Phương pháp tính

Chương 4: Phương pháp nội suy và phương pháp bình phương cực tiểu

Chu Bình Minh

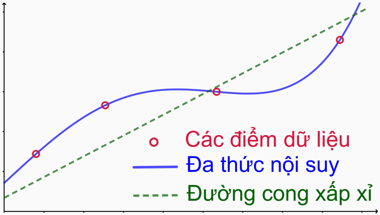
Khoa khoa học cơ bản, Trường ĐH Kinh tế - Kỹ thuật Công nghiệp

Table of contents

GIỚI THIỆU

Các dữ liệu thu được thường được cho dưới dạng bảng:

x_1	x_2	x_3	$\cdots$	x_n
y_1	y_2	y_3	$\cdots$	y_n



TÌM ĐƯỜNG CONG XẤP XỈ CÁC ĐIỂM DỮ LIỆU?

- Dữ liệu chính xác cao: Sử dụng đường cong đi qua các điểm
- Dữ liệu nhiễu lớn: Sử dụng đường cong đi qua nhóm các điểm

NỘI SUY ĐA THỨC

Nội suy đa thức là trường hợp phổ biến nhất của phương pháp nội suy. Với n điểm dữ liệu cho trước, luôn tồn tại duy nhất đa thức bậc $n-1$ đi qua các điểm dữ liệu này. Chẳng hạn, với hai điểm dữ liệu thì ta luôn xác định được một đường thẳng (đa thức bậc một) qua hai điểm này; với ba điểm dữ liệu thì tồn tại một parabola (đa thức bậc hai) đi qua ba điểm này. Mục này sẽ giới thiệu một số cách xây dựng đa thức nội suy đi qua n điểm dữ liệu.

Đa thức nội suy Lagrange

Cho n điểm dữ liệu

x_1	x_2	x_3	$\cdots$	x_n
y_1	y_2	y_3	$\cdots$	y_n

$$P_{n-1}(x) = \sum_{i=1}^{n} y_i l_i(x), \tag{1}$$

với

$$\begin{aligned} l_i(x) &= \frac{x - x_1}{x_i - x_1} . \frac{x - x_2}{x_i - x_2} \cdots \frac{x - x_{i-1}}{x_i - x_{i-1}} . \frac{x - x_{i+1}}{x_i - x_{i+1}} \cdots \frac{x - x_n}{x_i - x_n} \\ &= \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \end{aligned} \tag{2}$$

được gọi là *các hàm cơ bản*.

Đa thức nội suy Lagrange

Cho 2 điểm dữ liệu

x_1	x_2
y_1	y_2

$P_1(x) = y_1 l_1(x) + y_2 l_2(x)$, với

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

Cho 2 điểm dữ liệu

x_1	x_2
y_1	y_2

$P_1(x) = y_1 l_1(x) + y_2 l_2(x)$, với

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

Cho 3 điểm dữ liệu

x_1	x_2	x_3
y_1	y_2	y_3

$P_1(x) = y_1 l_1(x) + y_2 l_2(x) + y_3 l_3(x)$, với

$$l_1(x) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)}; l_2(x) = \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)};$$

$$l_3(x) = \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)}.$$

VÍ DỤ Cho các điểm dữ liệu

x	0	2	3
y	7	11	28

sử dụng phương pháp Lagrange xác định y tại $x = 1$.

GIẢI

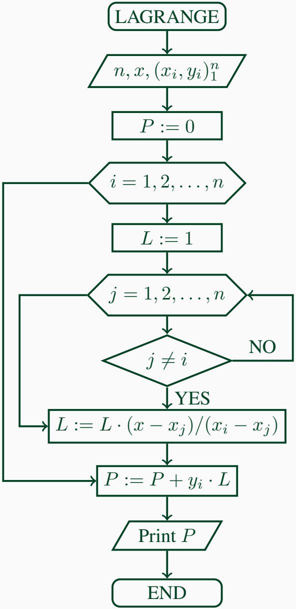
$$l_1(1) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} = \frac{(1-2)(1-3)}{(0-2)(0-3)} = \frac{1}{3}$$

$$l_2(1) = \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} = \frac{(1-0)(1-3)}{(2-0)(2-3)} = 1$$

$$l_3(1) = \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} = \frac{(1-0)(1-2)}{(3-0)(3-2)} = -\frac{1}{3}$$

$$y(1) = y_1 l_1(1) + y_2 l_2(1) + y_3 l_3(1) = \frac{7}{3} + 11 - \frac{28}{3} = 4.$$

