第六章　数据拟合与函数逼近

第六章　数据拟合与函数逼近

- 数据拟合的最小二乘法
- 正交多项式
- 最佳平方逼近

6.1 数据拟合的最小二乘法

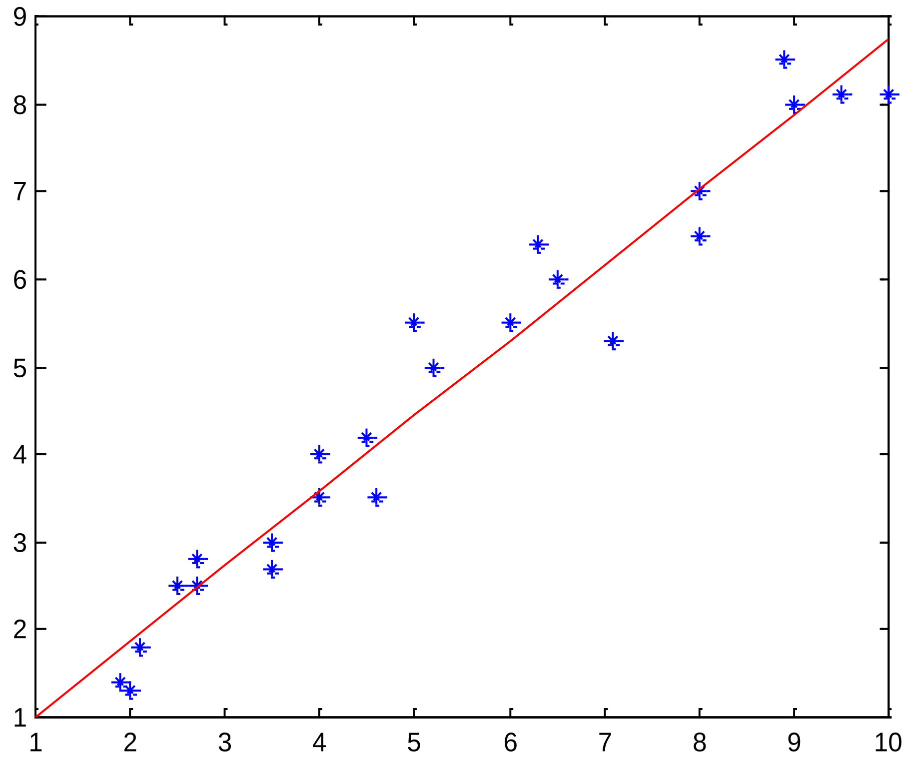
一、 曲线拟合的数学描述与问题求解

例：考察某种纤维的强度与其拉伸倍数的关系,下表是实际测定的24个纤维样品的强度与相应的拉伸倍数的记录:

编 号	拉伸倍数 x_i	强 度 y_i	编 号	拉伸倍数 x_i	强 度 y_i
1	1.9	1.4	13	5	5.5
2	2	1.3	14	5.2	5
3	2.1	1.8	15	6	5.5
4	2.5	2.5	16	6.3	6.4
5	2.7	2.8	17	6.5	6
6	2.7	2.5	18	7.1	5.3
7	3.5	3	19	8	6.5
8	3.5	2.7	20	8	7
9	4	4	21	8.9	8.5
10	4	3.5	22	9	8
11	4.5	4.2	23	9.5	8.1
12	4.6	3.5	24	10	8.1

纤维强度随拉伸倍数增加而增加.

24个点大致分布在一条直线附近.



故可认为强度y与拉伸倍数x的主要关系应为线性关系：

$$y(x)=\beta_0+\beta_1 x$$

其中β_0,β_1为待定参数.

我们希望$y(x)=\beta_0+\beta_1 x$与所有的数据点(样本点)(x_i, y_i)越接近越好.

必须找到一种度量标准来衡量什么曲线最接近所有数据点.

1、数据拟合问题

研究内容：从一大堆看上去杂乱无章的数据中找出规律性来，即设法构造一条曲线（拟合曲线）反映所给数据点总的趋势，以消除其局部波动.这种要求曲线尽可能逼近给定数据的过程称“拟合”.

给定一组值：

x	x_1	x_2 ……	x_m
$f(x)$	y_1	y_2 ……	y_m

求函数

$$\varphi(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\ldots\ldots+a_n\varphi_n(x)=\sum_{j=0}^{n}a_j\varphi_j(x)$$

使得

$$\sum_{i=1}^{m}[\varphi(x_i)-y_i]^2=\sum_{i=1}^{m}[\sum_{j=0}^{n}a_j\varphi_j(x_i)-y_i]^2$$

最小.

