CURVE FITTING

Md. Mehedi Hasan

Lecturer (Mathematics)

Department of Natural Sciences

Daffodil International University

Curve fitting:-

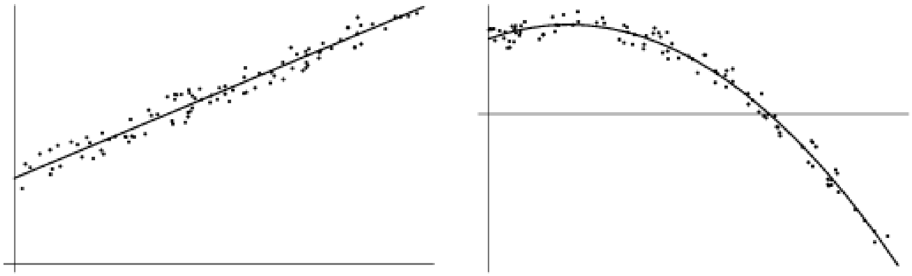
Curve fitting is the process of constructing a curve (named as an approximating curve) or mathematical function that has the best fit to a given set of data points, possibly subject to constraints. In this case the curve drawn is such that the discrepancy between the data points and the curve is least. The method of least squares is most commonly used in fitting curve.

Or

Curve Fitting is most often used by scientists and engineers to visualize and plot the curve that best describes the shape & behavior of their data. Curve fitting is the procedure in finding a curve which matches a series of data points and possibly other constraints.

Or

A procedure in which the basic problem is to pass a curve through a set of points, representing experimental data, in such a way that the curve shows as well as possible the relationship between the two quantities plotted. It is always possible to pass some smooth curve through all the points plotted, but since there is assumed to be some experimental error present, such a procedure would ordinarily not be desirable.



The method curve fitting was suggested early in the 19th century by the French mathematician Adrien Legendre.

Least Squares Method:

The least squares method is the most systematic procedure to fit a unique curve through the given data points and its widely used universally in practical computations. The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (*least square error*) from a given set of data.

Suppose that the data points are $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots\ldots\ldots\ldots\ldots, (x_n, y_n)$ where x is the independent variable and y is the dependent variable. The fitting curve $y = f(x)$ has the deviation (error/residual) d from each data point, i.e., $d_1 = y_1 - f(x_1), d_2 = y_2 - f(x_2), \ldots\ldots\ldots\ldots\ldots\ldots\ldots, d_n = y_n - f(x_n)$. It is clear that some of the residuals will be positive and the remaining will be negative. Hence to give equal importance to positive and negative residuals we square each of them and form the sum of squares.

Now the sum of the squares of the errors or deviations is,

$$S = d_1^{\ 2} + d_2^{\ 2} + \cdots\cdots\cdots\cdots\cdots\cdots\cdots + d_n^{\ 2}$$

$$S = \{y_1 - f(x_1)\}^2 + \{y_2 - f(x_2)\}^2 + \cdots\cdots\cdots\cdots\cdots\cdots + \{y_n - f(x_n)\}^2$$

$$S = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 \qquad \textbf{[Must be Minimum for Best Fitting]}$$

The quantity S provides a measure of the goodness of fit of the curve to the given data if it is very minimum and if it is large then the curve fitting is bad. For $S = 0$ each of the given points lies on $y = f(x)$ and it will decrease in value depending on the closeness of the points to the curves.

Therefore, the best representative curve to the given data set of points is that for which the sum of squares of the errors S is minimum. This is known as the least Square method /Criterion or the principle of least squares.

Note:

Least squares curves fitting are of two types such as linear and nonlinear least squares fitting to given data $(x_i, y_i), i = 1,2,\ldots\ldots,n$ according to the choice of approximating curves $f(x)$ as linear or nonlinear. The constant occurring in the equation $y = f(x)$ of the approximating curve can be found by several methods mentioned in the followings:

1.Graphical Method

2.The method of group average

3.Method of least squares

Linear curve fitting

Fitting a straight line:

Fitting a straight line means finding the values of the parameters a and b of the straight line $y = ax + b$ as well as actually constructing the line itself. The graphical method and least square method are two useful methods for finding a straight line.

