计算方法

第一章 插值方法

胡 敏

合肥工业大学 计算机与信息学院

jsjxhumin@hfut.edu.cn

uhnim@163.com

1.9 曲线拟合的最小二乘法

1．教学内容：

曲线拟合的概念、直线拟合、多项式拟合、正则方程组。

2．重点难点：

拟合曲线的类型、正则方程组的建立、拟合多项式的求解。

3．教学目标：

了解曲线拟合的概念、对给出的一组数据点，能判断其拟合曲线的类型、建立相应的正则方程组、求得拟合多项式

引言

一、问题的提法

在科学实验和生产实践中，经常要从一组实验数据出发，寻找函数$y=f(x)$的一个近似公式（称为经验公式）。已有的多项式插值法解决这类问题有明显的缺陷：实验数据有误差；实验数据量大等。

二、目　的

实际应用中并不刻意要求曲线经过所有的观测点，而是在符合数据分布特征的某类曲线中，在某个函数类中寻找一个“最好”的函数来拟合这组数据。

三、方 法

曲线拟合方法. 数据拟合最常用的近似标准是最小二乘法则.

最小二乘法

一、基本概念：残差

$$e_i = y_i - \hat{y}_i \quad (i = 1, 2, \cdots, N)$$

拟合的目的：使得残差最小，其中 $\hat{y} = \varphi(x)$ 为所要找的函数。

二、残差的选取方法（原则）

1、选取 $\varphi(x)$ 使残差绝对值之和最小，即

$$\sum_{i=1}^{N} |e_i| = \sum_{i=1}^{N} |y_i - \hat{y}_i| = \min$$

2、选取 φ(*x*)，使残差最大绝对值最小，即

$$\max_i |e_i| = \max_i |y_i - \hat{y}_i| = \min$$

3、选取 φ(*x*)，使残差平方和最小，即

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_i - \hat{y}_i]^2 = \min$$

三、最小二乘原则（方法）

1、定义：使“残差平方和最小”的原则称为最小二乘原则。

2、定义：按照最小二乘原则选取拟合曲线的方法，称为最小二乘法。

1、直线拟合

假设给定的数据点 (x_i, y_i)， $i = 1, 2, \ldots, n$ 的分布大致成一直线，虽然我们不能要求所做的拟合直线

$$y = a + bx$$

严格地通过所有的数据点 (x_i, y_i)，但总希望它尽可能地从所给数据点附近通过，即要求近似成立

$$y_i \approx a + bx_i, \qquad i = 1, 2, \ldots, n$$

由于数据点数目通常远远大于待定系数的个数，因此，拟合直线的构造实际上是求解超定方程（矛盾）方程组的代数问题。

设

$$\hat{y}_i \approx a + bx_i, \qquad i = 1,2, \ldots ,n$$

表示按拟合直线 y=a+bx 求得的近似值，它一般不同于观测值 y_i

两者之差

$$e_i = y_i - \hat{y}_i$$

称为**残差**

显然，残差的大小是衡量拟合好坏的重要标志。

具体地说，构造拟合曲线可以采用下列三种准则之一：

（1）使残差的最大绝对值为最小：

$$\max_i |e_i|$$

（2）使残差的绝对值之和为最小：

$$\sum_i |e_i|$$

（3）使残差的平方和为最小：

$$\sum_i e_i^2$$

（1）、（2）两种由于含有绝对值运算不便于实际应用。

基于准则（3）来选取拟合曲线的方法称为曲线拟合的**最小二乘法**

直线拟合问题可用数学语言描述如下：

问题10 对于给定的数据点 (x_i, y_i)， $i = 1, 2, \ldots, n$ 求作一次式 y=a+bx，使总误差

$$Q = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$$

为最小。

要使 Q 达到极值，参数a，b 应满足

$$\frac{\partial Q}{\partial a}=2\sum_{i=1}^{n}[y_i-(a+bx_i)]\frac{\partial[y_i-(a+bx_i)]}{\partial a}=-2\sum_{i=1}^{n}[y_i-(a+bx_i)]$$

即

$$\sum_{i=1}^{N}[y_i-(a+bx_i)]=0$$

$$\sum_{i=1}^{N}[y_i-(a+bx_i)]x_i=0$$

由此可得：

$$\begin{cases} aN+b\sum_{i=1}^{N}x_i=\sum_{i=1}^{N}y_i \\ a\sum_{i=1}^{N}x_i+b\sum_{i=1}^{N}x_i^2=\sum_{i=1}^{N}x_iy_i \end{cases} \qquad (42)$$

解线性方程组（42）既可得到a，b

例： 炼钢是个氧化脱碳的过程，钢液含碳量的多少直接影响冶炼时间的长短，下表是某平炉的生产记录，表中 i 是次数，x_i 为全部炉料熔化完毕时的钢液的含碳量，y_i 为熔毕至出钢所许的冶炼时间。