说明：

（1）若$\varphi(x)$为一元函数，则函数曲线为平面图形，称**曲线拟合**.

（2）$\varphi(x)$为拟合函数，上式最小为拟合条件（即要求拟合曲线与各数据点在y方向的误差平方和最小）.

（3）函数类的选取：$\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\}$

据实验数据分布特点选取，可选幂函数类、指数函数类、三角函数类等.

2、最小二乘法：

以残差平方和最小问题的解来确定拟合函数的方法.

令 $\delta_i = \varphi(x_i) - y_i \quad (i=1,2,\ldots m)$

——在回归分析中称为残差

残差向量：

$$\vec{\delta} = \begin{bmatrix} \varphi(x_1) - y_1 \\ \varphi(x_2) - y_2 \\ \cdots\cdots \\ \varphi(x_m) - y_m \end{bmatrix}$$

$= \|\vec{\delta}\|_2^2$

残差向量的各分量平方和记为：

$$S(a_0, a_1, \cdots, a_n) = \sum_{i=1}^{m}[\sum_{j=0}^{n} a_j \varphi_j(x_i) - y_i]^2$$

由多元函数求极值的必要条件，有

$$\frac{\partial S}{\partial a_k}=0 \qquad (k=0,1,\cdots,n)$$

可得

$$2\sum_{i=1}^{m}\varphi_k(x_i)[\sum_{j=0}^{n}a_j\varphi_j(x_i)-y_i]=0$$

即

$$\sum_{i=1}^{m}[\sum_{j=0}^{n}a_j\varphi_k(x_i)\varphi_j(x_i)-\varphi_k(x_i)y_i]=0$$

$$\sum_{i=1}^{m}\sum_{j=0}^{n}a_j\varphi_k(x_i)\varphi_j(x_i)=\sum_{i=1}^{m}\varphi_k(x_i)y_i$$

由
$$\sum_{i=1}^{m}\sum_{j=0}^{n} a_j \varphi_k(x_i)\varphi_j(x_i) = \sum_{i=1}^{m}\varphi_k(x_i) y_i$$

得
$$\sum_{j=0}^{n}\left[\sum_{i=1}^{m}\varphi_k(x_i)\varphi_j(x_i)\right]a_j = \sum_{i=1}^{m}\varphi_k(x_i) y_i$$

$$(k = 0, 1, \cdots, n)$$

即

$$a_0\sum_{i=1}^{m}\varphi_k(x_i)\varphi_0(x_i) + a_1\sum_{i=1}^{m}\varphi_k(x_i)\varphi_1(x_i) + \cdots$$
$$+a_n\sum_{i=1}^{m}\varphi_k(x_i)\varphi_n(x_i) = \sum_{i=1}^{m}\varphi_k(x_i) y_i \qquad (k = 0, 1, \cdots, n)$$

上式为由$n+1$个方程组成的方程组，称正规方程组.

引入记号 $\varphi_r=(\varphi_r(x_1),\varphi_r(x_2),\cdots,\varphi_r(x_m))$

$$f=(y_1,y_2,\cdots,y_m)$$

则由内积的概念可知

$$(\varphi_k,\varphi_j)=\sum_{i=1}^{m}\varphi_k(x_i)\varphi_j(x_i),\qquad (\varphi_k,f)=\sum_{i=1}^{m}\varphi_k(x_i)y_i$$

显然内积满足交换律 $(\varphi_k,\varphi_j)=(\varphi_j,\varphi_k)$

正规方程组便可化为

$$a_0(\varphi_k,\varphi_0)+a_1(\varphi_k,\varphi_1)+\cdots+a_n(\varphi_k,\varphi_n)=(\varphi_k,f)$$

$$(k=0,1,\cdots,n)$$

这是一个系数为(φ_k,φ_j),常数项为(φ_k,f)的线性方程组.

