{N+1-M}],$$

由于 $F_{N+1}=F_N(1+r)-q$，可以得到如下的方程

$$\frac{q}{r}[1-(1+r)^{N+1-M}]+[r(1-)^{N}\frac{p}{r}+1](1+r)-q=0,$$

化简得

$$(1+r)^M-(1+\frac{q}{p})(1+r)^{M-N}+\frac{q}{p}=0,$$

记 $x=1+r$，代入数据得

$$x^{600}-12.41x^{180}+11.41=0.$$

利用Matlab程序，求得 $x=1.0049$，因而投资收益率 $r=0.49\%$。

计算的 Matlab 程序如下：

```
clc, clear
M=600; N=420; p=200; q=2282;
eq=@(x) x^M-(1+q/p)*x^(M-N)+q/p;
x=fzero(eq,[1.0001,1.5])
```

11.2.2 Leslie 种群增长模型

研究人口问题最简单和常用的 Malthus 和 Logistic 模型简单方便，对人口数量的发展变化可以给出预测。但这两类模型的两个明显的不足是：（1）仅有人口总数，不能满足需要；（2）没有考虑到社会成员之间的个体差异，即不同年龄、不同体质的人在死亡、生育方面存在的差异。完全忽略这些差异显然是不合理的。但我们不可能对每个人的情况逐个加以考虑，故可以把人口适当分组，考虑每一组人口的变化情况。年龄是一个合理的分类标准，相同年龄的人口在生育、死亡方面的可能大致接近。所以可以按年龄对人口进行分组来建立模型。在讨论其他生物的数量变化时，也可以根据生物的体重、高度、大小等因素对其分组，建立更加仔细的模型，给出更丰富的预测信息。下面介绍 Leslie 模型的一些结论。

我们以人口为例来进行叙述，其方法和思路适用于类似生物种群数量变化规律的研究。

由于男、女性人口通常有一定的比例，为了简单起见，只考虑女性人口数。现将女性人口按年龄划分成 m 个年龄组，即 $1,2,\cdots,m$ 组。每组年龄段可以是 1 岁，亦可是给定的几岁为

一组，如每 5 年为一个年龄组。现将时间也离散为时段 t_k， $k=1,2,\cdots$。

记时段 t_k 第 i 年龄组的种群数量为 $x_i(k)$，第 i 年龄组的繁殖率为 α_i；第 i 年龄组的死亡率为 d_i， $\beta_i=1-d_i$，称为第 i 年龄组的存活率。基于上述符号和假设，在已知 t_k 时段的各值后，在 t_{k+1} 时段，第一年龄组种群数量是时段 t_k 各年龄组繁殖数量之和，即

$$x_1(k+1)=\sum_{i=1}^{m}\alpha_i x_i(k),$$

t_{k+1} 时段第 $i+1$ 年龄组的种群数量是时段 t_k 第 i 年龄组存活下来的数量，即

$$x_{i+1}(k+1)=\beta_i x_i(k),\quad i=1,2,\cdots,m.$$

记 t_k 时段种群各年龄组的分布向量为

$$X(k)=\begin{bmatrix}x_1(k)\\ \vdots\\ x_m(k)\end{bmatrix},$$

并记

$$L=\begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_{n-1} & \alpha_m\\ \beta_1 & 0 & \cdots & 0 & 0\\ 0 & \beta_2 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & \beta_{m-1} & 0\end{bmatrix},$$

则有

$$X(k+1)=LX(k),\quad k=0,1,\cdots.$$

当第 t_0 时段各年龄组的人数已知时，即 $X(0)$ 已知时，可以求得 t_k 时段的按年龄组的分布向量 $X(k)$ 为

$$X(k)=L^kX(0),\quad k=0,1,\cdots.$$

由此可算出各时段的种群总量。

例 11.3 假设种群按年龄分为 5（这里 $m=5$）组，出生率向量和存活率向量分别为

$$\alpha=[\alpha_1,\cdots,\alpha_5]^T=[1.1,1.5,2.2,2.7,1.3]^T,\quad \beta=[\beta_1,\cdots,\beta_4]^T=[0.4,0.1,0.1,0.5]^T,$$

初始种群数量 $X(0)=[10,20,30,25,15]^T$，研究该种群的发展变化情况，特别要给出该种群当 $k\to+\infty$ 的极限状态。

（1）模型分析

为估计种群增长过程的动态趋势，首先研究状态转移矩阵 Leslie 矩阵的特征值和特征向量

令 $p(\lambda)$ 为 Leslie 矩阵的特征多项式，则

$$p(\lambda)=|\lambda I-L|=\lambda^m-\alpha_1\lambda^{m-1}-\alpha_2\beta_1\lambda^{m-2}-\alpha_3\beta_1\beta_2\lambda^{m-3}-\cdots-\alpha_m\beta_1\beta_2\cdots\beta_m.$$

则有下述两个结论。

定理 11.1 Leslie 矩阵 L 有唯一的正特征根 λ_1，它是单根，且相应的特征向量 v 的所有元素均为正数。

定义 11.1 设 λ_1 是方阵 L 的一个正的特征根，若对 L 的其它特征根 λ，恒有 $|\lambda|\le\lambda_1$（$|\lambda|<\lambda_1$），则称 λ_1 为 L 的占优特征根（严格占优特征根）。

定理 11.2 如果 Leslie 矩阵 L 的第一行中有两个相邻的元素 α_i 和 α_{i+1} 不为零，则 L 的正特征根是严格占优的。

于是，如果种群有两个相邻的有生育能力的年龄类，则它的 Leslie 矩阵有一个严格占优的特征根。实际上，只要年龄类的区间分得足够小，总会满足这个条件。以后总假设定理 11.2 的条件满足。

现在来研究种群年龄分布的长期性态。为使讨论简单，设 L 可对角化，且有 m 个特征根 $\lambda_1,\lambda_2,\cdots,\lambda_m$，以及对应于它们的线性无关的特征向量 $v_1,v_2,\cdots,v_m$，这些特征向量组成矩阵 $P=[v_1,v_2,\cdots,v_m]$，则 L 的对角化可由下式给出

$$L=P\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_m\end{bmatrix}P^{-1},$$

由此推得

$$L^k=P\begin{bmatrix}\lambda_1^k & & \\ & \ddots & \\ & & \lambda_m^k\end{bmatrix}P^{-1},$$

对于任何一个给定的初始年龄分布向量 $X(0)$，有

$$X(k)=L^kX(0)=P\begin{bmatrix}\lambda_1^k & & \\ & \ddots & \\ & & \lambda_m^k\end{bmatrix}P^{-1}X(0),$$