Đa thức nội suy Lagrange



Phương pháp Lagrange

- Ưu điểm: Đơn giản
- Nhược điểm: thêm mốc nội suy sẽ phải tính lại từ đầu

Phương pháp Lagrange

- Ưu điểm: Đơn giản
- Nhược điểm: thêm mốc nội suy sẽ phải tính lại từ đầu

Cho n điểm dữ liệu

x_1	x_2	x_3	$\cdots$	x_n
y_1	y_2	y_3	$\cdots$	y_n

đa thức nội suy Newton:

$$P_{n-1}(x) = a_1 + (x - x_1)a_2 + (x - x_1)(x - x_2)a_3 + \cdots + (x - x_1)(x - x_2)\ldots(x - x_{n-1})a_n.$$

Đa thức nội suy Newton

$n = 4$, ta có đa thức nội suy bậc 3 là

$$\begin{aligned} P_3(x) &= a_1 + (x - x_1)a_2 + (x - x_1)(x - x_2)a_3 + (x - x_1)(x - x_2)(x - x_3)a_4 \\ &= a_1 + (x - x_1)\left\{a_2 + (x - x_2)[a_3 + (x - x_3)a_4]\right\}. \end{aligned}$$

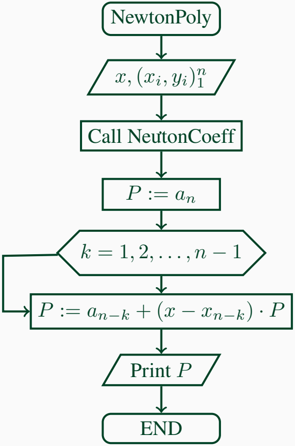
Ta có thể tính $P_3(x)$ bằng cách thay thế ngược

$$\begin{aligned} P_0(x) &= a_4 \\ P_1(x) &= a_3 + (x - x_3)P_0(x) \\ P_2(x) &= a_2 + (x - x_2)P_1(x) \\ P_3(x) &= a_1 + (x - x_1)P_2(x). \end{aligned}$$

ĐA THỨC NỘI SUY NEWTON

Với n tuỳ ý, ta có

$$P_0(x) = a_n, \quad P_k(x) = a_{n-k} + (x - x_{n-k})P_{k-1}(x), \quad k = 1, 2, \ldots n - 1. \tag{3}$$



$$P_{n-1}(x) = a_1 + (x - x_1)a_2 + \cdots + (x - x_1)(x - x_2)\ldots(x - x_{n-1})a_n.$$

Tìm hệ số a_i?

$$P_{n-1}(x) = a_1 + (x - x_1)a_2 + \cdots + (x - x_1)(x - x_2)\ldots(x - x_{n-1})a_n.$$

Tìm hệ số a_i?

$$a_1 = y_1, \quad a_2 = \nabla y_2, \quad a_3 = \nabla^2 y_3, \ldots, a_n = \nabla^{n-1} y_n, \tag{4}$$

với

$$\begin{aligned}
\nabla y_i &= \frac{y_i - y_1}{x_i - x_1}, \quad i = 2, 3, \ldots, n \\
\nabla^2 y_i &= \frac{\nabla y_i - \nabla y_2}{x_i - x_2}, \quad i = 3, 4, \ldots, n \\
\nabla^3 y_i &= \frac{\nabla^2 y_i - \nabla^2 y_3}{x_i - x_3}, \quad i = 4, 5, \ldots, n \\
&\vdots \\
\nabla^{n-1} y_n &= \frac{\nabla^{n-2} y_n - \nabla^{n-2} y_{n-1}}{x_n - x_{n-1}}
\end{aligned} \tag{5}$$

ĐA THỨC NỘI SUY NEWTON

Với $n = 5$

Bảng 1: Bảng tính toán hệ số cho đa thức nội suy Newton với $\boldsymbol{n = 5}$.

x_1	y_1				
x_2	y_2	∇y_2			
x_3	y_3	∇y_3	$\nabla^2 y_3$		
x_4	y_4	∇y_4	$\nabla^2 y_4$	$\nabla^3 y_4$	
x_5	y_5	∇y_5	$\nabla^2 y_5$	$\nabla^3 y_5$	$\nabla^4 y_5$

VÍ DỤ Các điểm dữ liệu

x	-2	1	4	-1	3	-4
y	-1	2	59	4	24	-53

nằm trên một đa thức. Xác định bậc của đa thức này bằng cách xây dựng bảng sai phân.

