第一章 插值与拟合

1.1 插值

在数学建模过程中，通常要处理由试验、测量得到的大量数据或一些过于复杂而不便于计算的函数表达式，针对此情况，很自然的想法就是，构造一个简单的函数作为要考察数据或复杂函数的近似。插值和拟合就可以解决这样的问题。

给定一组数据，需要确定满足特定要求的曲线，如果所求曲线通过所给定有限个数据点，这就是插值。插值函数类的取法有很多，可以是代数多项式，也可以是三角函数多项式或有理函数。由于代数多项式最简单，所以常用它来近似表达一些复杂的函数。

定义 1.1

已知未知函数在 $n+1$ 个互不相同的观测点 $x_0, x_1, x_2, \ldots, x_n$ 处的函数值：

$$f(x_0), f(x_1), \ldots, f(x_n)$$

寻求一个近似函数 $\phi(x)$，使之满足：

$$\phi(x_i) = y_i,\ i = 0, 1, \ldots, n$$

即：求一条近似曲线 $\phi(x)$，使其通过所有数据点：(x_i, y_i)，$i = 0, 1, \ldots, n$。

♣

观测点 $x_0, x_1, x_2, \ldots, x_n$ 称为插值节点，$f(x)$ 称为被插函数或原函数，为插值函数，$\phi(x_i) = y_i$ 称为插值条件，含 x_i 的最小区间 $[a, b]$ 称作插值区间，其中，$a = \min\{x_0, x_1, \ldots, x_n\}$，$b = \max\{x_0, x_1, \ldots, x_n\}$。

对任意插值点 $\hat{x}$，要估计该点的函数值 $f(\hat{x})$，就可以用 $\phi(\hat{x})$ 的值作为近似估计值。通常称此类建模问题为插值问题，而构造近似函数的方法就称为插值方法。

若 $\hat{x} \in [a, b]$，则称为内插，否则称为外推。值得注意的是，插值方法一般用于插值区间内部点的函数值估计或预测，利用该方法进行趋势外推预测时，可进行短期预测估计，对中长期预测并不适用。若插值函数为代数多项式，则该插值方法称为多项式插值。

1.1.1 利用待定系数法确定插值多项式

插值条件共含有 $n+1$ 个约束方程，而 n 次多项式也恰好有 $n+1$ 个待定系数。因此若已知 $y=f(x)$ 在 $n+1$ 个互不相同的观测点 $x_0, x_1, \ldots, x_n$ 处的观测值或函数值 $f(x_0), f(x_1), \ldots, f(x_n)$，则可以确定一个次数不超过 n 的多项式：

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

满足条件：

$$P_n(x_k) = y_k,\ k = 0, 1, \ldots, n$$

称 $P_n(x)$ 为 $f(x)$ 的 n 次插值多项式。

定理 1.1

n 次插值问题的解是存在而且唯一的。

证明：由 $P_n(x_k) = y_k,\ k = 0, 1, \ldots, n$ 可得：

$$\begin{cases} a_n {x_0}^n + a_{n-1}{x_0}^{n-1} + \cdots + a_1 x_0 + a_0 = y_0, \\ a_n {x_1}^n + a_{n-1}{x_1}^{n-1} + \cdots + a_1 x_1 + a_0 = y_1, \\ \ldots\ldots \\ a_n {x_n}^n + a_{n-1}{x_n}^{n-1} + \cdots + a_1 x_n + a_0 = y_n \end{cases}$$

注意到，系数矩阵的行列式：

$$|A| = \begin{vmatrix} x_0^n & \cdots & x_0 & 1 \\ x_1^n & \cdots & x_1 & 1 \\ \vdots & & \vdots & \vdots \\ x_n^n & \cdots & x_n & 1 \end{vmatrix}$$

为范德蒙德行列式，当 $x_0, x_1, \ldots, x_n$ 互不相同时，$|A| \neq 0$。由克莱姆法则，方程组的解存在且唯一。

说明：当插值次数越高，插值结果越偏离原函数，即龙格震荡现象。除此之外，待定系数法无法直接构造出插值多项式的表达式，插值多项式的次数每提高一次，都要重新求解，从而影响了方法的推广。

1.1.2 线性插值 ($n = 1$)