将其表示成矩阵形式：

$$\begin{pmatrix} (\varphi_0,\varphi_0) & (\varphi_0,\varphi_1) & \cdots & (\varphi_0,\varphi_n) \\ (\varphi_1,\varphi_0) & (\varphi_1,\varphi_1) & \cdots & (\varphi_1,\varphi_n) \\ \vdots & \vdots & \vdots & \vdots \\ (\varphi_n,\varphi_0) & (\varphi_n,\varphi_1) & \cdots & (\varphi_n,\varphi_n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} (\varphi_0,f) \\ (\varphi_1,f) \\ \vdots \\ (\varphi_n,f) \end{pmatrix}$$

其系数矩阵为对称阵.

由于$\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$为函数类$\Phi$的基,

因此$\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$必然线性无关.

正规方程组的系数矩阵非奇异，即 $\det[(\varphi_i,\varphi_j)_{n\times n}]\neq 0.$

根据Crame法则，正规方程组有唯一解，称其为

最小二乘解.

作为一种简单的情况，常使用多项式函数$P_n(x)$作为(x_i,y_i) $(i=1,2,\ldots,m)$的拟合函数.

拟合函数$\varphi(x)=P_n(x)$的基函数为：

$$\varphi_0(x)=1,\varphi_1(x)=x,\cdots,\varphi_k(x)=x^k,\cdots,\varphi_n(x)=x^n$$

基函数之间的内积为：

$$(\varphi_k,\varphi_j)=\sum_{i=1}^{m}\varphi_k(x_i)\varphi_j(x_i)=\sum_{i=1}^{m}x_i^k x_i^j=\sum_{i=1}^{m}x_i^{k+j}$$

$$(\varphi_k,f)=\sum_{i=1}^{m}\varphi_k(x_i)y_i=\sum_{i=1}^{m}x_i^k y_i$$

即正规方程组为

$$\begin{bmatrix} m & \sum_{k=1}^{m} x_k & \cdots & \sum_{k=1}^{m} x_k^n \\ \sum_{k=1}^{m} x_k & \sum_{k=1}^{m} x_k^2 & \cdots & \sum_{k=1}^{m} x_k^{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{k=1}^{m} x_k^n & \sum_{k=1}^{m} x_k^{n+1} & \cdots & \sum_{k=1}^{m} x_k^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{m} y_k \\ \sum_{k=1}^{m} x_k y_k \\ \vdots \\ \sum_{k=1}^{m} x_k^n y_k \end{bmatrix}$$

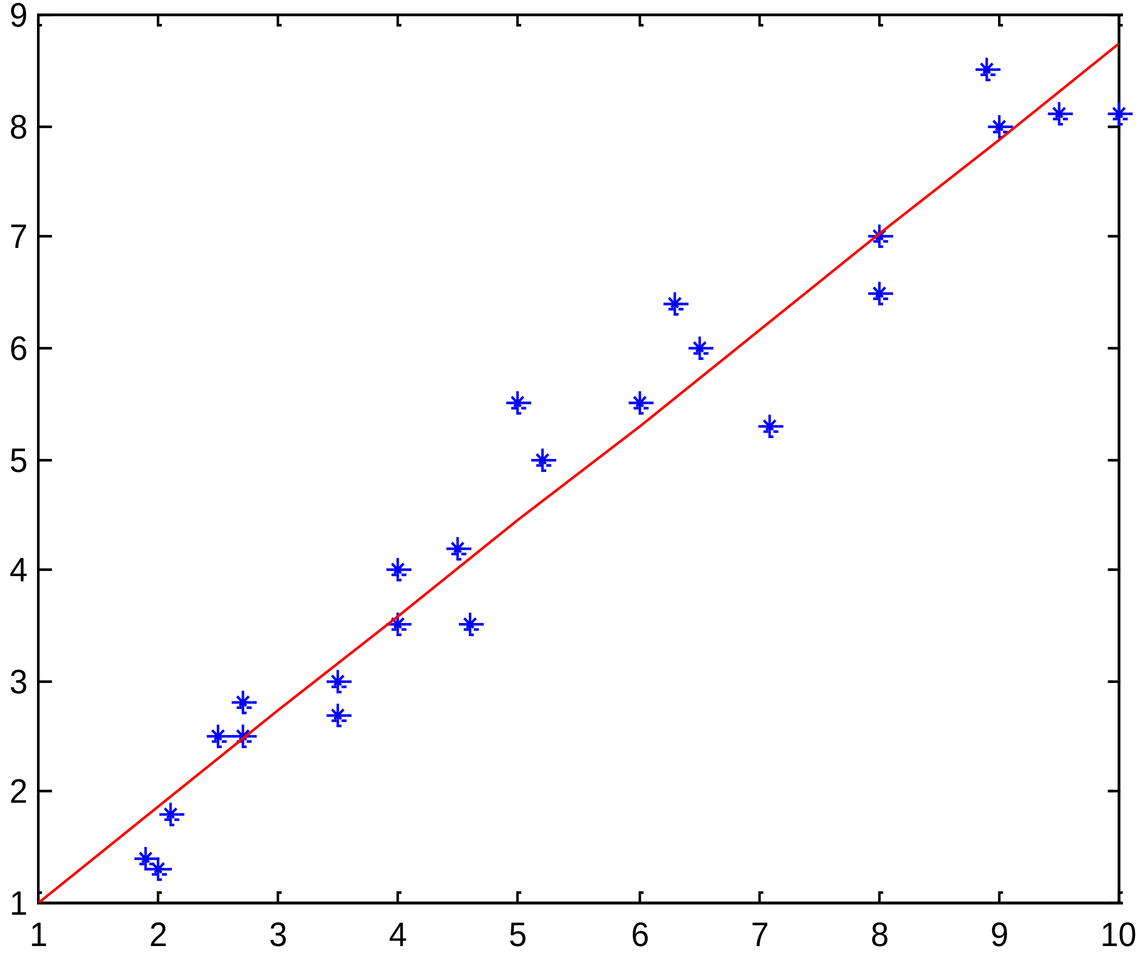
例. 回到本节开始的实例:

考察某种纤维的强度与其拉伸倍数的关系,下表是实际测定的24个纤维样品的强度与相应的拉伸倍数的记录:

编 号	拉伸倍数 x_i	强 度 y_i	编 号	拉伸倍数 x_i	强 度 y_i
1	1.9	1.4	13	5	5.5
2	2	1.3	14	5.2	5
3	2.1	1.8	15	6	5.5
4	2.5	2.5	16	6.3	6.4
5	2.7	2.8	17	6.5	6
6	2.7	2.5	18	7.1	5.3
7	3.5	3	19	8	6.5
8	3.5	2.7	20	8	7
9	4	4	21	8.9	8.5
10	4	3.5	22	9	8
11	4.5	4.2	23	9.5	8.1
12	4.6	3.5	24	10	8.1

纤维强度随拉伸倍数增加而增加.

其散点图为：



从散点图可以看出，纤维强度和拉伸倍数之间近似线性关系，故可选取线性函数

$$y(x) = a_0 + a_1 x$$

为拟合函数建立正规方程组，其基函数为

$$\varphi_0(x) = 1 \qquad \varphi_1(x) = x$$

根据内积公式，可得

$$(\varphi_0, \varphi_0) = 24 \qquad (\varphi_0, \varphi_1) = 127.5 \qquad (\varphi_1, \varphi_1) = 829.61$$

$$(\varphi_0, f) = 113.1 \qquad (\varphi_1, f) = 731.6$$

正规方程组为

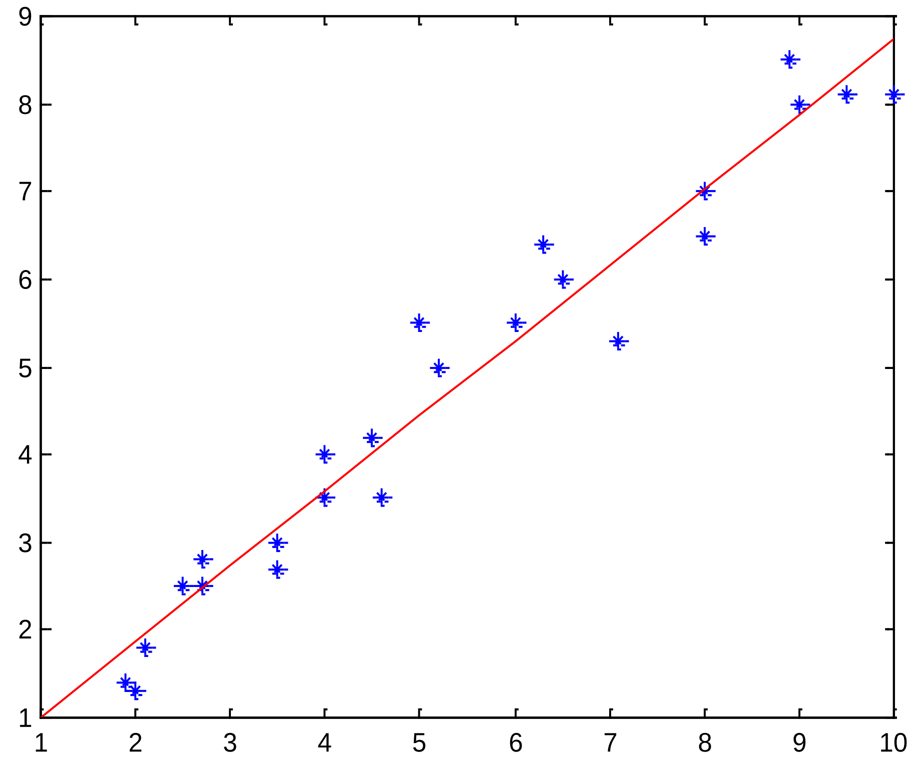
$$\begin{pmatrix} 24 & 127.5 \\ 127.5 & 829.61 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 113.1 \\ 731.6 \end{pmatrix}$$