由于 λ_1 为严格占优的特征根，故 $|\lambda_i/\lambda_1|<1$，$i=2,3,\cdots,m$。从而

$$\lim_{k\to+\infty}\left(\frac{\lambda_i}{\lambda_1}\right)^k=0,\quad i=2,3,\cdots,m,$$

由此知

$$\lim_{k\to+\infty}\frac{X(k)}{\lambda_1^k}=P\begin{bmatrix}1&0&\cdots&0\\0&0&\cdots&0\\\vdots&\vdots&\vdots&\vdots\\0&0&\cdots&0\end{bmatrix}P^{-1}X(0),$$

记列向量 $P^{-1}X(0)$ 的第一个元素为 c，即 $P^{-1}X(0)=[c,*,\cdots,*]^T$，则

$$P\begin{bmatrix}1&0&\cdots&0\\0&0&\cdots&0\\\vdots&\vdots&\vdots&\vdots\\0&0&\cdots&0\end{bmatrix}P^{-1}X(0)=[v_1,v_2,\cdots,v_m]\begin{bmatrix}c\\0\\\vdots\\0\end{bmatrix}=cv_1,$$

其中，c 为常数，仅与初始年龄分布有关，则

$$\lim_{k\to+\infty}\frac{X(k)}{\lambda_1^k}=cv_1,$$

因此当 k 很大时，

$$X(k)\approx c\lambda_1^kv_1,$$

而 $X(k-1)\approx c\lambda_1^{k-1}v_1$，所以对充分大的 k，有

$$X(k)\approx\lambda_1X(k-1),$$

这意味着对于充分大的时间，每一个年龄分布向量就是它前一期年龄分布向量的 λ_1 倍。

进一步得出，对时间充分大时种群的年龄分布有三种可能情况。1）若 $\lambda_1>1$，则种群最终为增加；2）若 $\lambda_1<1$，则种群数量最终为减少；3）若 $\lambda_1=1$，则种群为稳定。

（2）算例分析

本题中 Leslie 矩阵

$$L=\begin{bmatrix}1.1&1.5&2.2&2.7&1.3\\0.4&0&0&0&0\\0&0.1&0&0&0\\0&0&0.1&0&0\\0&0&0&0.5&0\end{bmatrix},$$

L 的最大特征值为 $\lambda_1=1.5325$，对应的特征向量

$$v_1=[0.9675,\ 0.2525,\ 0.0165,\ 0.0011,\ 0.0004].$$

计算得 $c=112.21$。

因而，当 $k\to+\infty$ 时，该种群的数量趋于无穷。

计算的 Matlab 程序如下

```
clc, clear
x0=[10 20 30 25 15];
alpha=[1.1 1.5 2.2 2.7 1.3];
beta=[0.4 0.1 0.1 0.5];
a=[alpha;[diag(beta),zeros(4,1)]]; %构造系数矩阵
[vec,val]=eig(a)
cc=inv(vec)*x0'; c=cc(1)  %注意在数学上 c 一定为正，Matlab 计算时 c 未必为正，可以进行修正，因为特征向量矩阵前乘以-1 仍为特征向量矩阵。
```

11.3 马尔可夫预测

马尔科夫预测法是应用概率论中马尔科夫链的理论和方法来研究随机事件变化，并借此分析预测未来变化趋势的一种方法。所谓马尔科夫链是一种随机序列，它在将来取何值只与它现在的取值有关，而与它过去取何值无关，即无后效性，具备这个性质的离散随机过程，就是马尔科夫链。首先介绍马尔科夫链的一些基本概念及性质。

定义 11.2 设随机过程 $\{X(t), t \in T\}$ 的状态空间为 I。如果对于时间 t 的任意 n 个数值 $t_1 < t_2 < \cdots < t_n$，$t_i \in T$，$n \geq 3$，在条件 $X(t_i) = x_i$，$x_i \in I$（$i = 1, 2, \cdots, n-1$）下，$X(t_n)$ 的条件分布函数恰好等于在条件 $X(t_{n-1}) = x_{n-1}$ 下 $X(t_n)$ 的条件分布函数，则称过程具有马尔科夫性，或称此过程为马尔科夫过程。如果时间和状态都是离散的，马尔科夫过程称为马尔科夫链，记为 $\{X_n = X(n), n = 0, 1, 2, \cdots\}$。

称条件概率 $p_{ij}(m, m+n) = p(X_{m+n} = a_j \mid X_m = a_i)$ 为马尔科夫链在时刻 m 处于状态 a_i 条件下，在时刻 $m+n$ 转移到状态 a_j 的转移概率；由转移概率组成的矩阵 $P(m, m+n) = (p_{ij}(m, m+n))$ 称为马尔科夫链的转移概率矩阵。当转移概率 $p_{ij}(m, m+n)$ 只与 i, j 及时间间距 n 有关，记作 $p_{ij}(n) = p_{ij}(m, m+n)$，称此转移具有平稳性，同时称链是齐次的。在马尔科夫链是齐次的情形下，$p_{ij}(n)$ 称为马氏链的 n 步转移概率，$P(n) = (p_{ij}(n))$ 为 n 步转移概率矩阵。

定理 11.3 设 $\{X_n = X(n), n = 0, 1, 2, \cdots\}$ 是一齐次马氏链，则对任意的 $u, v \in T$，有

$$p_{ij}(u+v) = \sum_{k=1}^{\infty} p_{ik}(u) p_{kj}(v), \quad i, j = 1, 2, \cdots$$

由定理 1 容易得到一步转移概率与 n 步转移概率的关系：$P(n) = (P(1))^n$。

定义 11.3 一个有 n 个状态的马氏链如果存在正整数 m，使从任意状态 a_i 经 m 次转移都以大于零的概率到达状态 a_j，则称此为正则链。

定理 11.4 设齐次马氏链 $\{X_n, n \geq 1\}$ 的状态空间为 $I = \{a_1, a_2, \cdots, a_N\}$，$P$ 是它的一步转移概率矩阵，如果存在正整数 m，使对任意的 $a_i, a_j \in I$ 都有 $p_{ij}(m) > 0$，则此链为正则链；则正则链的极限分布 $\pi = (\pi_1, \pi_2, \cdots, \pi_N)$，是方程组 $\pi = \pi P$ 满足条件 $\pi_j > 0$，$\sum_{j=1}^{N} \pi_j = 1$ 的唯一解。

例 11.4（服务网点的设置问题）为适应日益扩大的旅游事业的需要，某城市的甲、乙、丙三个照相馆组成一个联营部，联合经营出租相机的业务。游客可由甲、乙、丙三处任何一处租出相机，用完后，还在三处中任意一处即可。估计其转移概率如**表 11.2** 所示。