给定插值节点 x_0, x_1，其值分别为：$y_0 = f(x_0), y_1 = f(x_1)$。求线性插值多项式：

$$L_1(x) = a_0 + a_1 x$$

满足：

$$L_1(x_0) = y_0,\ L_1(x_1) = y_1$$

注意到，所求的多项式 $y = L_1(x)$ 为过点 $(x_0, y_0), (x_1, y_1)$ 的直线。由两点式方程可得：

$$\frac{x - x_0}{x_1 - x_0} = \frac{y - y_0}{y_1 - y_0}$$

等价于：

$$y = \frac{x - x_0}{x_1 - x_0} y_1 + \frac{x_1 - x}{x_1 - x_0} y_0$$

即：

$$L_1(x) = \frac{x_1 - x}{x_1 - x_0} y_0 + \frac{x - x_0}{x_1 - x_0} y_1$$

记

$$l_0(x) = \frac{x_1 - x}{x_1 - x_0},\ l_1(x) = \frac{x - x_0}{x_1 - x_0}$$

则有：

$$l_0(x_0) = 1,\ l_0(x_1) = 0$$

$$l_1(x_0) = 0,\ l_1(x_1) = 1$$

即：

$$l_k(x_j) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases},\ j, k = 0, 1$$

称 $l_0(x)$ 及 $l_1(x)$ 为线性插值基函数。

1.1.3 抛物插值 ($n = 2$)

问题：给定插值节点为 x_0, x_1, x_2，求二次插值多项式 $L_2(x)$，满足：

$$L_2(x_j) = y_j,\ j = 0, 1, 2$$

采用基函数方法。设

$$L_2(x) = l_0(x)y_0 + l_1(x)y_1 + l_2(x)y_2$$

此时，基函数 $l_0(x), l_1(x), l_2(x)$ 是二次函数，且在节点上满足：

$$\begin{cases} l_0(x_0) = 1,\ l_0(x_1) = 0,\ l_0(x_2) = 0 \\ l_1(x_0) = 0,\ l_1(x_1) = 1,\ l_1(x_2) = 0 \\ l_2(x_0) = 0,\ l_2(x_1) = 0,\ l_2(x_2) = 1 \end{cases}$$

即：

$$l_k(x_j) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases},\ j, k = 0, 1, 2.$$

满足上式的插值基函数很容易求出。如求 $l_0(x)$，因 x_1, x_2 为其零点，故可表为：

$$l_0(x) = A(x - x_1)(x - x_2)$$

其中 A 为待定系数，由 $l_0(x_0) = 1$ 可得：

$$A = \frac{1}{(x_0 - x_1)(x_0 - x_2)}$$

因此，

$$l_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}$$

同理可得：

$$l_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)},\ l_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

故有抛物插值公式：

$$L_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}y_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}y_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}y_2$$

1.1.4 Lagrange 插值多项式

问题：设有 $n+1$ 个互异节点 $x_0 < x_1 < \cdots < x_n$，满足：

$$y_i = f(x_i),\ i = 0, 1, 2, \ldots, n$$

构造 $L_n(x)$，使得：

$$L_n(x_j) = y_j,\ j = 0, 1, 2, \ldots, n$$

插值基函数：若 n 次多项式 $l_j(x),\ j = 0, 1, \ldots, n$ 在 $n+1$ 个节点 $x_0 < x_1 < \cdots < x_n$ 上满足条件：

$$l_k(x_j) = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases},\ j, k = 0, 1, \ldots, n$$

则称这 $n+1$ 个 n 次多项式 $l_0(x), l_1(x), \ldots, l_n(x)$ 为节点 $x_0, x_1, \ldots, x_n$ 上的 n 次插值基函数。

记

$$l_k(x) = \prod_{i=0, i\neq k}^{n} \frac{(x - x_i)}{(x_k - x_i)},\ k = 0, 1, 2, \ldots, n$$

则过 $n+1$ 个互异点 $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ 的插值函数为：

$$L_n(x) = l_0(x)y_0 + l_1(x)y_1 + \cdots + l_n(x)y_n$$

对于一个确定的区间，如果插值节点之间的距离较小，自然插值节点就增多，如果用一个多项式插值，自然次数就会升高。但是否次数越高，插值多项式的逼近效果越好呢？20 世纪初，Runge 就给出了一个等距节点插值多项式不收敛的例子。