VÍ DỤ Các điểm dữ liệu

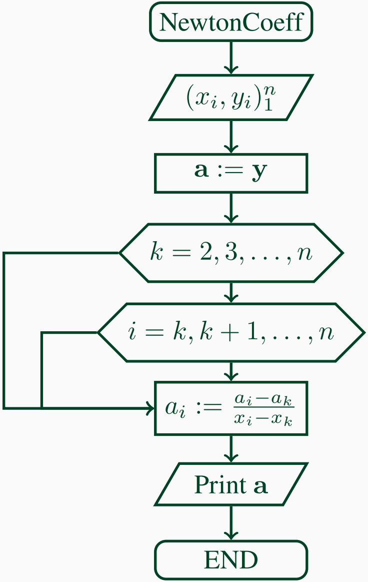
x	-2	1	4	-1	3	-4
y	-1	2	59	4	24	-53

nằm trên một đa thức. Xác định bậc của đa thức này bằng cách xây dựng bảng sai phân.

GIẢI

i	x_i	y_i	∇y_i	$\nabla^2 y_i$	$\nabla^3 y_i$	$\nabla^4 y_i$	$\nabla^5 y$
1	-2	-1					
2	1	2	1				
3	4	59	10	3			
4	-1	4	5	-2	1		
5	3	24	5	2	1	0	
6	-4	-53	26	-5	1	0	0

Đa thức nội suy Newton



Phương pháp nội suy Newton

- Ưu điểm: Không cần tính lại khi thêm mốc nội suy.
- Nhược điểm: Không hiệu quả khi tính giá trị đa thức tại một điểm cho trước (Bước 1: Lập đa thức nội suy; Bước 2: Thay giá trị vào đa thức).

Phương pháp nội suy Newton

- Ưu điểm: Không cần tính lại khi thêm mốc nội suy.
- Nhược điểm: Không hiệu quả khi tính giá trị đa thức tại một điểm cho trước (Bước 1: Lập đa thức nội suy; Bước 2: Thay giá trị vào đa thức).

$P_k[x_i, x_{i+1}, \ldots, x_{i+k}](x)$ là đa thức bậc k đi qua $k+1$ điểm dữ liệu $(x_i, y_i), (x_{i+1}, y_{i+1}), \ldots, (x_{i+k}, y_{i+k})$.

$$P_k[x_i, x_{i+1}, \ldots, x_{i+k}](x) = \frac{(x - x_{i+k})P_{k-1}[x_i, x_{i+1}, \ldots, x_{i+k-1}](x) + (x_i - x)P_{k-1}[x_{i+1}, x_{i+2}, \ldots, x_{i+k}](x)}{x_i - x_{i+k}}. \tag{6}$$

Với một điểm dữ liệu ta có

$$P_0[x_i] = y_i. \tag{7}$$

Nội suy dựa vào hai điểm dữ liệu là

$$P_1[x_i, x_{i+1}](x) = \frac{(x - x_{i+1})P_0[x_i] + (x_i - x)P_0[x_{i+1}]}{x_i - x_{i+1}}.$$

Nội suy ba điểm là

$$P_2[x_i, x_{i+1}, x_{i+2}](x) = \frac{(x - x_{i+2})P_1[x_i, x_{i+1}](x) + (x_i - x)P_1[x_{i+1}, x_{i+2}](x)}{x_i - x_{i+2}}.$$

VÍ DỤ Cho các điểm dữ liệu

x	4.0	4.9	3.8	3.7
y	-0.06604	-0.02724	0.01282	0.05383

xác định nghiệm của $y(x) = 0$ bằng phương pháp Neville.

VÍ DỤ Cho các điểm dữ liệu

x	4.0	4.9	3.8	3.7
y	-0.06604	-0.02724	0.01282	0.05383

xác định nghiệm của $y(x) = 0$ bằng phương pháp Neville.
GIẢI Thay vì tính y tại giá trị x
cho trước, chúng ta tính x tại giá trị y cho trước (trường hợp này $y = 0$).

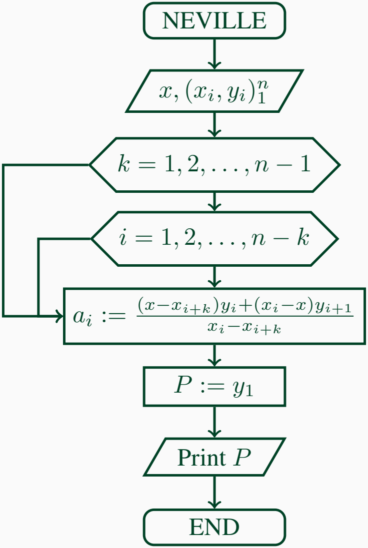
i	y_i	$P_0[] = x_i$	$P_1[,]$	$P_2[,,]$	$P_3[,,,]$
1	-0.06604	4.0	3.8298	3.8316	3.8317
2	-0.02724	3.9	3.8320	3.8318	
3	0.01282	3.8	3.8313		
4	0.05383	3.7			

Hai tính toán sau minh hoạ các tính toán trong bảng trên.