（1）若第一次租相机的概率分布为 $P^{(1)} = [0.2, 0.4, 0.4]^T$，画出在甲乙丙三处租赁相机的概率随租赁次数的变化图。

（2）今欲选择其中之一附设相机维修点，问该点设在哪一个照相馆为最好？

表 11.2　状态转移矩阵

		还相机处		
		甲	乙	丙
租相机处	甲	0.2	0.8	0
	乙	0.8	0	0.2
	丙	0.1	0.3	0.6

由于旅客还相机的情况只与该次租机地点有关，而与相机以前所在的店址无关，所以可用 X_n 表示相机第 n 次被租时所在的店址；“ $X_n=1$ ”、“ $X_n=2$ ”、“ $X_n=3$ ”分别表示相机第 n 次被租用时在甲、乙、丙馆。则 $\{X_n, n=1,2,\cdots\}$ 是一个马尔可夫链，其转移矩阵 P 由上表给出。

（1）记 $p_1(n)$、$p_2(n)$、$p_3(n)$ 分别表示第 n 次在甲乙丙店租赁相机的概率，容易写出第 $n+1$ 次在甲乙丙店租赁相机的概率满足

$$\begin{cases} p_1(n+1)=0.2p_1(n)+0.8p_2(n)+0.1p_3(n), \\ p_2(n+1)=0.8p_1(n)+0.3p_3(n), \\ p_3(n+1)=0.2p_2(n)+0.6p_3(n). \end{cases}$$

记向量 $P^{(n)}=\left[p_1(n), p_2(n), p_3(n)\right]^T$，矩阵

$$P=\begin{bmatrix} 0.2 & 0.8 & 0 \\ 0.8 & 0 & 0.2 \\ 0.1 & 0.3 & 0.6 \end{bmatrix},$$

则 $P^{(n+1)}=P^TP^{(n)}$，由给定的第一次租赁相机的概率分布 $P^{(1)}$，就可以迭代出各次在甲乙丙店租赁相机的概率，概率变化趋势如图 11.1 所示。

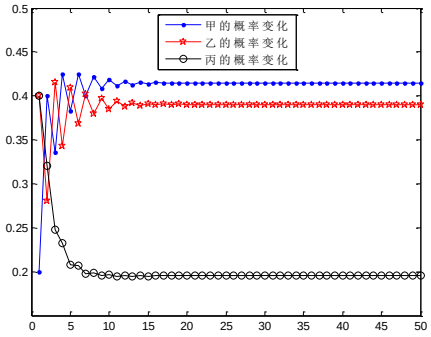

图 11.1　概率变化趋势图

计算及画图的 Matlab 程序如下：

```
clc, clear
n=50; %迭代 n 次
P(1,[1:3])=[0.2 0.4 0.4];
PP=[0.2 0.8 0; 0.8 0 0.2; 0.1 0.3 0.6];
for i=2:n
    P(i,[1:3])=P(i-1,[1:3])*PP;
end
P  %显示概率变换数据，第 1，2，3 列分别为甲乙丙的概率变化数据
plot(P(:,1),'.-')  %画甲的概率变化
hold on, plot(P(:,2),'P-r')
plot(P(:,3),'o-k')
legend('甲的概率变化','乙的概率变化','丙的概率变化','Location','north')
```

（2）考虑维修点的设置地点问题，实际上要计算这一马尔可夫链的极限概率分布。转移矩阵满足定理 11.4 的条件，极限概率存在，解方程组

$$\begin{cases} p_1 = 0.2p_1 + 0.8p_2 + 0.1p_3 \\ p_2 = 0.8p_1 + 0.3p_3 \\ p_3 = 0.2p_2 + 0.6p_3 \\ p_1 + p_2 + p_3 = 1 \end{cases}$$

得极限概率 $p_1 = \frac{17}{41}$， $p_2 = \frac{16}{41}$， $p_3 = \frac{8}{41}$。

由计算看出，经过长期经营后，该联营部的每架照相机还到甲、乙、丙照相馆的概率分别为$\frac{17}{41}$、$\frac{16}{41}$、$\frac{8}{41}$。由于还到甲馆的照相机较多，因此维修点设在甲馆较好。但由于还到乙馆的相机与还到甲馆的相差不多，若是乙的其它因素更为有利的话，比如，交通较甲方便，便于零配件的运输，电力供应稳定等等，亦可考虑设在乙馆。

计算极限分布的解线性方程组的 Matlab 程序如下：

```
format rat
p=[0.2 0.8 0; 0.8 0 0.2; 0.1 0.3 0.6];
a=[p'-eye(3);ones(1,3)];
b=[zeros(3,1);1];
p_limit=a\b
format  %恢复到短小数的数据格式
```

或者利用求状态转移概率矩阵 P 的转置矩阵 P^T 的最大特征值 1 对应的归一化特征向量，求得极限概率。编写程序如下：

```
clc, clear, format rat  %有理数的数据格式
p=[0.2 0.8 0; 0.8 0 0.2; 0.1 0.3 0.6];
[v,lamda]=eigs(p',1)  %注意矩阵转置；求模最大的特征值及对应的特征向量
v=v/sum(v)  %把特征向量归一化
format  %恢复到短小数的数据格式
```

11.4 插值与拟合

例 11.5 乡镇企业 1990-1996 年的生产利润如**表 11.3** 所示。试预测 1997 年和 1998 年的利润。

表 11.3 乡镇企业利润数据表

年份	1990	1991	1992	1993	1994	1995	1996
利润（万元）	70	122	144	152	174	196	202

作已知数据的的散点图，发现该乡镇企业的年生产利润几乎直线上升。因此，我们可以拟合线性函数 $y = a_1x + a_0$ 来预测该乡镇企业未来的年利润。编写程序如下：

```
clc, clear
x0=[1990  1991  1992  1993  1994  1995  1996];
y0=[70  122  144  152  174  196  202];
plot(x0,y0,'*')
ft=fittype('a1*x+a0'); %定义拟合的函数类
ft=fit(x0',y0',ft,'Start',rand(1,2)) %拟合函数
yhat=ft([1997,1998])
```

求得 $a_1 = 20.5$， $a_0 = -4.071\times10^4$，1997 年和 1998 年利润的预测值分别为 233.4286，253.9286。

一般地，插值不用外插（所求的插值点在所给已知点自变量最大变化区间的外部）。下面我们试着用分段线性插值做一下预测。

计算的 Matlab 程序如下：

```
clc, clear
```

```
x0=[1990  1991  1992  1993  1994  1995  1996];
y0=[70  122  144  152  174  196  202];
F=griddedInterpolant(x0,y0)  %进行分段线性插值
yhat=F([1997,1998])  %求预测值
```