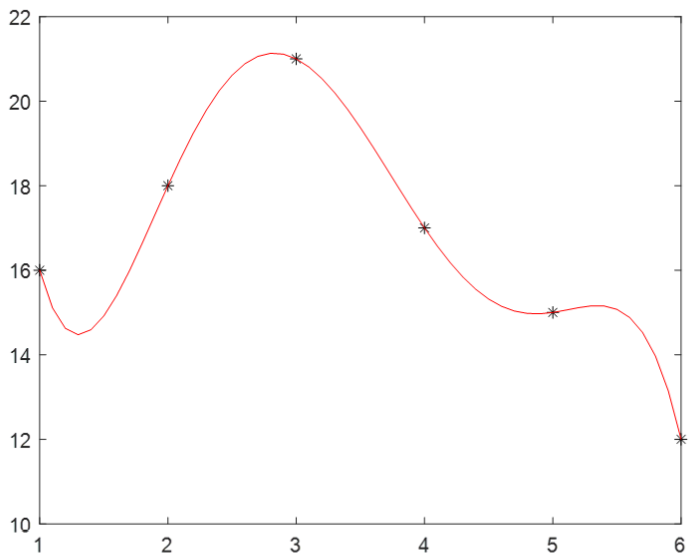
例题 1.1 已知未知函数 $y = f(x)$ 的 6 个观测点 $(x_i, y_i)(i = 0, 1, \ldots, 5)$ 的值如下表所示，试求插值函数，并求 $x = 3.5$ 处函数的估计值。

x_i	1	2	3	4	5	6
y_i	16	18	21	17	15	12

可以求得：5 次插值多项式为：

$$y = -0.2417x^5 + 4.3333x^4 - 28.9583x^3 + 87.6667x^2 - 115.8x + 69$$

插值函数的图像如图所示。



在 $x = 3.5$ 处，函数的估计值为：14.9180。代码图下图所示。

```
clc
clear
x = [1 2 3 4 5 6];
y = [16 18 21 17 15 12];
p = polyfit(x, y, 5);
plot(x, y, '*k')
hold on;
xx = 1: 0.1: 6;
yy = polyval(p, xx);
plot(xx, yy, '-r');
polyval(p, 3.5)
```

例题 1.2 某天文学家要确定一颗小行星绕太阳运行的轨道，他在轨道平面内建立以太阳为原点的直角坐标系，两坐标轴上的单位长度取为 1 天文测量单位（1 天文测量单位为地球到太阳的平均距离：1.496×10^8 千米）。在 5 个不同的时间对小行星作了 5 次观察，测得轨道上 5 个点的坐标数据见下表。由开普勒第一定律知，小行星的轨道为一椭圆，其一般方程可表示为：