I	1	2	3	4	5
x_i	165	123	150	123	141
y_i	187	126	172	125	148

解：

设所求的拟合直线为 y=a+bx

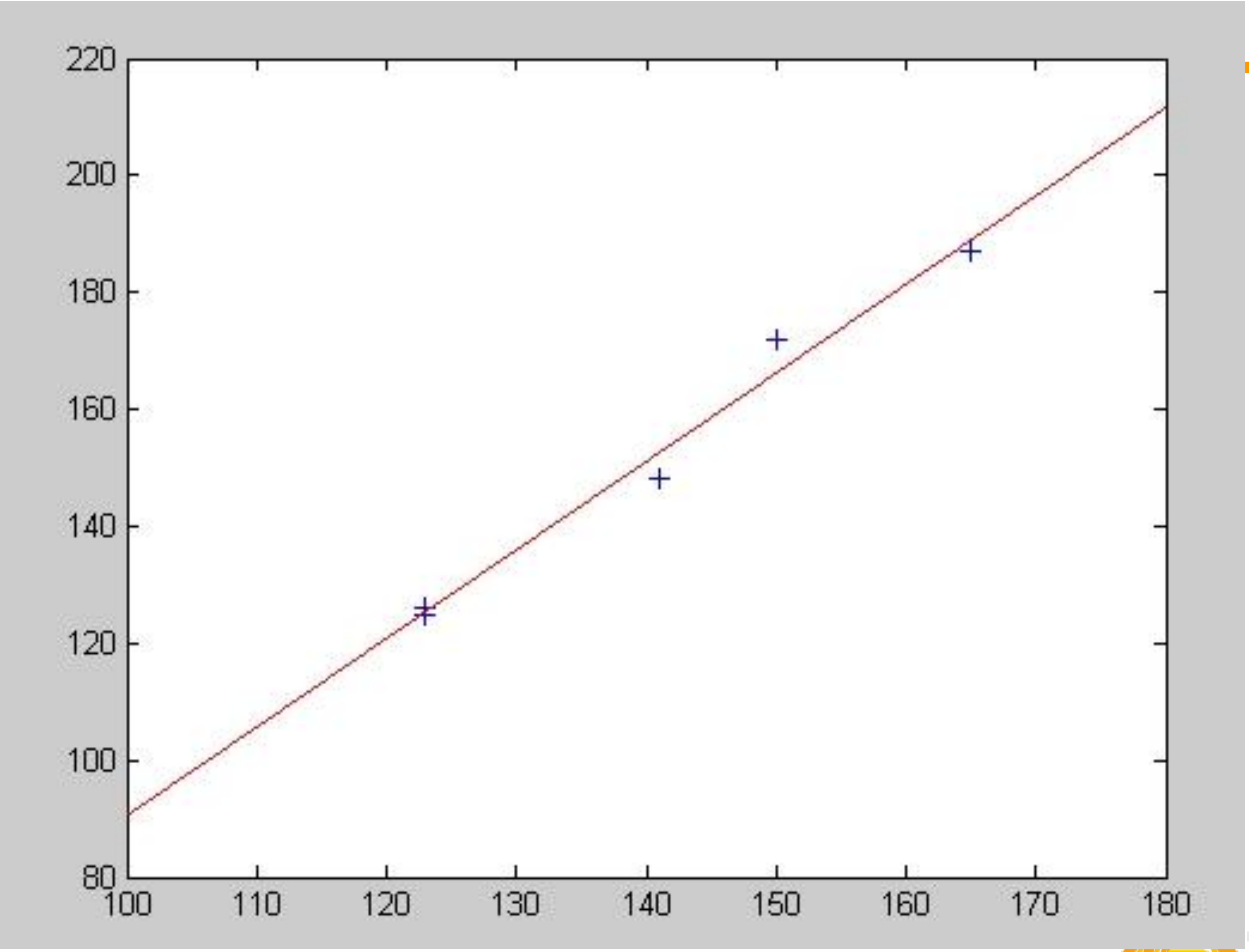
由（42）式可得关于a，b的线性方程组

$$\begin{cases} 5a + 702b = 758 \\ 702a + 99864b = 108396 \end{cases}$$

解此线性方程组得：a=-60.9392，b=1.5138

故拟合直线为：

$$y = -60.9392 + 1.5138x$$


220
200
180
160
140
120
100
80
100
110
120
130
140
150
160
170
180

2、多项式拟合

多项式拟合,是最流行的数据处理方法之一.它常用于把观测数据(离散的数据)归纳总结为经验公式(连续的函数),以利于进一步的推演分析或应用.