求得1997年和1998年的预测值分别为208，214。通过预测值可以看出外插是利用最近两点的函数值（1995年和1996年的数据相差6）进行线性外推。

11.5 灰色预测模型

灰色预测模型是通过少量的、不完全的信息，建立数学模型并作出预测的一种方法。常用的预测方法如回归分析，需要较大容量的样本，若样本容量较小，常造成很大的误差，从而使得预测目标失效。灰色预测的主要特点是模型使用的不是原始数据序列，而是生成的数据序列。其核心体系是灰色模型（Grey Model，简称 GM），即对原始数据作累加生成（或其它方法生成）得到近似的指数规律再进行建模的方法。优点是不需要很多的数据，一般只需要 4 个数据就够，能解决历史数据少、序列的完整性及可靠性低的问题；能利用微分方程来充分挖掘系统的本质，精度高；能将无规律的原始数据进行生成得到规律性较强的生成序列，运算简便，易于检验，具有不考虑分布规律，不考虑变化趋势。缺点是只适用于中短期的预测，只适合指数增长的预测。

11.5.1 GM(1,1)预测模型

M(1,1)表示模型是 1 阶微分方程，且只含 1 个变量的灰色模型。

定义 11.4 已知参考数据列 $x^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n))$，1 次累加生成序列（1—AGO）

$$\begin{aligned}x^{(1)}&=(x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n))\\&=(x^{(0)}(1),x^{(0)}(1)+x^{(0)}(2),\cdots,x^{(0)}(1)+\cdots+x^{(0)}(n))\,,\end{aligned}$$

其中 $x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i)$ （ $k=1,2,\cdots,n$ ）。 $x^{(}$ 的均值生成序列

$$z^{(1)}=(z^{(1)}(2),z^{(1)}(3),\cdots,z^{(1)}(n))\,,$$

其中 $z^{(1)}(k)=0.5x^{(1)}(k)+0.5x^{(1)}(k-1)$， $k=2,3,\cdots n$ 。

建立灰微分方程

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

相应的白化微分方程为

$$\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b\,. \tag{4}$$

记 $u=[a,b]^T$， $Y=[x^{(0)}(2),x^{(0)}(3),\cdots,x^{(0)}(n)]^T$， $B=\begin{bmatrix}-z^{(1)}(2) & 1\\ -z^{(1)}(3) & 1\\ \vdots & \vdots\\ -z^{(1)}(n) & 1\end{bmatrix}$，则由最小二乘法，求得使 $J(u)=(Y-Bu)^T(Y-Bu)$ 达到最小值的 u 的估计值

$\hat{u}=[\hat{a},\hat{b}]^T=(B^TB)^{-1}B^TY$ 。于是求解方程（13-4）得

$$\hat{x}^{(1)}(k+1)=\left(x^{(0)}(1)-\frac{\hat{b}}{\hat{a}}\right)e^{-\hat{a}k}+\frac{\hat{b}}{\hat{a}}\,,\quad k=0,1,\cdots,n-1,\cdots.$$

GM(1,1)模型预测步骤

1. 数据的检验与处理

首先，为了保证建模方法的可行性，需要对已知数据列作必要的检验处理。设参考数据为 $x^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n))$，计算序列的级比

$$\lambda(k)=\frac{x^{(0)}(k-1)}{x^{(0)}(k)},\quad k=2,3,\cdots,n.$$

如果所有的级比 $\lambda(k)$ 都落在可容覆盖 $\Theta=(e^{-\frac{2}{n+1}},e^{\frac{2}{n+2}})$ 内，则序列 $x^{(0)}$ 可以作为模型 GM(1,1)的数据进行灰色预测。否则，需要对序列 $x^{(0)}$ 做必要的变换处理，使其落入可容覆盖内。即取适当的常数 c，作平移变换

$$y^{(0)}(k)=x^{(0)}(k)+c,\quad k=1,2,\cdots,n,$$

使序列 $y^{(0)}=(y^{(0)}(1),y^{(0)}(2),\cdots,y^{(0)}(n))$ 的级比

$$\lambda_y(k)=\frac{y^{(0)}(k-1)}{y^{(0)}(k)}\in\Theta,\quad k=2,3,\cdots,n$$

符合要求。

2. 建立模型

按（13-4）式建立 GM(1,1) 模型，则可以得到预测值

$$\hat{x}^{(1)}(k+1)=\left(x^{(0)}(1)-\frac{\hat{b}}{\hat{a}}\right)e^{-\hat{a}k}+\frac{\hat{b}}{\hat{a}},\quad k=0,1,\cdots,n-1,\cdots,$$

而且 $\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)$，$k=1,2,\cdots,n-1,\cdots$。

3. 检验预测值

（1）残差检验

令残差为 $\varepsilon(k)$，计算

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

这里 $\hat{x}^{(0)}(1)=x^{(0)}(1)$，如果 $\varepsilon(k)<0.2$，则可认为达到一般要求；如果 $\varepsilon(k)<0.1$，则认为达到较高的要求。

（2）级比偏差值检验

首先由参考数据 $x^{(0)}(k-1)$，$x^{(0)}(k)$ 计算出级比 $\lambda(k)$，再用发展系数 a 求出相应的级比偏差

$$\rho(k)=1-\left(\frac{1-0.5a}{1+0.5a}\right)\lambda(k),$$

如果 $\rho(k)<0.2$，则可认为达到一般要求；如果 $\rho(k)<0.1$，则认为达到较高的要求。

4. 预测预报

由 GM(1,1) 模型得到指定时区内的预测值，根据实际问题的需要，给出相应的预测预报。

例 11.6 某大型企业 1997—2000 年四年产值资料如表 11.4 所示，试建立 GM(1,1)预测模型，预测该企业 2001—2005 年的产值。

表 11.4 某大型企业 1997—2000 年四年产值资料

年份	1997	1998	1999	2000
产值/万元	27260	29547	32411	35388

1. 级比检验

建立原始序列数据

$$x^{(0)}=(x^{(0)}(1),x^{(0)}(2),x^{(0)}(3),x^{(0)}(4))=(27260,\ 29547,\ 32411,\ 35388).$$

（1）求级比 $\lambda(k)$

$$\lambda(k)=\frac{x^{(0)}(k-1)}{x^{(0)}(k)},$$

$$\lambda=(\lambda(2),\lambda(3),\lambda(4))=(\ 0.9226,\quad 0.9116,\quad 0.9159).$$

（2）级比判断

由于所有的 $\lambda(k)\in[0.6703,\ 1.3956]$，$k=2,3,4$，故可以用 $x^{(0)}$ 作满意的 GM(1,1)建模。

2. GM(1,1)建模

（1）对原始数据 $x^{(0)}$ 作一次累加，得到

$$x^{(1)}=(27260,\ 56807,\ 89218,\ 124606)$$

（2）构造数据矩阵 B 及数据向量 Y

$$B=\begin{bmatrix}-\frac{1}{2}(x^{(1)}(1)+x^{(1)}(2)) & 1\\ -\frac{1}{2}(x^{(1)}(2)+x^{(1)}(3)) & 1\\ -\frac{1}{2}(x^{(1)}(3)+x^{(1)}(4)) & 1\end{bmatrix},\quad Y=\begin{bmatrix}x^{(0)}(2)\\ x^{(0)}(3)\\ x^{(0)}(4)\end{bmatrix}.$$