$$a_1x^2 + a_2xy + a_3y^2 + a_4x + a_5y + 1 = 0.$$

请根据观测数据建立行星运行轨道的方程，并画出轨道曲线。

x	5.764	6.286	6.759	7.168	7.408
y	0.648	1.202	1.823	2.526	3.360

解：将天文学家所测的轨道上 5 个点的坐标数据代入椭圆轨道方程，可得下面的线性方程组：

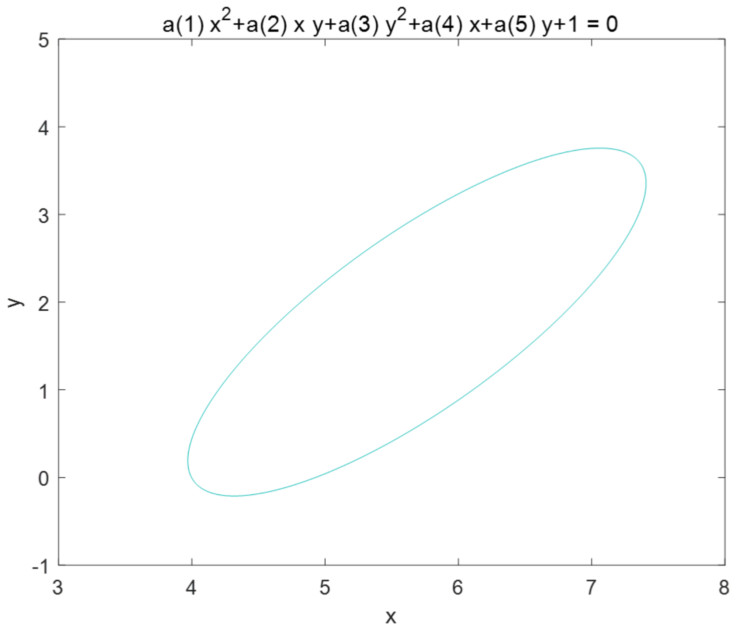
$$\begin{cases} a_1x_1^2 + a_2x_1y_1 + a_3y_1^2 + a_4x_1 + a_5y_1 = -1, \\ a_1x_2^2 + a_2x_2y_2 + a_3y_2^2 + a_4x_2 + a_5y_2 = -1, \\ a_1x_3^2 + a_2x_3y_3 + a_3y_3^2 + a_4x_3 + a_5y_3 = -1, \\ a_1x_4^2 + a_2x_4y_4 + a_3y_4^2 + a_4x_4 + a_5y_4 = -1, \\ a_1x_5^2 + a_2x_5y_5 + a_3y_5^2 + a_4x_5 + a_5y_5 = -1. \end{cases}$$

求解可得：

$$a_1 = 0.0508,\ a_2 = -0.0702,\ a_3 = 0.0381,\ a_4 = -0.4531,\ a_5 = 0.2643$$

即小行星轨道的椭圆方程为：

$$0.0508x^2 - 0.0702xy + 0.0381y^2 - 0.4531x + 0.2643y + 1 = 0$$



代码如下图所示。

```
clc
clear
x0 = [5.764; 6.286; 6.759; 7.168; 7.408];
y0 = [0.648; 1.202; 1.823; 2.526; 3.360];
A = [x0.^2, x0.*y0, y0.^2, x0, y0];
b = -ones(5,1);
a = A\b;
f = @(x,y) a(1)*x^2+a(2)*x* y+a(3)*y^2 + a(4)*x+ a(5)*y + 1;
ezplot(f, [3,8], [-1,5])
```

1.2 拟合

在数学建模过程中，通常要处理由试验、测量得到的大量数据或一些过于复杂而不便于计算的函数表达式，针对此情况，很自然的想法就是，构造一个简单的函数作为要考察数据或复杂函数的近似。插值和拟合就可以解决这样的问题。

给定一组二维数据，不要求曲线通过所有数据点 $(x_i, y_i)(i = 1, 2, \ldots, n)$，而是要求寻求一个函数（曲线）$y = f(x)$，使 $f(x)$ 在某种准则下与所有数据点最为接近，即反映对象整体的变化态势，这种要求曲线尽可能逼近给定数据的过程称为"拟合"。