$$P_1[y_1, y_2] = \frac{(y - y_2)P_0[y_1] + (y_1 - y)P_0[y_2]}{y_1 - y_2}$$
$$= \frac{(0 + 0.02724)4.0 + (-0.06604 - 0)3.9}{-0.06604 + 0.02724} = 3.8298.$$
$$P_2[y_2, y_3, y_4] = \frac{(y - y_4)P_1[y_2, y_3] + (y_2 - y)P_1[y_3, y_4]}{y_2 - y_4}$$
$$= \frac{(0 - 0.05383)3.8320 + (-0.02724 - 0)3.8313}{-0.02724 - 0.05383} = 3.8318.$$

Tất cả các Ps trong bảng trên đều được tính toán với $y = 0$ nên Ps là nghiệm của đa thức nội suy đi qua các điểm tương ứng. Chẳng hạn $P_1[y_1, y_2]$ là nghiệm của đa thức bậc một đi qua hai điểm đầu tiên, $P_2[y_2, y_3, y_4]$ là nghiệm của đa thức bậc hai đi qua ba điểm cuối cùng. Do đó, nghiệm của đa thức bậc ba đi qua bốn điểm dữ liệu là $x = P_3[y_1, y_2, y_3, y_4] = 3.8317$.

PHƯƠNG PHÁP NỘI SUY NEVILLE



PHƯƠNG PHÁP BÌNH PHƯƠNG NHỎ NHẤT

Các phương pháp nội suy:

- Ưu điểm: Đơn giản.
- Nhược điểm:
  1. Khi số điểm dữ liệu lớn thì bậc đa thức nội suy lớn;
  2. Khi các điểm dữ liệu có nhiễu thì nhiều khi đã thức nội suy sẽ không chính xác.

Các phương pháp nội suy:

- Ưu điểm: Đơn giản.
- Nhược điểm:
  1. Khi số điểm dữ liệu lớn thì bậc đa thức nội suy lớn;
  2. Khi các điểm dữ liệu có nhiễu thì nhiều khi đã thức nội suy sẽ không chính xác.

Phương pháp bình phương nhỏ nhất
tìm đường cong trơn phù hợp với điểm dữ liệu theo nghĩa trung bình. Đường cong này phải có dạng đơn giản (ví dụ: đa thức bậc thấp), để không tái tạo nhiễu.
Do vậy, việc tìm hàm xấp xỉ được thực hiện qua hai bước: trước tiên tìm dạng hàm số và sau đó xấp xỉ một cách tốt nhất các tham số bằng cách dựa vào các dữ liệu.

Điều này đặt ra câu hỏi: "Xấp xỉ tốt nhất" nghĩa là gì? Nếu các nhiễu chỉ hạn chế trên các biến y thì phương pháp phổ biến là phương pháp *bình phương nhỏ nhất*, đây là phương pháp cực tiểu hàm số

$$S(a_1, \ldots, a_m) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \tag{8}$$

phụ thuộc vào a_j. Do đó, các giá trị tối ưu của các tham số được tính bằng cách giải hệ phương trình

$$\frac{\partial S}{\partial a_k} = 0, k = 1, 2, \ldots, m. \tag{9}$$

Các đại lượng $r_i = y_i - f(x_i)$ trong (8) được gọi là phần dư, chúng biểu diễn sự khác biệt giữa các điểm dữ liệu và hàm xấp xỉ tại x_i. Hàm S cần cực tiểu là tổng của bình phương các phần dư. Phương trình (9) cho trường hợp tổng quát, nếu phi tuyến với các a_j thì bài toán rất khó giải.

Nếu hàm xấp xỉ $f(x)$ được chọn dưới dạng tổ hợp tuyến tính của các hàm đặc biệt $f_j(x)$:

$$f(x) = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_m f_m(x)$$

thì hệ phương trình (9) là hệ phương trình tuyến tính. Hàm xấp xỉ là đa thức khi $f_1(x) = 1, f_2(x) = x, f_3(x) = x^2, \ldots$.

Độ phân tán của dữ liệu để xấp xỉ hàm số được đo bởi đại lượng gọi là *độ lệch chuẩn* và được xác định bởi

$$\sigma = \sqrt{\frac{S}{n-m}}. \tag{10}$$

Chú ý rằng khi $m = n$ ta có nội suy đa thức. Trong trường hợp này, cả tử số và mẫu số của công thức (10) đều bằng không nên σ không có nghĩa.