（3）计算

$$\hat{u}=\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix}=(B^TB)^{-1}B^TY=\begin{bmatrix}-0.089995\\ 25790.2838\end{bmatrix},$$

于是得到 $\hat{a}=-0.089995$，$\hat{b}=25790.2838$。

（4）建立模型

$$\frac{dx^{(1)}}{dt}+\hat{a}x^{(1)}=\hat{b},$$

求解得

$$\hat{x}^{(1)}(k+1)=\left(x^{(0)}(1)-\frac{\hat{b}}{\hat{a}}\right)e^{-\hat{a}k}+\frac{\hat{b}}{\hat{a}}=313834e^{0.089995k}-286574.\qquad(13\text{-}5)$$

（5）求生成序列预测值 $\hat{x}^{(1)}(k+1)$ 及模型还原值 $\hat{x}^{(0)}(k+1)$，令 $k=1,2,3$，由（13-5）式的时间响应函数可算得 $\hat{x}^{(1)}$，其中取 $\hat{x}^{(1)}(1)=x^{(0)}(1)=$，由 $\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)$，取 $k=1,2,3$，得

$$\hat{x}^{(0)}=(\hat{x}^{(0)}(1),\hat{x}^{(0)}(2),\hat{x}^{(0)}(3),\hat{x}^{(0)}(4))=(27260,\ 29553.4421,\ 32336.4602,\ 35381.5524)$$

3. 模型检验

模型的各种检验指标值的计算结果见表 11.5。经验证，该模型的精度较高，可进行预测和预报。

表 11.5　GM(1,1)模型检验表

年份	原始值	预测值	残差	相对误差	级比偏差
1997	27260	27260	0	0	
1998	29547	29553.4421	-6.4421	0.0002	-0.0095
1999	32411	32336.4602	74.5398	0.0023	0.0025
2000	35388	35381.5524	6.4476	0.0002	-0.0022

4.预测值

2001—2005 年的预测值见表 11.6。

表 11.6　2001—2005 年的预测值

年份	2001	2002	2003	2004	2005
产值/万元	38713.3978	42358.9998	46347.9045	50712.4404	55487.9803

计算的 Matlab 程序如下

```
clc,clear
x0=[27260    29547    32411    35388]'; %注意这里为列向量
n=length(x0);
lamda=x0(1:n-1)./x0(2:n)   %计算级比
range=minmax(lamda')   %计算级比的范围
theta=[exp(-2/(n+1)),exp(2/(n+2))]   % 计算级比的容许区间
x1=cumsum(x0)   %累加运算
```

```
B=[-0.5*(x1(1:n-1)+x1(2:n)),ones(n-1,1)];
Y=x0(2:n);
u=B\Y   %拟合参数 u(1)=a,u(2)=b
syms x(t)
x=dsolve(diff(x)+u(1)*x==u(2),x(0)==x0(1)); %求微分方程的符号解
xt=vpa(x,6) %以小数格式显示微分方程的解
yuce1=subs(x,t,[0:n+4]); %求已知数据和未来 5 期的预测值
yuce1=double(yuce1); %符号数转换成数值类型，否则无法作差分运算
yuce=[x0(1),diff(yuce1)]   %差分运算，还原数据
epsilon=x0'-yuce(1:n)      %计算已知数据预测的残差
delta=abs(epsilon./x0')   %计算相对误差
rho=1-(1-0.5*u(1))/(1+0.5*u(1))*lamda'   %计算级比偏差值，u(1)=a
yhat=yuce(n+1:end)   %提取未来 5 期的预测值
xlswrite('data136.xls',[x0,yuce(1:n)',epsilon',delta',[0,rho]'])
```

11.5.2 GM(2,1)、DGM 和 Verhulst 模型

GM(1,1)模型适用于具有较强指数规律的序列，只能描述单调的变化过程，对于非单调的摆动发展序列或有饱和的 S 形序列，可以考虑建立 GM(2,1)，DGM 和 Verhulst 模型。下面我们只介绍 GM(2,1)模型。

定义 11.5 设原始序列

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

其 1 次累加生成序列（1－AGO） $x^{(1)}$ 和 1 次累减生成序列（1－IAGO） $\alpha^{(1)}x^{(0)}$ 分别为

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

和

$$\alpha^{(1)}x^{(0)}=(\alpha^{(1)}x^{(0)}(2),\cdots,\alpha^{(1)}x^{(0)}(n))\,,$$

其中

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

$x^{(1)}$ 的均值生成序列为

$$z^{(1)}=(z^{(1)}(2),z^{(1)}(3),\cdots,z^{(1)}(n))\,,$$

则称

$$\alpha^{(1)}x^{(0)}(k)+a_1x^{(0)}(k)+a_2z^{(1)}(k)=b$$

为 GM(2,1)模型。

定义 11.6 称

$$\frac{d^2x^{(1)}}{dt^2}+a_1\frac{dx^{(1)}}{dt}+a_2x^{(1)}=b$$

为 GM(2,1)模型的白化方程。

定理 11.5 设 $x^{(0)}$，$x^{(1)}$，$\alpha^{(1)}x^{(0)}$ 如定义 13.4 所述，且

$$B=\begin{bmatrix}-x^{(0)}(2) & -z^{(1)}(2) & 1\\ -x^{(0)}(3) & -z^{(1)}(3) & 1\\ \vdots & \vdots & \vdots\\ -x^{(0)}(n) & -z^{(1)}(n) & 1\end{bmatrix},\quad Y=\begin{bmatrix}\alpha^{(1)}x^{(0)}(2)\\ \alpha^{(1)}x^{(0)}(3)\\ \vdots\\ \alpha^{(1)}x^{(0)}(n)\end{bmatrix}=\begin{bmatrix}x^{(0)}(2)-x^{(0)}(1)\\ x^{(0)}(3)-x^{(0)}(2)\\ \vdots\\ x^{(0)}(n)-x^{(0)}(n-1)\end{bmatrix},$$