1.2.1 偏差的平方和最小准则

记拟合函数 $f(x)$ 在 x_i 点处的偏差（或残差）为：

$$\delta_i = f(x_i) - y_i,\ i = 1, 2, \cdots, n$$

为使 $f(x)$ 在整体上尽可能与给定数据最为接近，可以采用"偏差的平方和最小"作为判定准则，即通过使

$$J = \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

达到最小值。该原则称为最小二乘原则，根据该原则确定拟合函数 $f(x)$ 的方法称为最小二乘法。

拟合函数应是自变量 x 和待定参数 $a_0, a_1, \ldots, a_m$ 的函数，即：

$$f(x) = f(x, a_0, a_1, \ldots, a_m)$$

按照 $f(x)$ 关于参数 $a_0, a_1, \ldots, a_m$ 的线性与否，可将最小二乘法也分为线性最小二乘法和非线性最小二乘法两类。

1.2.2 线性最小二乘法

给定一个线性无关的函数系 $\{\varphi_k(x)|k=0,1,\ldots,m\}$，拟合函数为如下的线性组合的形式：

$$f(x)=\sum_{k=0}^{m}a_k\varphi_k(x)$$

如取函数系为：$\{x^k|k=0,1,\ldots,m\}$，则有：

$$f(x)=a_mx^m+a_{m-1}x^{m-1}+\cdots+a_1x+a_0$$

如取函数系为：$\{\cos(kx)|k=0,1,\ldots,m\}$，则有：

$$f(x)=\sum_{k=0}^{m}a_k\cos(kx)$$

进而，可得残差的平方和：

$$J=\sum_{i=1}^{n}\left[\sum_{k=0}^{m}a_k\varphi_k(x_i)-y_i\right]^2$$

由多元函数求极值的必要条件：

$$\frac{\partial J}{\partial a_k}=0,\ k=0,1,\ldots,m$$

可得：

$$2\sum_{i=1}^{n}\varphi_k(x_i)\left[\sum_{k=0}^{m}a_k\varphi_k(x_i)-y_i\right]=0$$

即：

$$\sum_{i=1}^{n}\left[\sum_{k=0}^{m}a_k\varphi_k(x_i)\varphi_k(x_i)-\varphi_k(x_i)y_i\right]=0$$

也即是：

$$\sum_{i=1}^{n}\sum_{k=0}^{m}a_k\varphi_k(x_i)\varphi_k(x_i)=\sum_{i=1}^{n}\varphi_k(x_i)y_i$$

进而，

$$\sum_{k=0}^{m}\left[\sum_{i=1}^{n}\varphi_k(x_i)\varphi_k(x_i)\right]a_k=\sum_{i=1}^{n}\varphi_k(x_i)y_i\ ,\ k=0,1,\ldots,m$$

该式为一个关于 $a_0,a_1,\ldots,a_m$ 的线性方程组，称为正规方程组。
记

$$R=\begin{pmatrix}\varphi_1(x_1)&\varphi_2(x_1)&\cdots&\varphi_m(x_1)\\\varphi_1(x_2)&\varphi_2(x_2)&\cdots&\varphi_m(x_2)\\\vdots&\vdots&&\vdots\\\varphi_1(x_n)&\varphi_2(x_n)&\cdots&\varphi_m(x_n)\end{pmatrix},\ A=\begin{pmatrix}a_1\\a_2\\\vdots\\a_m\end{pmatrix},\ Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_m\end{pmatrix}$$

则正规方程组可表示为：

$$R^TRA=R^TY$$

当矩阵 R 是列满秩时，系数矩阵 R^TR 是可逆的。于是，正规方程组有唯一解：

$$A=(R^TR)^{-1}R^TY$$

该解为拟合函数的系数，从而求得了最小二乘拟合函数。

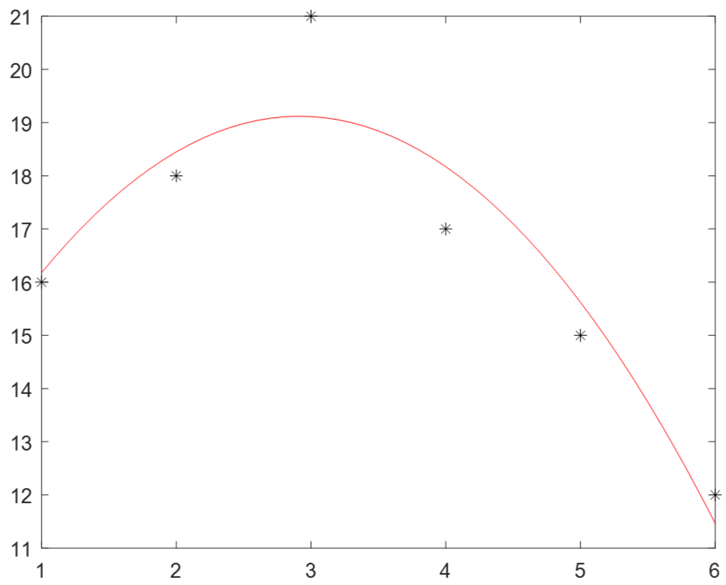
例题 1.3 已知未知函数 $y = f(x)$ 的 6 个观测点 $(x_i, y_i)(i = 0, 1, \ldots, 5)$ 的值如下表所示，分别给出 2 次和 4 次的拟合函数。