Xấp xỉ dữ liệu bởi đường thẳng

$$f(x) = a + bx \tag{11}$$

còn được biết đến với tên gọi *hồi quy tuyến tính*. Hàm cần cực tiểu là

$$S(a, b) = \sum_{i=1}^{n} [y_i - a - bx_i]^2 .$$

Hệ phương trình (9) trở thành

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} -2(y_i - a - bx_i) = 2\left(-\sum_{i=1}^{n} y_i + na + b\sum_{i=1}^{n} x_i\right) = 0$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} -2(y_i - a - bx_i)x_i = 2\left(-\sum_{i=1}^{n} x_i y_i + na\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2\right) = 0.$$

Chia hai vế các phương trình cho $2n$ ta có

$$a + \bar{x}b = \bar{y}, \qquad a\bar{x} + \left(\frac{1}{n}\sum_{1}^{n} x_i^2\right) b = \frac{1}{n}\sum_{i=1}^{n} x_i y_i,$$

với

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_1, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

là giá trị trung bình của dữ liệu x và y. Giải các phương trình cho ta các tham số là

$$a = \frac{\bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i}{\sum x_i^2 - n\bar{x}^2}, \qquad b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}. \tag{12}$$

Các biểu thức này dễ bị lỗi làm tròn nên các tham số nên được tính theo công thức

$$b = \frac{\sum y_i(x_i - \bar{x})}{\sum x_i(x_i - \bar{x})}, \qquad a = \bar{y} - \bar{x}b. \tag{13}$$

tương đương với công thức (12) mà sai số làm tròn giảm hơn.

VÍ DỤ Xấp xỉ cho dữ liệu dưới đây bằng đường thẳng và tính độ lệch chuẩn.

x	0.0	1.0	2.0	2.5	3.0
y	2.9	3.7	4.1	4.4	5.0

VÍ DỤ Xấp xỉ cho dữ liệu dưới đây bằng đường thẳng và tính độ lệch chuẩn.

x	0.0	1.0	2.0	2.5	3.0
y	2.9	3.7	4.1	4.4	5.0

GIẢI Trung bình của dữ liệu là

$$\bar{x} = \frac{1}{5}\sum x_i = \frac{0.0 + 1.0 + 2.0 + 2.5 + 3.0}{5} = 1.7$$
$$\bar{y} = \frac{1}{5}\sum y_i = \frac{2.9 + 3.7 + 4.1 + 4.4 + 5.0}{5} = 4.02.$$

Từ (13) ta có

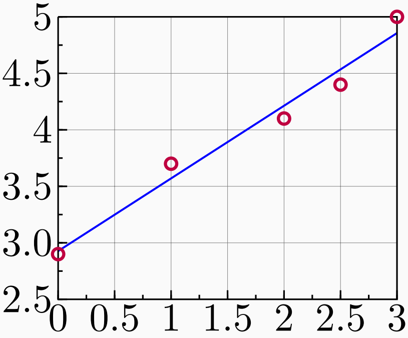
$$b = \frac{\sum y_i(x_i - \bar{x})}{\sum x_i(x_i - \bar{x})} = 0.6431$$
$$a = \bar{y} - \bar{x}b = 2.927.$$

Do đó, đường hồi quy là $f(x) = 2.927 + 0.6431x$.

Phương trình hồi quy tuyến tính

x	0.0	1.0	2.0	2.5	3.0
y	2.9	3.7	4.1	4.4	5.0

Phương trình hồi quy là $f(x) = 2.927 + 0.6431x$



Chúng ta bắt đầu đánh giá độ lệch chuẩn bằng cách tính các phần dư

y	2.900	3.700	4.100	4.400	5.000
$f(x)$	2.927	3.570	4.213	4.535	4.856
$y - f(x)$	-0.027	0.130	-0.113	-0.135	0.144

Tổng bình phương của các phần dư là

$$S = \sum [y_i - f(x_i)]^2$$
$$= (-0.027)^2 + (0.130)^2 + (-0.113)^2 + (-0.135)^2 + (0.144)^2 = 0.06936$$

nên độ lệch chuẩn trong (10) trở thành

$$\sigma = \sqrt{\frac{S}{n-m}} = \sqrt{\frac{0.06936}{5-2}} = 0.1520.$$

Xét phương pháp xấp xỉ bình phương nhỏ nhất dạng tuyến tính

$$f(x) = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_m f_m(x) = \sum_{j=1}^{m} a_j f_j(x), \tag{14}$$

với $f_i(x)$ là hàm số xác định trước của x, được gọi là các *hàm cơ sở*. Thay vào (8) ta có

$$S = \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{m} a_j f_j(x_i) \right]. \tag{a}$$

Do đó, hệ phương trình (9) là

$$\frac{\partial S}{\partial a_k} = -2 \left\{ \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{m} a_j f_j(x_i) \right] f_k(x_i) \right\} = 0, k = 1, 2, \ldots, m.$$

Bỏ -2 và chuyển vế ta có

$$\sum_{j=1}^{m}\left[\sum_{i=1}^{n} f_j(x_i)f_k(x_i)\right]a_j = \sum_{i=1}^{n} f_k(x_i)y_i, k = 1, 2, \ldots, m.$$

