Curve fitting

Lecture-2

Objective:

The purpose of curve fitting is to find the parameters values of the model function that closely match the data.

Methodologies:

The method of least squares may be one of the most systematic procedures to fit a curve through given data points.

Curve Fitting by Least Squares Method

Consider the problem of fitting a set of *n* data points

$$(x_r, y_r), \quad r = 1.2.3, \cdots, n$$

to a curve $Y = f(x)$ whose values depends on *m* parameters $c_1, c_2, c_3, \cdots, c_m$. The values of the function at a point depends on the values of the parameter involved. In least square method we determine a set of values of the parameter $c_1, c_2, c_3, \cdots, c_m$ such that the sum of the squares of the error

$$E(c_1, c_2, \cdots, c_m) = \sum_{ii=1}^{n} \left[f(x_i, c_1, c_2, \cdots, c_m) - y_i \right]^2$$

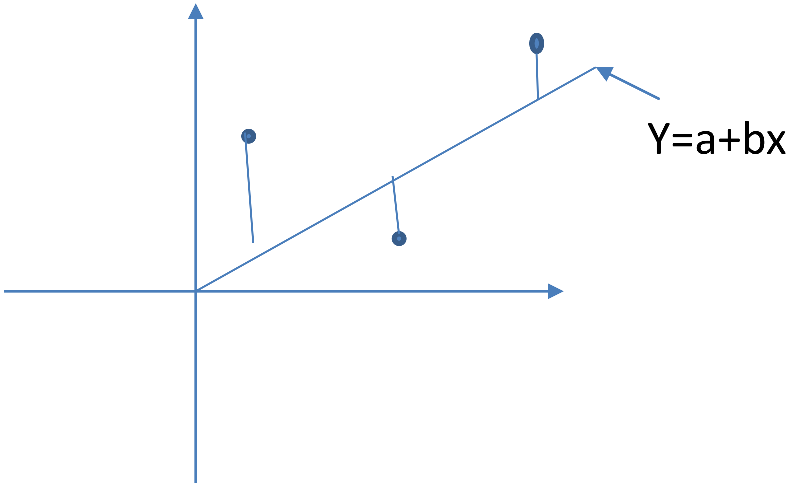
is minimum.

The necessary conditions for *E* to have a minimum is that

$$\frac{\partial E}{\partial c_r} = 0, \qquad r = 1,2,3, \cdots, m$$

This condition gives a system of *m* equations, ***called normal equations***, in *m* unknowns $c_1, c_2, c_3, \cdots, c_m$.

If the parameters appear in the function in non-linear form, the normal equations become non-linear and are difficult to solve. This difficulty may be avoided if $f(x)$ is transformed to a form which is linear in parameters.



Let the curve to be fitted $y = a + bx$

$$E(a, b) = \sum (a + bx - y)^2$$

To minimize error (i.e. E(a, b)), we have $\frac{\partial E}{\partial a} = 0 \ and \ \frac{\partial E}{\partial b} = 0$

Example: Given the following set of values of *x* and *y*:

X	1	2	3	4	5	6
Y	1.553	1.638	0.685	–0.428	–0.679	0.164

A physicist wants to approximate the data using a periodic curve $y = a + b\sin x$

Estimate the parameters *a* and *b* to 2 decimal places using least squares method.

Solution:

$$E(a,b) = \sum_{i=1}^{6} (a + b\sin x_i - y_i)^2$$

Normal equations

$$\frac{\partial E}{\partial a} = 0 \ \ and \ \ \frac{\partial E}{\partial b} = 0$$

$$\rightarrow \sum_{i=1}^{6} 2(a + b\sin x_i - y_i)1 = 0$$

$$\rightarrow \sum_{i=1}^{6} 2(a + b\sin x_i - y_i)\sin x_i = 0$$

$$\rightarrow a\sum 1 + b\sum \sin x_i = \sum y_i$$

$$\rightarrow \ a\sum \sin x_i + b\sum \sin^2 x_i = \sum y_i \sin x_i$$

$$\rightarrow \ 6\,a + b\sum \sin x_i = \sum y_i$$

The sum can be calculated as follows

X	Y	$\sin x$	$\sin^2 x$	$y \sin x$
1	1.553	0.8415	0.7081	1.3068
2	1.638	0.9093	0.8268	1.4894
3	0.685	0.1411	0.0199	0.0967
4	-0.428	-0.7568	0.5727	0.3239
5	-0.679	-0.9589	0.9195	0.6511
6	0.164	-0.2794	0.0781	-0.0458
Sum	2.933	-0.1032	3.1251	3.8221

The normal equations are

$$6a - 0.1032\, b = 2.933$$

$$-0.1032\, a + 3.1251\, b = 3.8221$$

$$6\, a + b \sum \sin x_i = \sum y_i$$

$$a \sum \sin x_i + b \sum \sin^2 x_i = \sum y_i \sin x_i$$

The normal equations are

$$6a - 0.1032\,b = 2.933$$

$$-0.1032\,a + 3.1251\,b = 3.8221$$

By dividing each equation by the coefficient of *a*, we have

$$a - 0.0172\,b = 0.4888$$

$$a - 30.282\,b = -37.0359$$

Subtracting the equations

$$30.2648\,b = 37.5247$$

Solving we have

$$b = 1.2399 \approx 1.24$$

$$a = 0.5108 \approx 0.51$$

Now the fitted curve is $y = 0.51 + 1.24\sin x$

Example

The height of a child is measured at different ages and listed below:

t (yrs)	3	6	9	12	15
H (ft)	2.87	3.60	4.28	4.88	5.35

It is believed that height follows saturation growth model

$$H = \frac{6.45}{1 + a_2 \exp(-a_3 t)}$$

i. Use a suitable substitution to reduce the above relation to a linearized form in parameters.

ii. Use least square method to find the normal equation of the above data

iii. Estimate, to 2 decimal places, the values of a_2 and a_3

iv. Estimate the height when the child becomes 20 years old.

v. Use MATLAB function $\mathrm{a} = \mathrm{lsqcurvefit(fun, a_0, xdata, ydata)}$ to fit the general form like $H = \frac{a_1}{1+a_2 \exp(-a_3 t)}$.