解得 $a_0 = 0.1505$ $a_1 = 0.8587$

故 $y(x) = 0.1505 + 0.8587x$ 即为所求的最小二乘解.

残差平方和: $\|\delta^*\|_2^2 = 5.6615$

拟合曲线与散点的关系如右图:



二、 超定方程组的最小二乘解

将拟合函数以向量表示：

$$\varphi(x)=\begin{bmatrix}\varphi_0(x) & \varphi_1(x) & \cdots & \varphi_n(x)\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ \vdots\\ a_n\end{bmatrix}$$

令 $\varphi(x_i)=y_i$ （i=1,2,…,m）

可得

$$\begin{bmatrix}\varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_n(x_1)\\ \varphi_0(x_2) & \varphi_1(x_2) & \cdots & \varphi_n(x_2)\\ \cdots & \cdots & \cdots & \cdots\\ \varphi_0(x_m) & \varphi_1(x_m) & \cdots & \varphi_n(x_m)\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ \vdots\\ a_n\end{bmatrix}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_m\end{bmatrix}$$

若m>n+1，则此方程组称超定方程组（方程个数>未知数个数）

考虑正规方程组

$$\sum_{j=0}^{n}[\sum_{i=1}^{m}\varphi_k(x_i)\varphi_j(x_i)]a_j=\sum_{i=1}^{m}\varphi_k(x_i)y_i$$

（k=0,1,…n）

可知：

（1）未知数a_j的系数 $\sum_{i=1}^{m}\varphi_k(x_i)\varphi_j(x_i)$

为超定方程组中系数阵第k列与第j列对应积之和（即内积(φ_k,φ_j)）；

（2）右端向量 $\sum_{i=1}^{m}\varphi_k(x_i)y_i$

为系数阵第k列与m个函数值对应积之和.

故正规方程组矩阵形式为：　　$\Phi^T\Phi\vec{a}=\Phi^T\vec{y}$

若有唯一解，称其为超定方程组的**最小二乘解**.

注：最小二乘解并不能满足超定方程组中每个方程，但要求尽可能接近给定数据，即允许每个等式可以稍有偏差（即**残差**）.

求一般超定方程组$Ax=b$的主要过程：

（1）求出系数矩阵A的转置矩阵A^T；

（2）计算矩阵$D=A^TA$和向量$f=A^Tb$；

（3）求解正规方程组$Dx=f$.

例1 用多项式函数拟合下述给定数据：

x	1	2	3	4
y	4	10	18	26

解： 设

$$P(x)=a_0+a_1x+a_2x^2$$

得

$$\begin{cases} a_0+a_1+a_2=4 \\ a_0+2a_1+4a_2=10 \\ a_0+3a_1+9a_2=18 \\ a_0+4a_1+16a_2=26 \end{cases}$$

即

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}=\begin{bmatrix} 4 \\ 10 \\ 18 \\ 26 \end{bmatrix}$$

记系数矩阵为Φ，则

$$\Phi^T\Phi=\begin{bmatrix}4 & 10 & 30\\ 10 & 30 & 100\\ 30 & 100 & 354\end{bmatrix} \qquad \Phi^T\vec{y}=\begin{bmatrix}58\\ 182\\ 622\end{bmatrix}$$

故正规方程组为

$$\begin{bmatrix}4 & 10 & 30\\ 10 & 30 & 100\\ 30 & 100 & 354\end{bmatrix}\begin{bmatrix}a_0\\ a_1\\ a_2\end{bmatrix}=\begin{bmatrix}58\\ 182\\ 622\end{bmatrix}$$

解得

$$a_0=-\frac{3}{2},a_1=\frac{49}{10},a_2=\frac{1}{2}$$

拟合曲线：

$$P(x) = -\frac{3}{2} + \frac{49}{10}x + \frac{1}{2}x^2$$

- 注：具体用几次多项式拟合，可据实际情况而定.可先画草图，将已知点描上去，看与什么函数相近，就以什么函数拟合.