Dạng ma trận của hệ phương trình này là

$$\mathbf{Aa} = \mathbf{b}, \tag{15}$$

với

$$a_{kj} = \sum_{i=1}^{n} f_j(x_i)f_k(x_i), \qquad b_k = \sum_{i=1}^{n} f_k(x_i)y_i. \tag{16}$$

Hệ phương trình (15) được gọi là *hệ phương trình chuẩn* của phương pháp xấp xỉ bình phương nhỏ nhất và được giải bằng các phương pháp đã thảo luận trong Chương 3. Chú ý rằng ma trận hệ số có tính đối xứng, tức là $a_{kj} = a_{jk}$.

Nếu đa thức xấp xỉ có bậc là $m-1$, ta có $f(x) = \sum_{j=1}^{m} a_j x^{j-1}$. Các hàm cơ sở là

$$f_j(x) = x^{j-1}, j = 1, 2, \ldots, m. \tag{17}$$

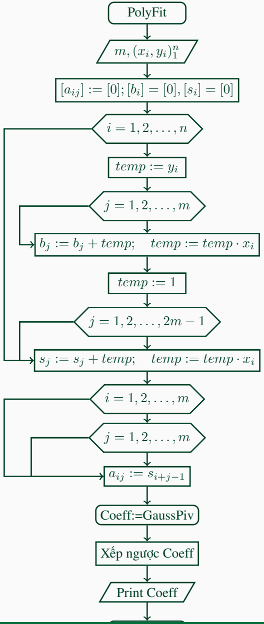
Khi đó, hệ số của hệ phương trình chuẩn là

$$a_{kj} = \sum_{i=1}^{n} x_i^{j+k-2}, \qquad b_k = \sum_{i=1}^{n} x_i^{k-1} y_i$$

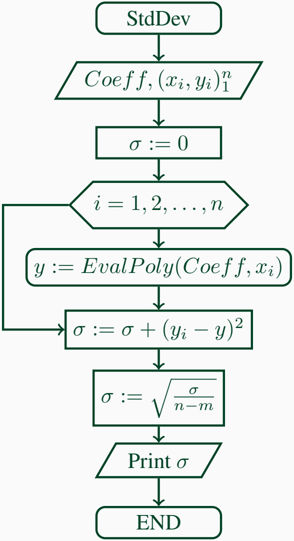
hoặc

$$\mathbf{A} = \begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m-1} \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_i^{m-1} & \sum x_i^m & \sum x_i^{m+1} & \cdots & \sum x_i^{2m-2} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^{m-1} y_i \end{bmatrix}, \tag{18}$$

XẤP XỈ DẠNG ĐA THỨC



XẤP XỈ DẠNG ĐA THỨC



Dữ liệu có trọng số

Trong một số trường hợp, niềm tin vào độ chính xác của dữ liệu thay đổi theo từng điểm. Chẳng hạn, công cụ thực hiện các phép đo có thể nhạy hơn trong một phạm vi dữ liệu nhất định. Đôi khi dữ liệu đại diện cho kết quả của một số thực nghiệm và mỗi thực nghiệm này lại được thực hiện dưới các tiêu chuẩn khác nhau. Trong các trường hợp này, chúng ta có thể gán hệ số tin cậy hoặc trọng số cho từng điểm dữ liệu và cực tiểu hóa tổng bình phương của *phần dư có trọng số* $r_i = W_i[y_i - f_i(x_i)]$, với W_i là các trọng số. Do đó, hàm cần được cực tiểu hoá là

$$S(a_1, \ldots, a_m) = \sum_{i=1}^{n} W_i^2 \left[y_i - f(x_i)\right]^2 . \tag{19}$$

Quy trình này buộc hàm xấp xỉ $f(x)$ gần hơn với các điểm dữ liệu có trọng số cao hơn.

Hồi quy tuyến tính có trọng số.

Nếu hàm xấp xỉ là đường thẳng $f(x) = a + bx$ thì (19) trở thành

$$S(a,b) = \sum_{i=1}^{n} W_i^2 \left[y_i - a - bx_i\right]^2 . \tag{20}$$

Điều kiện để cực tiểu hàm S là

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{n} W_i^2[y_i - a - bx_i],$$

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{n} W_i^2[y_i - a - bx_i]x_i$$

hoặc

$$a\sum_{i=1}^{n} W_i^2 + b\sum_{i=1}^{n} W_i^2 x_i = \sum_{i=1}^{n} W_i^2 y_i \tag{21}$$

$$a\sum_{i=1}^{n} W_i^2 x_i + b\sum_{i=1}^{n} W_i^2 x_i^2 = \sum_{i=1}^{n} W_i^2 x_i y_i. \tag{22}$$