Solution:

$$H = \frac{6.45}{1 + a_2 \exp(-a_3 t)}$$

$$or, \frac{6.45}{H} = 1 + a_2 e^{-a_3 t}$$

$$or, \frac{6.45}{H} - 1 = a_2 e^{-a_3 t}$$

$$\ln\left(\frac{6.45}{H} - 1\right) = \ln a_2 - a_3 t$$

This can be written as

$$Y = A + BX$$

Where,

$$Y = \ln\left(\frac{6.45}{H} - 1\right), A = \ln a_2, B = -a_3, X = t$$

(ii) Now,

$$E(A, B) = \sum_{i=1}^{5} (A + BX_i - Y_i)^2$$

Normal equations are,

$$\frac{\partial E}{\partial A} = 0 \quad and \quad \frac{\partial E}{\partial B} = 0$$

$$\sum_i 2(A + BX_i - Y_i)1 = 0$$

$$\sum_i 2(A + BX_i - Y_i)X_i = 0$$

Now,

$$A \sum 1 + B \sum X_i = Y_i$$

$$A \sum X_i + B \sum X_i^2 = \sum X_i Y_i$$

$$Y = \ln\left(\frac{6.45}{H} - 1\right), A = \ln a_2\,, B = -a_3, X = t$$

The sum can be calculated in a tabular form as shown below:

N	T	H	X	Y	XY	X^2
1	3	2.87	3	0.221	0.663	9
2	6	3.60	6	-0.234	-1.402	36
3	9	4.28	9	-0.679	-6.113	81
4	12	4.88	12	-1.134	-13.609	144
5	15	5.35	15	-1.582	-23.727	225
Sum			45	-3.408	-44.187	495

Normal Equations

5	A	+	45	B	=	-3.408
45	A	+	495	B	=	-44.187

iii.

Solutions:	A	+	9	B	=	-0.682
	A	+	11	B	=	-0.982
			-2	B	=	0.3

B	=	-0.150	a_3 =	0.15
A	=	0.668	a_2 =	1.95

iv. The fitting curve is $H = \frac{6.45}{1+1.45\exp(-0.15t)}$.

From the equation of the curve, we get

when $t = 20$ then $H = 5.88.$

MATLAB Code

v. >> xd=[3 6 9 12 15]; % state x-values

>> yd=[2.87 3.60 4.28 4.88 5.35]; % staet y-values

Define fitting curve in terms of parameters as vector a

>> Fd=@(a,xd) a(1)./(1+a(2).*exp(-a(3).*xd));

>> a0=[6,2,0.2]; % guess parameter values

% To fit the curve use MATLAB function **lsqcurvefit** with following syntax

>> a=lsqcurvefit(Fd,a0,xd,yd)

Advantages and Drawbacks: Least Square Method

Advantages of Least Square Method:

- Simplicity: It is very easy to explain and to understand.
- Applicability: There are hardly any applications where least squares doesn’t make sense.
- Theoretical Underpinning: It is the maximum-likelihood solution and, if the Gauss-Markov conditions apply, the best linear unbiased estimator.

Drawbacks of Least Square Method:

- Sensitivity to outliers.
- Test statistics might be unreliable when the data is not normally distributed (but with many datapoints that problem gets mitigated).
- Tendency to overfit data.

Sample MCQ

1. What is the purpose of curve fitting?

(a) Find the parameter value

(b) Find the solution

(c) Find the data

(d) Find the parameter which closely match the data

2. To find the parameter value which method we use in curve fitting?

(a) Interpolation

(b) Least square method

(c) Lagrange method

(d) None

3. $\sum_{i=1}^{20} 1 = ?$

a) 10

b) 20

c) 30

d) 40

Exercise

1. Average price, *P*, of a certain type of second-hand car is believed to be related to its age, *t* years, by an equation of the form

$$P = \frac{50}{a + be^{\frac{t}{4}}}$$

Where *a* and *b* are constants. Data from a recent newspaper give the following average price (in Taka) for used car of this type,

t (yrs)	2	4	6	8
P (lac)	20.50	17.25	14.50	11.75

(i) Estimate the values of *a* and *b* rounded to 3 significant figures.
(ii) Estimate the values of a car of this type that is 10 years old and the original new price.

2. A bowl of hot water is kept in a room of constant temperature 25^0C. At 5 minutes interval temperature of the water is recorded and listed as given below.

t in minute	5	10	15	20	25
T in ^{0}C	76.8	70.4	64.2	58.8	54.1

The law of cooling can be assumed to be of the form $T = 27 + ae^{-kt}$.

(i) Find, to 2 significant figures, the best values of *a* and *k*.

(ii) Estimate the initial temperature.

(iii) Estimate the time, to the nearest minute, when the temperature of the water in the bowl will be 50^0C.

3. The equation $v = 70 - ce^{-kt}$ can be used for calculating the speed of a moving car, where c and k are constant.

t	4	8	12	16	20
v	23.21	28.52	33.07	36.96	40.29

(a) Estimate the values of *c* and *k* rounded to 2 significant figures.

(b) Find the time, to the nearest second, when the speed is 45 ms^{-1}.