MATLAB实现多项式拟合实例：

某种产品在生产过程中的废品率y与它所含的某种物质量x有关，现将试验所得16组数据记录列于下表.

x	34	36	37	38	39	39	39	40
y	1.30	1.00	0.73	0.90	0.81	0.70	0.60	0.50
x	40	41	42	43	43	45	47	48
y	0.44	0.56	0.30	0.42	0.35	0.40	0.41	0.60

要求拟合y与x的函数关系.

MATLAB实现：

```
clear,clc
x=[34 36 37 38 39 39 39 40 40 41 42 43 43 45 47 48];
y=[1.30 1.00 0.73 0.90 0.81 0.70 0.60 0.50 0.44 0.56
    0.30 0.42 0.35 0.40 0.41 0.60];
subplot(2 2 1),plot(x,y,'o'),title('散点图','fontsize',14);
s={'二次多项式','三次多项式','四次多项式'};
```

```
for i=2:4
p=polyfit(x,y,i)；
xi=linspace(34,48,1000);    %绘图的X轴数据
Z=polyval(p,xi);    %得到多项式在数据点处的值
z=polyval(p,x);
E(i-1)=sum((y-z).^2)
subplot(2,2,i),
plot(x,y,'ko',xi,Z,'r-')
title(s{i-1},'fontsize',14)
end
set(gcf,'color','w')
```