问题11 对于给定的数据点 (x_i, y_i), $i = 1, 2, \ldots, n$ 求作m（m<<N）次多项式

$$y = \sum_{j=0}^{m} a_j x^j$$

使总误差

$$Q = \sum_{i=1}^{N} [y_i - \sum_{j=0}^{m} a_j x_i^j]^2$$

为最小。

由于Q可以看成是关于 $a_j(j=0, 1, \ldots, m)$ 的多元函数，故上述拟合多项式的构造问题可归结为多元函数的极值问题。

令
$$\frac{\partial Q}{\partial a_k}=0, \quad k=0, 1, \ldots, m$$

得：

$$\sum_{i=1}^{N}\left(y_i-\sum_{j=0}^{m}a_j x_i^j\right)x_i^k=0, \quad k=0, 1, \ldots, m$$

既

$$\begin{cases} a_0 N + a_1 \sum_{i=1}^{N} x_i + \ldots + a_m \sum_{i=1}^{N} x_i^m = \sum_{i=1}^{N} y_i \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + \ldots + a_m \sum_{i=1}^{N} x_i^{m+1} = \sum_{i=1}^{N} x_i y_i \\ \cdots\cdots\cdots \\ a_0 \sum_{i=1}^{N} x_i^m + a_1 \sum_{i=1}^{N} x_i^{m+1} + \ldots + a_m \sum_{i=1}^{N} x_i^{2m} = \sum_{i=1}^{N} x_i^m y_i \end{cases} \qquad (43)$$

这个关于系数 a_j 的线性方程组称为**正则方程组**

定理7　正则方程组（43）有唯一解

证： 用反证法，若不然，则对应的齐次方程组

$$\begin{cases} a_0 N + a_1 \sum_{i=1}^{N} x_i + \cdots + a_m \sum_{i=1}^{N} x_i^m = 0 \\ a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + \cdots + a_m \sum_{i=1}^{N} x_i^{m+1} = 0 \\ \cdots\cdots \\ a_0 \sum_{i=1}^{N} x_i^m + a_1 \sum_{i=1}^{N} x_i^{m+1} + \cdots + a_m \sum_{i=1}^{N} x_i^{2m} = 0 \end{cases}$$

有非零解

而

$$a_0\sum_{i=1}^{N}x_i^k+a_1\sum_{i=1}^{N}x_i^{k+1}+\cdots+a_m\sum_{i=1}^{N}x_i^{k+m}=\sum_{j=0}^{m}a_j\sum_{i=0}^{N}x_i^{k+j}$$

从而有

$$\sum_{j=0}^{m}a_j\sum_{i=1}^{N}x_i^{k+j}=0,\quad k=0,\ 1,\ \ldots,\ m$$

所以

$$\sum_{k=0}^{m}a_k(\sum_{j=0}^{m}a_j\sum_{i=1}^{N}x_i^{k+j})=0$$

而

$$\sum_{k=0}^{m} a_k \left(\sum_{j=0}^{m} a_j \sum_{i=1}^{N} x_i^{k+j}\right) = \sum_{i=1}^{N} \sum_{k=0}^{m} \sum_{j=0}^{m} a_k a_j x_i^{k+j}$$

$$= \sum_{i=1}^{N} \left(\sum_{k=0}^{m} a_k x_i^{k}\right)\left(\sum_{j=0}^{m} a_j x_i^{j}\right) = \sum_{i=1}^{N} \left(\sum_{j=0}^{m} a_j x_i^{j}\right)^2$$

因此有:

$$\sum_{j=0}^{m} a_j x_i^j = 0 \quad (i = 1,2, \ldots ,N)$$

即拟合多项式 $y = \sum_{j=0}^{m} a_j x^j$

有N个零点 $x_i (i = 1,2, \ldots ,N)$

当N>m时，由代数学基本定理知必有 $\sum_{j=0}^{m} a_j x^j \equiv 0$

从而

$$a_j = 0 \qquad (j = 1,2, \ldots ,m)$$

故与正则方程组的题设矛盾，定理得证。

定理8

设 $a_j\ (j=0,1,\cdots,m)$ 为正则方程（43）的解，则 $y=\sum\limits_{j=0}^{m}a_jx^j$ 必为问题11的解。

证： 任给一组值 $b_j \quad (j=0,1,\cdots,\ m)$ 有

$$\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m}b_jx_i^j)^2-\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m}a_jx_i^j)^2$$

$$=\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m}a_jx_i^j+\sum_{j=0}^{m}a_jx_i^j-\sum_{j=0}^{m}b_jx_i^j)^2-\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m}a_jx_i^j)^2$$

$$=\sum_{i=1}^{N}(y_i - \ldots x_i^j)(\sum_{j=0}^{m}a_jx_i^j-\sum_{j=0}^{m}b_jx_i^j)+$$

$$(\sum_{j=0}^{m}a_jx_i^j - \ldots x_i^j)^2$$

利用正则方程组（43）可以知道，该项应该为零

$$=2\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m} \ldots x_i^j)(\sum_{j=0}^{m}a_jx_i^j-\sum_{j=0}^{m}b_jx_i^j)+(\sum_{j=0}^{m}a_jx_i^j-\sum_{j=0}^{m}b_jx_i^j)^2$$

$$\geq 2\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m}a_jx_i^j)(\sum_{j=0}^{m}a_jx_i^j-\sum_{j=0}^{m}b_jx_i^j)=0$$