x_i	1	2	3	4	5	6
y_i	16	18	21	17	15	12

取 $n = 2$，则可得到 2 次拟合函数：

$$y = -0.8036x^2 + 4.6821x + 12.3000$$

拟合的效果如图所示。



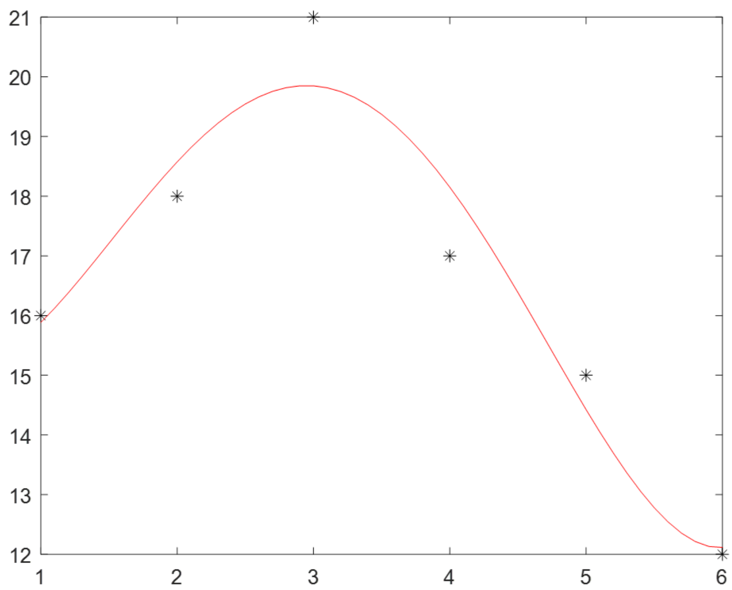
代码如下。

```
clc
clear
x = [1 2 3 4 5 6];
y = [16 18 21 17 15 12];
p = polyfit(x, y, 2);
plot(x, y, '*k')
hold on;
xx = 1: 0.1: 6;
yy = polyval(p, xx);
plot(xx, yy, '-r')
```

取 $n = 4$，则可得到 4 次拟合函数：

$$y = 0.1042x^4 - 1.3009x^3 + 4.4931x^2 - 3.2447x + 15.8333$$

拟合的结果及代码分别如下。
代码如下。




```
clc
clear
x = [1 2 3 4 5 6];
y = [16 18 21 17 15 12];
p = polyfit(x, y, 4);
plot(x, y, '*k')
hold on;
xx = 1: 0.1: 6;
yy = polyval(p, xx);
plot(xx, yy, '-r')
```

1.2.3 非线性最小二乘法

对于给定一个线性无关的函数系 $\{\varphi_k(x)|k=0,1,\dots,m\}$，拟合函数不是线性组合的形式，称

$$f(x)=f(x,a_0,a_1,\dots,a_m)$$

为关于参数 $a_0,a_1,\dots,a_m$ 的非线性函数。

将 $f(x)$ 的表达式代入 $J=\sum\limits_{i=1}^{n}(f(x_i)-y_i)^2$，则问题转化为非线性函数的极小化问题。为得到最小二乘拟合函数的具体表达式，可用非线性优化方法求解出参数 $a_0,a_1,\dots,a_m$。

数据拟合时，首要也是最关键的一步就是选取恰当的拟合函数。如果能够根据问题的背景通过机理分析得到变量之间的函数关系，那么只需估计相应的参数即可。但很多情况下，问题的机理并不清楚。此时，一个较为自然的方法是先做出数据的散点图，从直观上判断应选用什么样的拟合函数。一般来讲，如果数据分布接近于直线，则宜选用线性函数拟合；如果数据分布接近于抛物线，则宜选用二次多项式拟合；如果数据分布特点是开始上升较快随后逐渐变缓，则可选用双曲线型函数：

$$f(x)=\frac{x}{a_1x+a_2}$$

或指数型函数

$$f(x)=a_1e^{-\frac{a_2}{x}}$$

如果数据分布特点是开始下降较快随后逐渐变缓，则可选用如下的函数：

$$f(x)=\frac{1}{a_1x+a_2},\ f(x)=\frac{1}{a_1x^2+a_2},\ f(x)=a_1e^{-a_2x}$$

除此之外，常被选用的非线性拟合函数有对数函数：

$$y=a_1+a_2\ln x$$

S 形曲线函数

$$y=\frac{1}{a+be^{-x}}$$

1.3 模型的应用

参见优秀论文 2023-C-01。