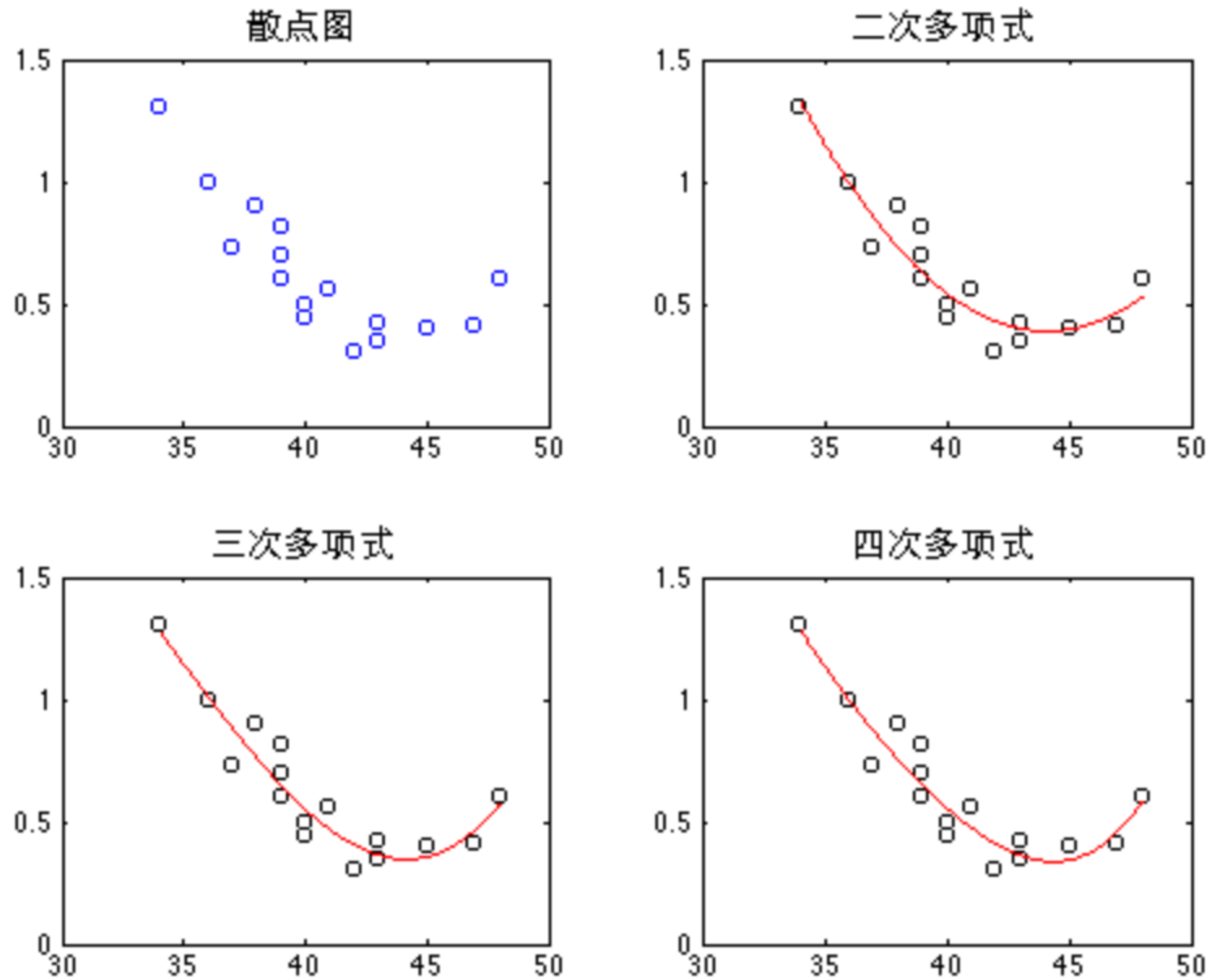
程序计算返回结果为：

E =

0.1304 0.1174 0.1165

可知当多项式拟合的次数越高，与原数据的误差平方和越小，但是次数的增高并不能明显地减小误差平方和. 其实二次多项式拟合已经能够达到较好的效果.

绘出各种拟合曲线图如下：



6.2　正交多项式　　——函数逼近的重要工具

n维向量空间： $\mathbf{R}^n$，所有n维实向量的集合，按向量加法及向量与数的乘法构成实数域上的线性空间；

多项式空间： H_n，次数不超过n（n为正整数）的实系数多项式全体，按多项式与多项式加法及数与多项式乘法构成实数域上的线性空间；

函数空间： $C[a,b]$，所有定义在$[a,b]$上的连续函数集合，按函数加法及数与函数乘法构成实数域上的线性空间；

$C^p[a,b]$：具有p阶连续导数的函数空间.

定义1 设集合S是数域P上的线性空间，元素$x_1, x_2, \cdots, x_n \in S$，如果存在不全为零的数$a_1, a_2, \cdots, a_n \in P$，使得

$$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = 0$$

则称$x_1, x_2, \cdots, x_n$线性相关，否则，若等式只对$a_1 = a_2 = \cdots = a_n = 0$成立，则称$x_1, x_2, \cdots, x_n$线性无关.

若线性空间S是由n个线性无关元素$x_1, x_2, \cdots, x_n$生成的，即对$\forall x \in S$,都有

$$x = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

则称$x_1, x_2, \cdots, x_n$为空间S的一组基，记为

$$S = span\{x_1, x_2, \cdots, x_n\}.$$

对于多项式集合H_n，其元素$p(x)\in H_n$，表示为

$$p(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n$$

它由$n+1$个系数$a_0,a_1,\cdots,a_n$唯一确定.

$1,x,x^2,\cdots,x^n$线性无关，构成H_n的一组基，故

$$H_n=span\{1,x,\cdots,x^n\}.$$

魏尔斯特拉斯(Weierstrass)定理:

设$f(x)\in C[a,b]$，则对任何$\varepsilon>0$，总存在一个代数多项式，使

$$\max_{a\le x\le b}|f(x)-p(x)|<\varepsilon \qquad \text{在}[a,b]\text{上一致成立.}$$

权函数的概念：

定义1 在区间(a, b)上，若非负函数$\rho(x)$满足

（1）对一切整数$n \geq 0, \int_a^b x^n \rho(x) dx$存在；

（2）对(a,b)上的非负连续函数$f(x)$，若$\int_a^b \rho(x) f(x) dx$ $=0$，则在区间(a,b)上$f(x) \equiv 0$.

则称为$\rho(x)$为区间(a,b)上的权函数.