因而有

$$\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m}a_jx_i^j)^2\le\sum_{i=1}^{N}(y_i-\sum_{j=0}^{m}b_jx_i^j)^2$$

所以，只有 $a_j\,(j=0,1,\cdots,m)$ 使得残差的平方和最小

故

$$y=\sum_{j=0}^{m}a_jx^j$$

必为问题11的解。

多项式拟合的一般方法可归纳为：

(1)根据具体问题，确定拟合多项式的次数*n*；（描点）

(2)计算正则方程组的系数和右端项

$$S_k=\sum_{i=0}^{m}x_i^k,\qquad t_k=\sum_{i=0}^{m}x_i^k\,y_i.$$

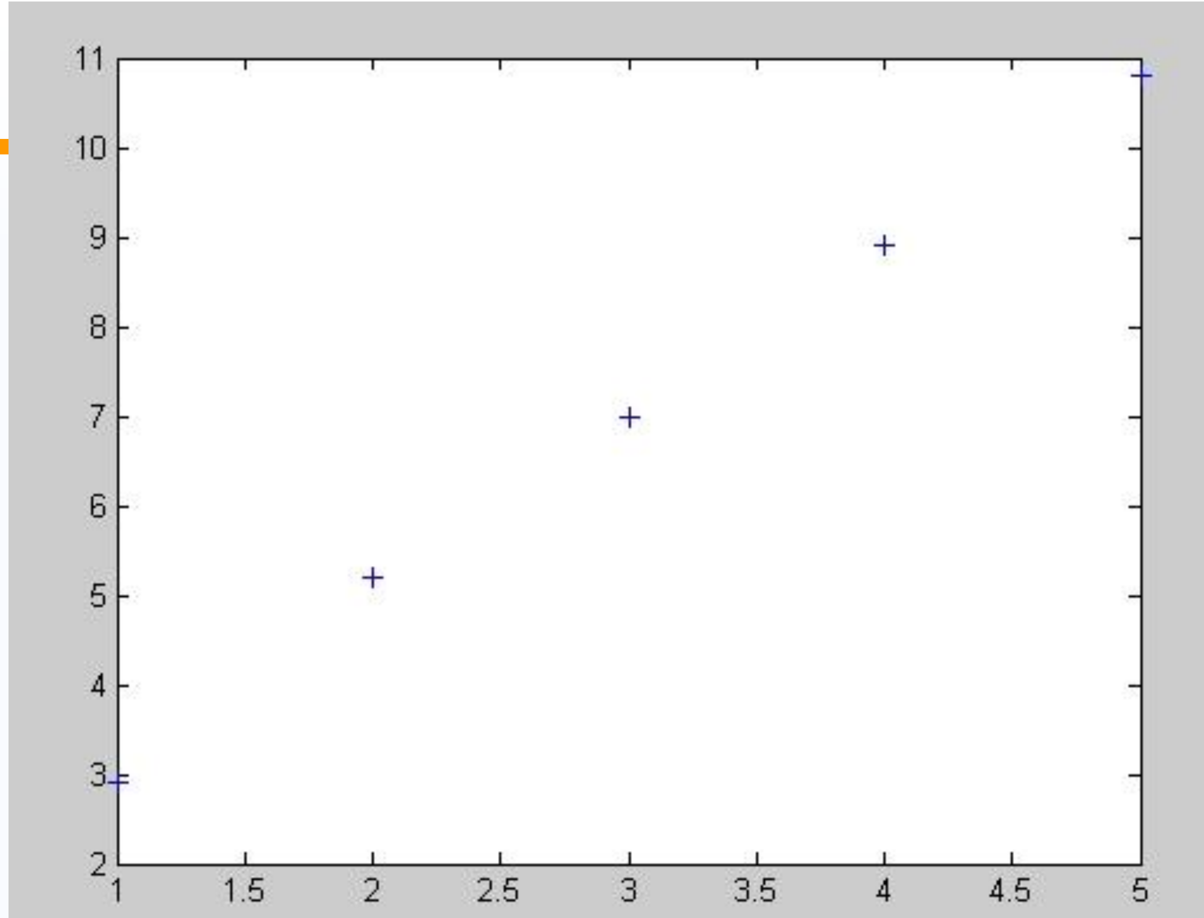
(3)写出正则方程组

(4)解正则方程组,求出$a_0,a_1,\ldots,a_n$;

(5)写出拟合多项式$P_n(x)$

例： 试求一个多项式拟合下列数据。

x	1	2	3	4	5
y	2.9	5.2	7	8.9	10.8



解：

如图所示，它们大体分布在一条直线上，故考虑用线性函数拟合这些数据。

设所求的拟合直线为 y=a+bx

由（42）式

$$\begin{cases} aN + b\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \\ a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \end{cases} \qquad (42)$$