则 GM(2,1)模型参数序列 $u=[a_1,a_2,b]^T$ 的最小二乘估计为

$$\hat{u}=(B^TB)^{-1}B^TY\,.$$

例 11.7 已知 $x^{(0)}=(41,49,61,78,96,104)$，试建立 GM(2,1)模型。

解 $x^{(0)}$ 的 1－AGO 序列 $x^{(1)}$ 和 1－IAGO 序列 $\alpha^{(1)}x^{(0)}$ 分别为

$$x^{(1)}=(41,90,151,229,325,429)\,,$$

$$\alpha^{(1)}x^{(0)}=(8,12,17,18,8)\,,$$

$x^{(1)}$ 的均值生成序列

$$z^{(1)} = (65.5, 120.5, 190, 277, 377)$$

$$B = \begin{bmatrix} -x^{(0)}(2) & -z^{(1)}(2) & 1 \\ -x^{(0)}(3) & -z^{(1)}(3) & 1 \\ \vdots & \vdots & \vdots \\ -x^{(0)}(6) & -z^{(1)}(6) & 1 \end{bmatrix} = \begin{bmatrix} -49 & -65.5 & 1 \\ -61 & -120.5 & 1 \\ -78 & -190 & 1 \\ -96 & -277 & 1 \\ -104 & -377 & 1 \end{bmatrix},$$

$$Y = [8, 12, 17, 18, 8]^T,$$

$$\hat{u} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{b} \end{bmatrix} = (B^T B)^{-1} B^T Y = \begin{bmatrix} -1.0922 \\ 0.1959 \\ -31.7983 \end{bmatrix},$$

故得 GM(2,1)白化模型

$$\frac{d^2 x^{(1)}}{dt^2} - 1.0922 \frac{dx^{(1)}}{dt} + 0.1959 x^{(1)} = -31.7983.$$

利用边界条件 $x^{(1)}(1) = 41$， $x^{(1)}(6) = 429$，解之得

$$x^{(1)}(t) = 203.85 e^{0.22622t} - 0.5325 e^{0.86597t} - 162.317,$$

于是 GM(2,1)时间响应式

$$\hat{x}^{(1)}(k+1) = 203.85 e^{0.22622k} - 0.5325 e^{0.86597k} - 162.317.$$

所以

$$\hat{x}^{(1)} = (41, 92, 155, 232, 325, 429).$$

做 IAGO 还原，有

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k),$$

$$\hat{x}^{(0)} = (41, 51, 63, 77, 92, 104).$$

计算结果见表 11.7。

表 11.7　误差检验表

序号	实际数据 $x^{(0)}$	预测数据 $\hat{x}^{(0)}$	残差 $x^{(0)} - \hat{x}^{(0)}$	相对误差 Δ_k
2	49	51	−2	4.1%
3	61	63	−2	3.3%
4	78	77	1	1.3%
5	96	92	4	4.2%
6	104	104	0	0

计算的 Matlab 程序如下

```
clc,clear
x0=[41,49,61,78,96,104]; %原始序列
n=length(x0);
x1=cumsum(x0)   %计算1次累加序列
a_x0=diff(x0)' %计算1次累减序列
z=0.5*(x1(2:end)+x1(1:end-1))'; %计算均值生成序列
B=[-x0(2:end)',-z,ones(n-1,1)];
u=B\a_x0     %最小二乘法拟合参数
syms x(t)
x=dsolve(diff(x,2)+u(1)*diff(x)+u(2)*x==u(3),x(0)==x1(1),x(5)==x1(6)); %求符号解
xt=vpa(x,6) %显示小数形式的符号解
yuce=subs(x,t,0:n-1); %求已知数据点1次累加序列的预测值
yuce=double(yuce) %符号数转换成数值类型，否则无法作差分运算
x0_hat=[yuce(1),diff(yuce)]; %求已知数据点的预测值
x0_hat=round(x0_hat) %四舍五入取整数
epsilon=x0-x0_hat       %求残差
```

delta=abs(epsilon./x0)　%求相对误差

11.6　回归分析

11.6.1　多元线性回归

1.多元线性回归的数学模型

设随机变量 y 与一组（k 个）变量 $x_1,\cdots,x_k$ 有关系式

$$y=b_0+b_1x_1+\cdots+b_kx_k+\varepsilon,$$

其中 $\varepsilon\sim N(0\sigma^2$ 。取一个容量为 n 的子样 $(y_1,x_{11},\cdots,x_{k1})$，…，$(y_n,x_{1n},\cdots,x_{kn})$，则有 $y_i=\sum_{j=1}^{k}b_jx_{ji}+\varepsilon_i$，$i=1,2,\cdots,n$，$b_1,b_2,\cdots,b_k$ 是未知参数，且假定 $\varepsilon_i\sim N(0,\sigma^2)$，$E(\varepsilon_i\varepsilon_j)=0$，$i,j=1,2,\cdots,n$，$i,j=1,2,\cdots,n$，$i\neq j$，$\sigma^2$ 未知。

2.最小二乘估计与正规方程组

令 $Q=\sum_{i=1}^{n}[y_i-(b_0+b_1x_{i1}+\cdots+b_kx_{ki})]^2$，两边分别关于 $b_0,b_1,\cdots,b_k$ 求偏导并令其为零，整理后得到正规方程组

$$\begin{cases}nb_0+\sum_{i=1}^{n}x_{1i}b_1+\cdots+\sum_{i=1}^{n}x_{ki}b_k=\sum_{i=1}^{n}y_i,\\ \sum_{i=1}^{n}x_{1i}b_0+\sum_{i=1}^{n}x_{1i}^2b_1+\cdots+\sum_{i=1}^{n}x_{ki}x_{1i}b_k=\sum_{i=1}^{n}y_ix_{1i},\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ \sum_{i=1}^{n}x_{ki}b_0+\sum_{i=1}^{n}x_{1i}x_{ki}b_1+\cdots+\sum_{i=1}^{n}x_{ki}^2b_k=\sum_{i=1}^{n}y_ix_{ki}.\end{cases}$$

若记

$$X=\begin{bmatrix}1&x_{11}&x_{12}&\cdots&x_{1k}\\1&x_{21}&x_{22}&\cdots&x_{2k}\\\vdots&\vdots&\vdots&&\vdots\\1&x_{n1}&x_{n2}&\cdots&x_{nk}\end{bmatrix},\quad Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},\quad B=\begin{bmatrix}b_0\\b_1\\\vdots\\b_k\end{bmatrix},$$

则正规方程组可写成 $X^TXB=X^TY$，若矩阵 X^TX 是满秩的，则其解为 $\hat{B}=(X^TX)^{-1}X^TY$，从而对 $x_1,x_2,\cdots,x_k$ 的线性回归方程为 $\hat{y}=\hat{b}_0+\hat{b}_1x_1+\cdots+\hat{b}_kx_k$。

3.整体线性相关性的 F 检验

首先提出假设：H_0：$b_1=b_2=\cdots=b_k=0$；

选取检验统计量：$F=\dfrac{U/k}{Q/(n-k-1)}$，其中 $U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，若 H_0 真，则 $F\sim F(k,n-k-1)$。