常见的权函数：

$$\rho(x)\equiv 1，\ a\le x\le b$$

$$\rho(x)\equiv\frac{1}{\sqrt{1-x^2}},-1\le x\le 1;$$

$$\rho(x)\equiv\sqrt{1-x^2},-1\le x\le 1;$$

$$\rho(x)\equiv e^{-x},0\le x\le\infty;$$

$$\rho(x)\equiv e^{-x^2},-\infty\le x\le\infty;$$

正交函数系的概念：

定义2 设 $f(x), g(x) \in C[a, b]$, $\rho(x)$是区间$[a,b]$上的非负函数，则称

$$(f, g) = \int_a^b \rho(x) f(x) g(x) dx$$

为$[a,b]$上以$\rho(x)$为权函数的**内积**.

定义3 设 $f(x), g(x) \in C[a, b]$, $\rho(x)$是区间$[a,b]$上的权函数，若

$$(f, g) = \int_a^b \rho(x) f(x) g(x) dx = 0$$

成立，则称$f(x), g(x)$在$[a, b]$上带权$\rho(x)$**正交**. 当$\rho(x)=1$时，简称正交.

若函数系 $\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\}$

满足关系

$$\left(\varphi_j(x),\varphi_k(x)\right)=\int_a^b \rho(x)\varphi_j(x)\varphi_k(x)dx=\begin{cases}0 & j\neq k\\ A_k>0 & j=k\end{cases}$$

则称$\{\varphi_k(x)\}$是$[a,b]$上带权$\rho(x)$的正交函数系.

若$A_k\equiv 1$，则称之为标准正交函数系.

例1 验证 $\varphi_0(x)=1$, $\varphi_1(x)=x$ 在$[-1, 1]$上正交.

解:

$$\int_{-1}^{1}\varphi_0(x)\varphi_1(x)dx=\int_{-1}^{1}1\cdot xdx=0$$

所以 $\varphi_0(x)$与 $\varphi_1(x)$在$[-1, 1]$上正交.

例2 证明:当$m\neq n$时,$\cos(m\theta)$ 和 $\cos(n\theta)$在区间$[-\pi, \pi]$上正交.

证

$$\int_{-\pi}^{\pi}\cos(m\theta)\cos(n\theta)d\theta$$

$$=\frac{1}{2}\int_{-\pi}^{\pi}[\cos(m+n)\theta+\cos(m-n)\theta]d\theta=0$$

所以, $\cos\theta, \cos2\theta, \cos3\theta,\cdots, \cos(n\theta),\cdots\cdots$

是正交函数系.

例3 验证函数系$\{1, \cos\theta, \sin\theta, \ldots\ldots, \cos(n\theta), \sin(n\theta)\}$

在区间$[-\pi, \pi]$上是正交函数系.

证

$$\int_{-\pi}^{\pi} 1^2 d\theta = 2\pi,$$

$$\int_{-\pi}^{\pi} \cos^2(n\theta) d\theta = \int_{-\pi}^{\pi} [1 - \sin^2(n\theta)] d\theta$$

$$= 2\pi - \frac{1}{2}\int_{-\pi}^{\pi} [1 - \cos(2n\theta)] d\theta = 2\pi - \pi = \pi,$$

$$\int_{-\pi}^{\pi} \sin^2(n\theta) d\theta = \frac{1}{2}\int_{-\pi}^{\pi} [1 - \cos(2n\theta)] d\theta = \pi,$$

$$\int_{-\pi}^{\pi} 1\cdot\cos(n\theta)d\theta=0,\quad \int_{-\pi}^{\pi} 1\cdot\sin(n\theta)d\theta=0,$$

$$\int_{-\pi}^{\pi}\cos(m\theta)\cos(n\theta)d\theta$$

$$=\frac{1}{2}\int_{-\pi}^{\pi}[\cos(m+n)\theta+\cos(m-n)\theta]d\theta=0$$

$$\int_{-\pi}^{\pi}\sin(m\theta)\sin(n\theta)d\theta$$

$$=\frac{1}{2}\int_{-\pi}^{\pi}[\cos(m+n)\theta-\cos(m-n)\theta]d\theta=0$$

$$\int_{-\pi}^{\pi}\sin(m\theta)\cos(n\theta)d\theta$$

$$=\frac{1}{2}\int_{-\pi}^{\pi}[\sin(m+n)\theta+\sin(m-n)\theta]d\theta=0$$

故 函数系$\{1,\cos\theta,\sin\theta,\ldots\ldots,\cos(n\theta),\sin(n\theta)\}$在区间

$[-\pi,\pi]$上是正交函数系.