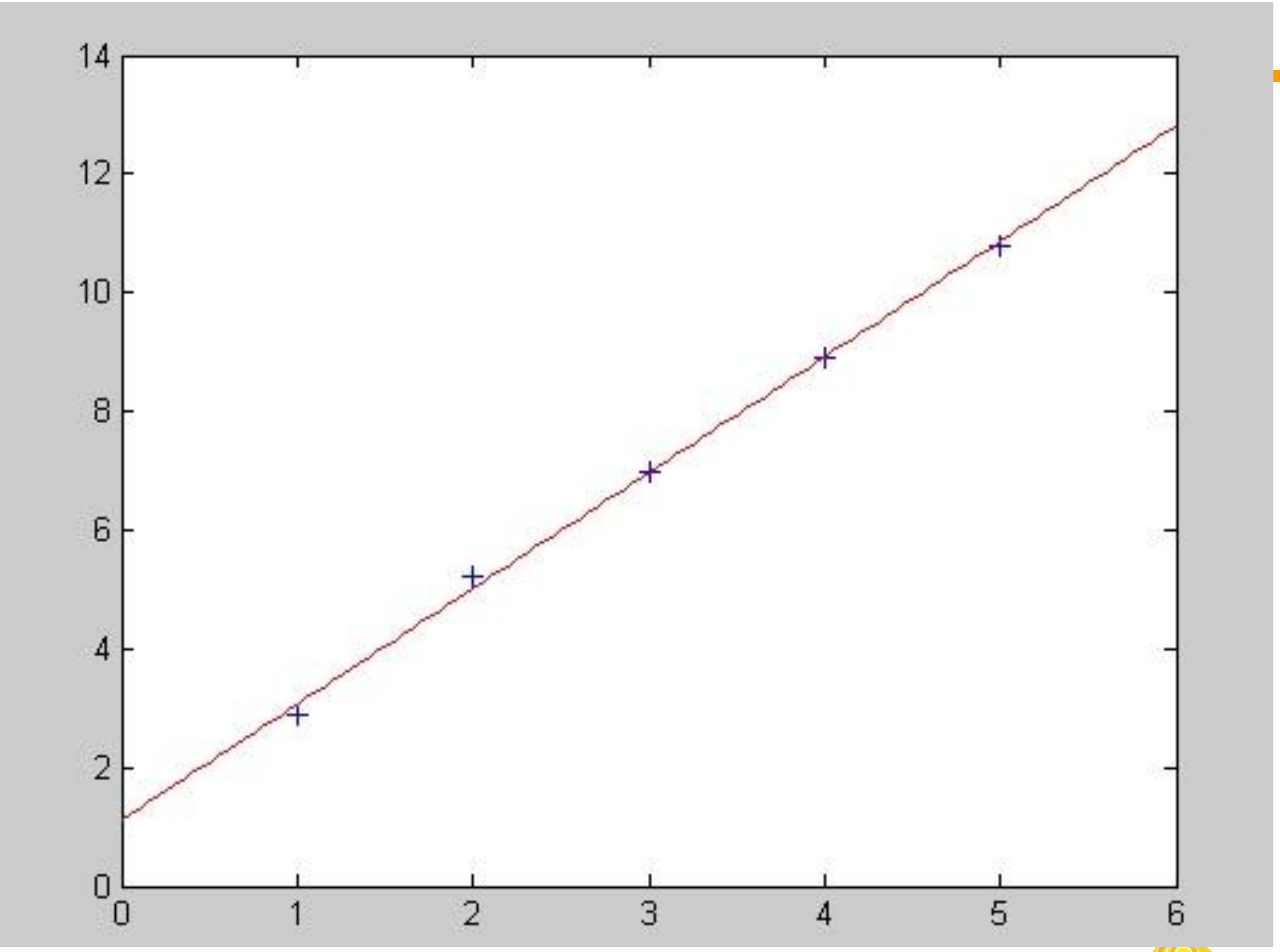
可得关于a，b的线性方程组

$$\begin{cases} 5a + 15b = 34.8 \\ 15a + 55b = 123.9 \end{cases}$$

解此方程组得：a=1.11，b=1.95

故所求拟合直线为：

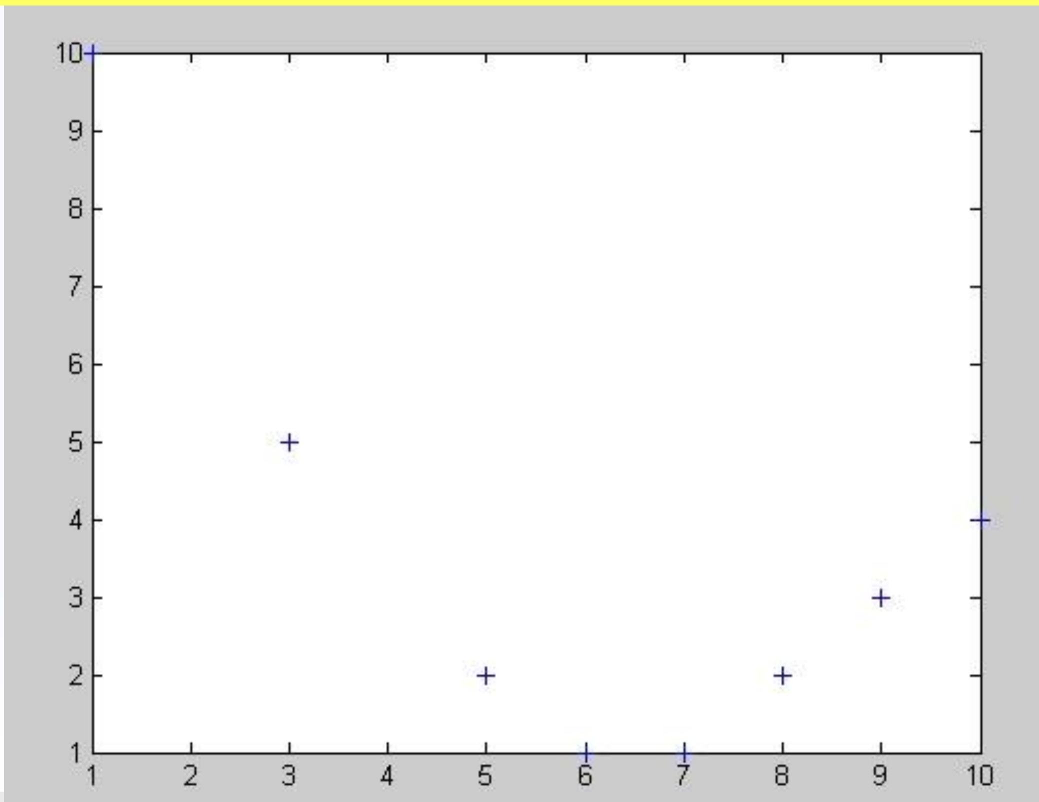
$$y = 1.11 + 1.95x$$


14
12
10
8
6
4
2
0
0
1
2
3
4
5
6

例： 试求一个多项式拟合下列数据。

x	1	3	5	6	7	8	9	10		

如图所示，它们大体分布在一条抛物线附近，故考虑用二次多项式函数拟合这些数据。



解：

设所求的拟合多项式为 $y = a_0 + a_1 x + a_2 x^2$

由（43）式

$$
\begin{cases}
a_0 N + a_1 \sum_{i=1}^{N} x_i + \ldots + a_m \sum_{i=1}^{N} x_i^m = \sum_{i=1}^{N} y_i \\
a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + \ldots + a_m \sum_{i=1}^{N} x_i^{m+1} = \sum_{i=1}^{N} x_i y_i \\
\cdots\cdots\cdots\cdots \\
a_0 \sum_{i=1}^{N} x_i^m + a_1 \sum_{i=1}^{N} x_i^{m+1} + \ldots + a_m \sum_{i=1}^{N} x_i^{2m} = \sum_{i=1}^{N} x_i^m y_i
\end{cases}
\qquad (43)
$$