对于给定的显著性水平 α，确定临界值 $F_\alpha(k,n-k-1)$，并与检验统计量 F 的观察值比较。最后给出判断：若 $F>F_\alpha$，则拒绝 H_0，即认为 $x_1,x_2,\cdots,x_k$ 对 y 具有显著的线性影响；若 $F<F_\alpha$，则接受 H_0，即认为 $x_1,x_2,\cdots,x_k$ 对 y 不具有显著的线性影响。

4.每个变量显著性的 t 检验

首先提出假设：H_j：$b_j=0$，$1\le j\le k$；

选取检验统计量：$T_j=\dfrac{b_j/\sqrt{c_{jj}}}{\sqrt{Q/(n-k-1)}}$，其中 c_{jj} 是矩阵 $(X^TX)^{-1}$ 的主对角线上的第 $j+1$ 个元素，若 H_j 为真，则 $T_j\sim t(n-k-1)$。

对给定的 α，若 $|T_j|<t_{\alpha/2}(n-k-1)$，则接受 H_j，即认为 x_j 对 y 的影响不显著；若 $|T_j|>t_{\alpha/2}(n-k-1)$，则拒绝 H_j，即认为 x_j 对 y 的影响显著。

11.6.2 逐步回归分析

建立模型一般只要使用3~5个自变量就可以了。如果模型中的自变量个数太多，所建的模型的稳定性等方面的性质都会很差的。

在Matlab统计工具箱中用作逐步回归的命令是stepwise，它提供了一个交互式画面，通过这个工具你可以自由地选择变量，进行统计分析，其通常用法是：

stepwise(x,y,inmodel,alpha)

其中x是自变量数据，y是因变量数据，分别为 $n\times m$ 和 $n\times 1$ 矩阵，inmodel是矩阵x的列数的指标，给出初始模型中包括的子集（缺省时设定为空），alpha为显著性水平。

Stepwise Regression 窗口，显示回归系数及其置信区间，和其它一些统计量的信息。绿色表明在模型中的变量，红色表明从模型中移去的变量。在这个窗口中有Export按钮，点击Export产生一个菜单，表明了要传送给Matlab工作区的参数，它们给出了统计计算的一些结果。

下面通过一个例子说明stepwise的用法。

例 11.8 某产品的销售额 y 与部门的全部市场销售额 x_1，给批发商的优惠 x_2，价格 x_3，开发预算 x_4，投资 x_5，广告 x_6，销售费用 x_7，部门全部广告的预算 x_8 有关。为预测未来的销售量，收集了 38 个样本点的有关数据见表 11.8，试建立 y 的经验公式。

表 11.8 原始数据表

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y
398	138	56.205	12.112	49.895	76.862	228.90	98.205	5540.39
369	118	59.044	9.330	16.595	88.805	177.45	224.953	5439.04
268	129	56.723	28.748	89.182	51.297	166.40	263.032	4290.00
484	111	57.862	12.891	106.738	39.747	285.05	320.928	5502.34
394	146	59.117	13.381	142.552	51.651	209.30	406.989	4871.77
332	140	60.111	11.085	61.287	20.547	180.05	246.996	4708.08
336	136	59.839	24.957	-30.385	40.153	213.20	328.436	4627.81
383	104	60.052	20.809	-44.586	31.645	200.85	298.456	4110.24
285	105	63.141	8.485	-28.373	12.457	176.15	218.110	4122.69
277	135	62.302	10.730	75.723	68.307	174.85	410.467	4842.25
456	128	64.922	21.874	144.030	52.453	252.85	93.006	5740.65
355	131	64.857	23.506	112.904	76.677	208.00	307.226	5094.10
364	120	63.591	13.894	128.347	96.067	195.00	106.792	5383.20
320	147	65.614	14.865	10.097	47.979	154.05	304.921	4888.17
311	143	67.022	22.494	-24.760	27.231	180.70	59.612	4003.13
362	145	66.904	23.269	116.748	72.668	219.70	238.986	4941.96
408	131	66.184	13.035	120.406	62.312	234.65	141.074	5312.80
433	124	67.865	8.033	121.823	24.712	258.05	290.832	5139.87
359	106	68.889	27.048	71.055	73.912	196.30	413.636	5397.36
476	138	71.417	18.220	4.186	63.273	278.85	206.454	5149.47
415	148	69.277	7.742	46.935	28.676	207.35	79.566	5150.83
420	136	69.733	10.136	7.621	91.363	213.20	428.982	4989.02
536	111	73.162	27.370	127.509	74.016	296.40	273.072	5926.86
432	152	73.365	15.528	-49.574	16.162	245.05	309.422	4703.88
436	123	73.050	32.491	100.098	42.998	275.60	280.139	5365.59
415	119	74.910	19.712	-40.183	41.134	211.25	314.548	4630.09
462	112	73.200	14.835	68.153	92.518	282.75	212.058	5711.86
429	125	74.161	11.369	87.963	83.287	217.75	118.065	5095.48
517	142	74.283	26.751	27.098	74.892	306.90	344.553	6124.37
328	123	77.140	19.603	59.343	87.510	210.60	140.872	4787.34
418	135	78.591	34.688	141.969	74.471	269.75	82.855	5035.62
515	120	77.093	23.202	126.420	21.271	328.25	398.425	5288.01
412	149	78.231	35.739	29.558	26.494	258.05	124.027	4647.01
455	126	77.929	21.589	18.007	94.631	232.70	117.911	5315.63
554	138	81.039	19.569	42.352	92.544	323.70	161.250	6180.06
441	120	79.848	15.503	-21.558	50.048	267.15	405.088	4800.97
417	120	80.639	34.923	148.450	83.180	257.40	110.740	5512.13
461	132	82.284	26.549	-17.584	91.221	266.50	170.392	5272.21

把表11.8中的全部数据保存在纯文本文件data118.txt中。下面用两种方式建立逐步回归线性模型。

1.初始时选入全部自变量

```
clc,clear
a=load('data118.txt');
x=a(:,[1:8]); %提取 x 的数据
y=a(:,end);   %提取 y 的数据
stepwise(x,y,[1:8])   %初始时选入全部自变量
```