Chia hai vế (21) cho $\sum W_i^2$ và đặt các trung bình trọng số

$$\hat{x} = \frac{\sum W_i^2 x_i}{\sum W_i^2}, \quad \hat{y} = \frac{\sum W_i^2 y_i}{\sum W_i^2} \tag{23}$$

ta có

$$a = \hat{y} - b\hat{x}. \tag{24}$$

Thay (24) vào (22) và tính toán ta thu được

$$b = \frac{\sum_{i=1}^{n} W_i^2 y_i (x_i - \hat{x})}{\sum_{i=1}^{n} W_i^2 x_i (x_i - \hat{x})}. \tag{25}$$

Chú ý rằng các công thức (24)-(25) tương tự như (13) trong trường hợp không có trọng số.

Xấp xỉ bởi hàm mũ.
Xét một ví dụ hàm xấp xỉ dạng

$$f(x) = ae^{bx}.$$

Nếu giải bằng phương pháp bình phương nhỏ nhất sẽ dẫn đến hệ phương trình phi tuyến đối với a, b. Nhưng nếu ta xấp xỉ $\ln y$ thay vì y thì bài toán sẽ được chuyển về dạng hồi quy tuyến tính: tìm hàm xấp xỉ

$$F(x) = \ln f(x) = \ln a + bx$$

với các điểm dữ liệu $(x_i, \ln y_i)$, $i = 1, 2, \ldots, n$. Cái giá của sự đơn giản này là: thay vì áp dụng phương pháp bình phương nhỏ nhất cho dữ liệu thì ta phải áp dụng phương pháp bình phương nhỏ nhất cho logarit của dữ liệu. Các phần dư của xấp xỉ logarit là

$$R_i = \ln y_i - F(x_i) = \ln y_i - \ln a - bx_i \tag{26}$$

trong khi phần dư được sử dụng trong xấp xỉ dữ liệu gốc là

$$r_i = y_i - f(x_i) = y_i - ae^{bx_i}. \tag{27}$$

Sự khác biệt này có thể được loại bỏ phần lớn bằng trọng số từ xấp xỉ logarit. Từ (27) ta có $\ln(r_i - y_i) = \ln ae^{bx_i} = \ln a + bx_i$ nên (26) được viết lại dưới dạng

$$R_i = \ln y_i - \ln(r_i - y_i) = \ln\left(1 - \frac{r_i}{y_i}\right).$$

Nếu phần dư r_i đủ nhỏ ($r_i \ll y_i$) ta có thể xấp xỉ $\ln(1 - r_i/y_i) \approx r_i/y_i$, do đó

$$R_i \approx \frac{r_i}{y_i}.$$

Bây giờ ta có thể thấy rằng, bằng cách cực tiểu hoá $\sum R_i^2$, chúng ta đã tình cờ giới thiệu trọng số $1/y_i$. Quá trình này có thể bị phủ nhận nếu chúng ta áp dụng các trọng số y_i khi xấp xỉ hàm $F(x)$ cho các dữ liệu $(\ln y_i, x_i)$ bằng cách cực tiểu hoá

$$S = \sum_{i=1}^{n} y_i^2 R_i^2. \tag{28}$$

DỮ LIỆU CÓ TRỌNG SỐ

Các ví dụ hữu ích khác khi lấy trong số $W_i = y_i$ được cho trong bảng sau.

$f(x)$	$F(x)$	Dữ liệu được xấp xỉ cho $F(x)$
axe^{bx}	$\ln[f(x)/x] = \ln a + bx$	$[x_i, \ln(y_i/x_i)]$
ax^b	$\ln f(x) = \ln a + b\ln x$	$(\ln x_i, \ln y_i)$

Bảng 2: Bảng xấp xỉ bình phương nhỏ nhất với trường hợp dữ liệu có trọng số.

VÍ DỤ Xác định tham số a, b sao cho hàm $f(x) = ae^{bx}$ là hàm xấp xỉ dữ liệu dưới đây bằng phương pháp bình phương nhỏ nhất

x	1.2	2.8	4.3	5.4	6.8	7.9
y	7.5	16.1	38.9	67.0	146.6	266.2

theo hai cách: (1) xỉ $\ln y_i$ và (2) xỉ $\ln y_i$ với các trọng số $W_i = y_i$. Tính độ lệch chuẩn cho mỗi trường hợp.

VÍ DỤ Xác định tham số a, b sao cho hàm $f(x) = ae^{bx}$ là hàm xấp xỉ dữ liệu dưới đây bằng phương pháp bình phương nhỏ nhất

x	1.2	2.8	4.3	5.4	6.8	7.9
y	7.5	16.1	38.9	67.0	146.6	266.2

theo hai cách: (1) xỉ $\ln y_i$ và (2) xỉ $\ln y_i$ với các trọng số $W_i = y_i$. Tính độ lệch chuẩn cho mỗi trường hợp.