得其正则方程组为：

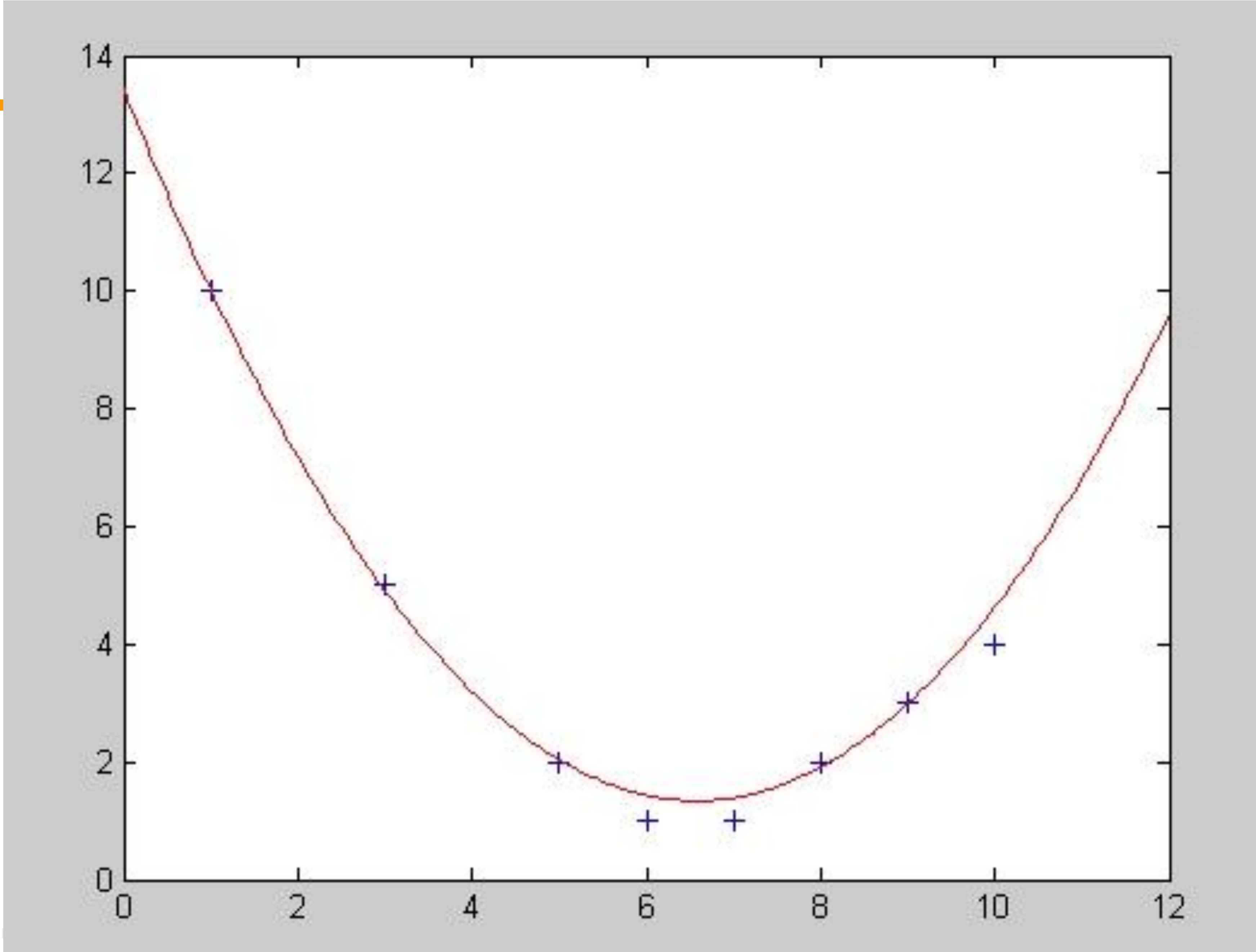
$$\begin{cases} 8a_0 + 49a_1 + 365a_2 = 28 \\ 49a_0 + 365a_1 + 2953a_2 = 131 \\ 365a_0 + 2953a_1 + 25061a_2 = 961 \end{cases}$$

解此方程组得：

$$a_0 = 13.43, \quad a_1 = -3.68, \quad a_2 = 0.28$$

所以，所求拟合多项式为：

$$y = 13.43 - 3.68x + 0.28x^2$$


14
12
10
8
6
4
2
0
0
2
4
6
8
10
12

3、观察数据的修匀　　（选学）

提高拟合多项式的次数不一定能改善逼近的效果，实际应用中常用不同的低次多项式去拟合不同的分段，这种方法称为**分段拟合**

对于给出的一组观察数据，不可避免地会产生随机干扰和误差，分段拟合的曲线在两段曲线交接的地方，也可能产生不够光滑的现象。因此，我们希望，根据数据分布的总趋势去剔除观察数据中的偶然误差，这就是所谓**数据修匀（或称数据平滑）**

考察相邻的五个节点

$$x_{-2} \langle x_{-1} \langle x_0 \langle x_1 \langle x_2$$

假设节点是等距的，节点间距为h，记 $x = x_0 + th$

有 $$t = \frac{(x - x_0)}{h}$$

$$t_i = \frac{(x_i - x_0)}{h} = i(i = -2,-1,0,1,2)$$

数据如下表

t_i	-2	-1	0	1	2
y_i	y_{-2}	y_{-1}	y_0	y_1	y_2

设用二项式作拟合

$$y = a + bt + ct^2$$

则其正则方程组为

$$\begin{cases} 5a + 10b = \sum_{i=-2}^{2} y_i \\ 10b = \sum_{i=-2}^{2} i y_i \\ 10a + 34c = \sum_{i=-2}^{2} i^2 y_i \end{cases}$$

解出a，b，c，即可得出在节点 $x = x_0$ 处的**五点二次修匀公式**

$$\hat{y}_0 = \frac{1}{35}(-3y_{-2} + 12y_{-1} + 17y_0 + 12y_1 - 3y_2)$$

小 结

插值问题

设函数f(x)在区间[a ，b]上有定义，且已知在一组互异点 $a \le x_0 < x_1 < \ldots < x_n \le b$ 上的函数 值 y_0， y_1， …， y_n，寻求一个简单的函数p(x)，使满足

$$p(x_i) = y_i, \quad i = 0, 1, 2, \ldots, n \quad (1.1)$$

并用p(x)近似代替f(x)，上述问题称为**插值问题**。

拉格朗日插值基函数

$$l_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{(x - x_j)}{(x_k - x_j)}$$

拉格朗日插值公式：

$$L_n(x) = \sum_{k=0}^{n} l_k(x) y_k = \sum_{k=0}^{n} \left(\prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j} \right) y_k$$