运行上述程序，得到图11.2所示的图形界面。然后点击“All Steps”剔除所有不显著的变量，得到图11.3所示的界面。

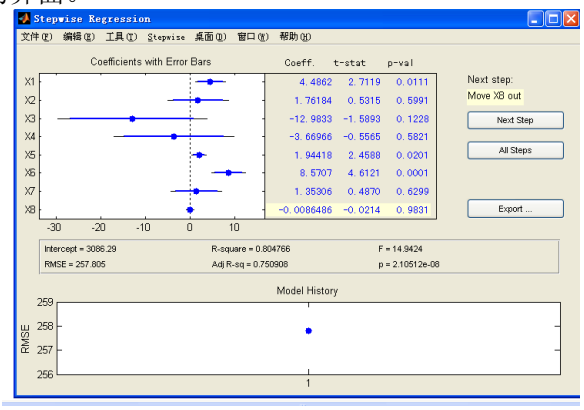

图 11.2　逐步回归交互式画面

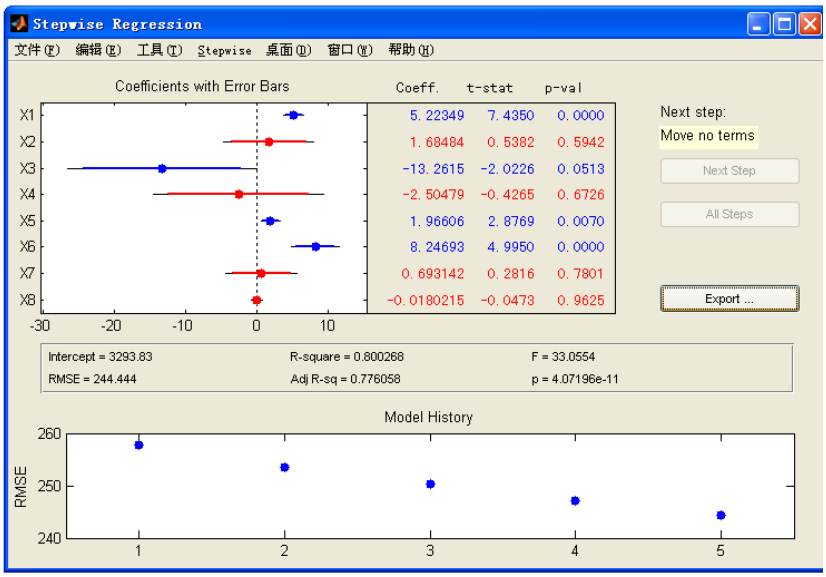

图 11.3　逐步回归计算结果一

从图11.3可以看出，最后只剩下变量 x_1, x_3, x_5, x_6 用于建立线性回归模型，回归模型为

$$y = 3293.83 + 5.22349x_1 - 13.2615x_3 + 1.96605x_5 + 8.24693x_6$$

模型的检验指标如下： F 统计量为33.0554，相关系数的平方 R^2 为0.800268，剩余标准差RMSE为244.444。

2.初始时变量集合为空集

```
clc,clear
a=load('data118.txt');
x=a(:,[1:8]); %提取x的数据
y=a(:,end);   %提取y的数据
stepwise(x,y) %初始时变量集合为空集
```

类似地，最后的求解结果如图 11.4 所示。

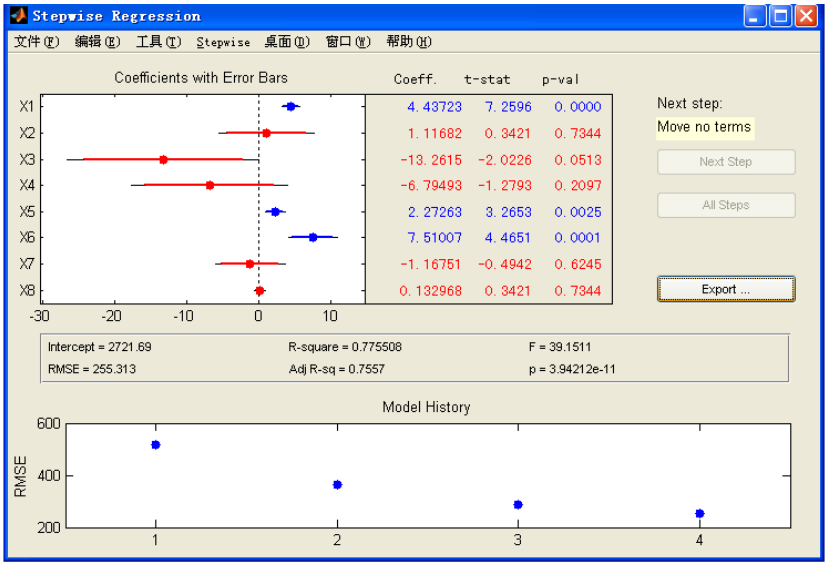

图 11.4　逐步回归计算结果二

从图11.4可以看出，最后只剩下变量 x_1, x_5, x_6 用于建立线性回归模型，回归模型为

$$y = 2721.69 + 4.43723x_1 + 2.27263x_5 + 7.51007x_6$$

模型的检验指标如下： F 统计量为 39.1511，相关系数的平方 R^2 为 0.775508，剩余标准差 RMSE 为 255.313。

也可以使用逐步线性回归的命令 stepwisefit 直接计算，使用 stepwisefit 可以看到变量剔除的先后次序（初始时选入全部自变量），或看到变量加入的先后次序（初始时自变量集为空集）。

计算的 Matlab 程序如下：

```
clc,clear
a=load('data118.txt');
x=a(:,[1:8]); %提取 x 的数据
y=a(:,end);   %提取 y 的数据
b1=stepwisefit(x,y,'inmodel',[1:8]) %初始时选入全部 8 个自变量
b2=stepwisefit(x,y) %初始时自变量集合为空
```

习题 11

11.1　1949 年—1994 年我国人口数据资料如表 11.9，试建模预测 1999 年、2005 年我国人口数量。

表 11.9　我国人口数据资料（单位：亿）

年份	49	54	59	64	69	74	79	84	89	94
人口	5.4	6.0	6.7	7.0	8.1	9.1	9.8	10.3	11.3	11.8

11.2　1960-2005 年美国出口额数据见表 11.10，试建模预测 2006—2010 年美国的出口额。

表 11.10　1960—2005 年美国出口额数据

序号	1	2	3	4	5	6	7	8	9	10
数据	19.65	20.108	20.781	22.272	25.501	26.461	29.31	30.666	33.626	36.414
序号	11	12	13	14	15	16	17	18	19	20
数据	42.469	43.319	49.381	71.410	98.306	107.088	114.745	120.816	142.075	184.439
序号	21	22	23	24	25	26	27	28	29	30
数据	224.25	237.044	211.157	201.799	219.926	215.915	223.344	250.208	320.23	359.916

序号	31	32	33	34	35	36	37	38	39	40
数据	387.401	414.083	439.631	456.943	502.859	575.204	612.113	678.366	670.416	683.965
序号	41	42	43	44	45	46				
数据	771.994	718.712	682.422	713.415	807.516	894.631				

11.3　某大型企业 1999 年至 2004 年的产品销售额如下表，试建立 GM(1,1)预测模型，并预测 2005 年的产品销售额。

表 11.11　产品销售额数据

年份	1999	2000	2001	2002	2003	2004
销售额（亿元）	2.67	3.13	3.25	3.36	3.56	3.72