GIẢI

Trường hợp (1). Bài toán trở thành tìm hàm xấp xỉ $\ln ae^{bx} = \ln a + bx$ theo dữ liệu

x	1.2	2.8	4.3	5.4	6.8	7.9
$z = \ln y$	2.015	2.779	3.661	4.205	4.988	5.584

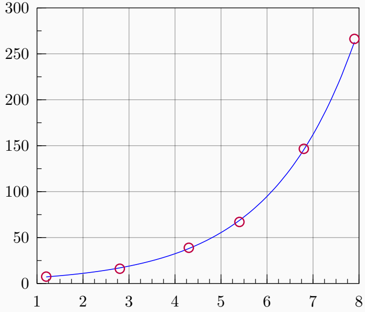
Dữ liệu có trọng số

Ta tìm đường hồi quy tuyến tính với tham số cần tìm là $A = \ln a$ và b.

$$\bar{x} = \frac{1}{6}\sum x_i = 4.733 \qquad \bar{z} = \frac{1}{6}\sum z_i = 3.872$$

$$b = \frac{\sum z_i(x_i - \bar{x})}{\sum x_i(x_i - \bar{x})} = \frac{16.716}{31.153} = 0.5366 \qquad A = \bar{z} - \bar{x}b = 1.3323.$$

Do đó, $a = e^A = 3.790$ và hàm xấp xỉ trở thành $f(x) = 3.790e^{0.5366x}$. Hàm $f(x)$ và các điểm dữ liệu được vẽ trong hình sau.



Tiếp theo ta tính độ lệch chuẩn:

y	7.5	16.1	38.9	67.0	146.6	266.2
$f(x)$	7.21	17.02	38.07	68.69	145.60	262.72
$y - f(x)$	0.29	−0.92	0.83	−1.69	1.00	3.48

$$S = \sum [y_i - f(x_i)]^2 = 17.59$$
$$\sigma = \sqrt{\frac{S}{n-m}} = 2.10.$$

Như đã trình bày phía trên, thay vì xấp xỉ y_i, chúng ta xấp xỉ $\ln y_i$ để được lời giải xấp xỉ của bài toán đã nêu.

Trường hợp (2). Chúng ta tiếp tục xấp xỉ hàm $\ln ae^{bx} = \ln a + bx$ nhưng các trọng số $W_i = y_i$ được sử dụng. Từ (23), các trung bình trọng số được tính (nhắc lại là ta xấp xỉ $z = \ln y$)

$$\hat{x} = \frac{\sum y_i^2 x_i}{\sum y_i^2} = \frac{737.5 \times 10^3}{98.67 \times 10^3} = 7.474,$$
$$\hat{z} = \frac{\sum y_i^2 z_i}{\sum y_i^2} = \frac{528.2 \times 10^3}{98.67 \times 10^3} = 5.353$$

và từ (25) ta có tham số

$$b = \frac{\sum_{i=1}^{n} y_i^2 z_i(x_i - \hat{x})}{\sum_{i=1}^{n} y_i^2 x_i(x_i - \hat{x})} == \frac{35.39 \times 10^3}{65.05 \times 10^3} = 0.5440,$$
$$\ln a = \hat{z} - b\hat{x} = 5.353 - 0.5440(7.474) = 1.287.$$

Vì

$$a = e^{\ln a} = e^{1.287} = 3.622$$

nên hàm xấp xỉ cần tìm là $f(x) = 3.622e^{0.5440}$. Ta thấy kết quả này trùng với kết quả thu được phần (1).

Dữ liệu có trọng số

Các phần dư và độ lệch chuẩn được tính như sau:

y	7.50	16.10	38.90	67.00	146.60	266.20
$f(x)$	6.96	16.61	37.56	68.33	146.33	266.20
$y-f(x)$	0.54	-0.51	1.34	-1.33	0.267	0.00

$$S = \sum [y_i - f(x_i)]^2 = 41.68$$
$$\sigma = \sqrt{\frac{S}{n-m}} = 1.023.$$

Ta quan sát thất rằng các phần dư và độ lệch chuẩn này nhỏ hơn phần (1) nên xấp xỉ tốt hơn. Ta thấy rằng nếu xấp xỉ trực tiếp y_i (xấp xỉ liên quan đến tìm nghiệm hàm siêu việt) ta sẽ thu được kết quả là $f(x) = 3.164e^{0.5442x}$ với độ lệch chuẩn là $\sigma = 1.022$, nhỏ hơn một chút so với phần (2).

THANK YOU FOR YOUR ATENTTION