拉格朗日插值多项式存在并且唯一，并有估计式

$$f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{k=0}^{n} (x - x_k)$$

n阶差商可以递推定义为：

$$f(x_0, x_1, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n) - f(x_0, x_1, \ldots, x_{n-1})}{x_n - x_0}$$

n阶差商的性质：

$$f(x_0, x_1, \ldots, x_n) = \sum_{k=0}^{n} \frac{f(x_k)}{\prod\limits_{\substack{j=0 \\ j \neq k}}^{n} (x_k - x_j)} \qquad (24)$$

n阶差商关于节点是对称的

差商与导数的关系

$$f(x_0, x_1, \cdots, x_n) = \frac{f^{(n)}(\xi)}{n!}$$

差商表的建立与使用

x	$f(x)$	一阶差商	二阶差商	三阶差商
x_0	$f(x_0)$			
x_1	$f(x_1)$	$f(x_0,x_1)$		
x_2	$f(x_2)$	$f(x_1,x_2)$	$f(x_0,x_1,x_2)$	
x_3	$f(x_3)$	$f(x_2,x_3)$	$f(x_1,x_2,x_3)$	$f(x_0,x_1,x_2,x_3)$

牛顿插值公式

$$p_n(x) = f(x_0) + f(x_0, x_1)(x - x_0) + f(x_0, x_1, x_2)(x - x_0)(x - x_1) + \ldots + f(x_0, x_1, \ldots, x_n)(x - x_0)(x - x_1)\ldots(x - x_{n-1})$$

差分的定义

$$\Delta^n y_i = \Delta^{n-1} y_{i+1} - \Delta^{n-1} y_i$$

有限差分公式

$$p_n(x_0 + th) = y_0 + \sum_{k=1}^{n} \frac{\Delta^k y_0}{k!} \prod_{j=0}^{k-1} (t - j)$$

差分表

x_i	y_i	一阶	二阶	三阶	…
x_0	y_0				
x_1	y_1	Δy_0			
x_2	y_2	Δy_1	$\Delta^2 y_0$		
x_3	y_3	Δy_2	$\Delta^2 y_1$	$\Delta^3 y_0$	
…	…	…	…	…	…

正则方程组

$$
\begin{cases}
a_0 N + a_1 \sum_{i=1}^{N} x_i + \ldots + a_m \sum_{i=1}^{N} x_i^m = \sum_{i=1}^{N} y_i \\
a_0 \sum_{i=1}^{N} x_i + a_1 \sum_{i=1}^{N} x_i^2 + \ldots + a_m \sum_{i=1}^{N} x_i^{m+1} = \sum_{i=1}^{N} x_i y_i \\
\cdots\cdots\cdots\cdots \\
a_0 \sum_{i=1}^{N} x_i^m + a_1 \sum_{i=1}^{N} x_i^{m+1} + \ldots + a_m \sum_{i=1}^{N} x_i^{2m} = \sum_{i=1}^{N} x_i^m y_i
\end{cases}
\tag{43}
$$

问题11的解唯一

多项式拟合的一般方法可归纳为：

(1)根据具体问题，确定拟合多项式的次数n；（描点）

(2)计算正则方程组的系数和右端项

$$S_k=\sum_{i=0}^{m}x_i^k,\qquad t_k=\sum_{i=0}^{m}x_i^k\,y_i.$$

(3)写出正规方程组

(4)解正规方程组,求出$a_0,a_1,\ldots,a_n$;

(5)写出拟合多项式$P_n(x)$

例： 令 $x_0=0$，$x_1=1$，写出 $f(x)=e^{-x}$ 的一次多项式插值，并估计误差。

解：

记 $x_0=0$，$x_1=1$，$y_0=e^{-0}=1$，$y_1=e^{-1}$

则 $f(x)=e^{-x}$ 以 x_0，x_1 为插值节点的一次多项式为

$$p_1(x)=y_0\frac{x-x_1}{x_0-x_1}+y_1\frac{x-x_0}{x_1-x_0}=1\times\frac{x-1}{0-1}+e^{-1}\times\frac{x-0}{1-0}$$

$$=-(x-1)+e^{-1}x=1+(e^{-1}-1)x$$

因为 $y'(x)=-e^{-x}, y''(x)=e^{-x}$

所以

$$y(x)-p_1(x)=\frac{1}{2}y''(\xi)(x-x_0)(x-x_1)$$

$$=\frac{1}{2}e^{-\xi}(x-0)(x-1),\quad \xi\in(0,1)$$

故

$$|y(x)-p_1(x)|\le\frac{1}{2}\max_{0\le x\le 1}|e^{-x}|\cdot\max_{0\le x\le 1}|(x-0)(x-1)|$$

$$\le\frac{1}{2}\times 1\times\frac{1}{4}=\frac{1}{8}$$

作业

P54　6、11、12、13、16、17、31、36、37