Let us consider n data points $(x_i, y_i), i = 1,2,\ldots\ldots,n$ and a linear function $y = ax + b$ in x and y that represents a straight line best fit to the given data. We have to find the constants a and b. For any x_i the expected value of y (Value calculated from the equations) is $a + bx_i$ and observed value of y is y_i.

Therefore, the deviation/error/residual $d_i = y_i - (ax_i + b)$,by giving values $i = 1,2,\ldots\ldots,n$ we get the various residuals.

Now the sum of the squares of the errors or deviations is,

$$S = d_1^{\ 2} + d_2^{\ 2} + \cdots\cdots\cdots\cdots\cdots\cdots\cdots + d_n^{\ 2}$$

$$S = \{y_1 - (ax_1 + b)\}^2 + \{y_2 - (ax_2 + b)\}^2 + \cdots\cdots\cdots\cdots\cdots\cdots + \{y_n - (ax_n + b)\}^2$$

$$S = \sum_{i=1}^{n} \{y_i - (ax_i + b)\}^2 \qquad \textbf{[Must be Minimum for Best Fitting]}$$

The quantity S provides a measure of the goodness of fit of the curve to the given data if it is very minimum.

For S to be minimum the conditions are $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$.

Partially differentiating S with respect to a and b, we get

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2\{y_i - (ax_i + b)\}(-x_i)$$

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} (-2x_i)\{y_i - (ax_i + b)\}$$

And

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2\{y_i - (ax_i + b)\}(-1)$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} (-2)\{y_i - (ax_i + b)\}$$

For satisfying the conditions above equation equating with zero, we find

$$\sum_{i=1}^{n} (-2x_i)\{y_i - (ax_i + b)\} = 0$$

$$-2\sum_{i=1}^{n} x_i\{y_i-(ax_i+b)\}=0$$

$$\sum_{i=1}^{n} x_i\{y_i-(ax_i+b)\}=0$$

$$\sum_{i=1}^{n}\{x_i y_i-(ax_i{}^2+bx_i)\}=0$$

$$\sum_{i=1}^{n} x_i y_i-\sum_{i=1}^{n}(ax_i{}^2+bx_i)=0$$

$$a\sum_{i=1}^{n} x_i{}^2+b\sum_{i=1}^{n} x_i=\sum_{i=1}^{n} x_i y_i \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(i)$$

And

$$\sum_{i=1}^{n}(-2)\{y_i-(ax_i+b)\}=0$$

$$(-2)\sum_{i=1}^{n}\{y_i-(ax_i+b)\}=0$$

$$\sum_{i=1}^{n}\{y_i-(ax_i+b)\}=0$$

$$\sum_{i=1}^{n} y_i-\sum_{i=1}^{n}(ax_i+b)=0$$

$$\sum_{i=1}^{n}(ax_i+b)=\sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} ax_i+\sum_{i=1}^{n} b=\sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i+bn=\sum_{i=1}^{n} y_i \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(ii)$$

We represent the equations (i) and (ii) in matrix form

$$\begin{bmatrix}\sum_{i=1}^{n} x_i{}^2 & \sum_{i=1}^{n} x_i\\ \sum_{i=1}^{n} x_i & n\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n} x_i y_i\\ \sum_{i=1}^{n} y_i\end{bmatrix}$$

$$\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}\sum_{i=1}^{n} x_i{}^2 & \sum_{i=1}^{n} x_i\\ \sum_{i=1}^{n} x_i & n\end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^{n} x_i y_i\\ \sum_{i=1}^{n} y_i\end{bmatrix}$$

$$\begin{bmatrix}a\\ b\end{bmatrix}=\frac{1}{n\sum_{i=1}^{n} x_i{}^2-\sum_{i=1}^{n} x_i\sum_{i=1}^{n} x_i}\begin{bmatrix}n & -\sum_{i=1}^{n} x_i\\ -\sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i{}^2\end{bmatrix}\begin{bmatrix}\sum_{i=1}^{n} x_i y_i\\ \sum_{i=1}^{n} y_i\end{bmatrix} \qquad \because \begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}, ad-bc\neq 0$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i} \\ \dfrac{\sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i} \end{bmatrix}$$

Now equating the two equal matrices we get,

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - n\bar{x}\, n\bar{y}}{n\sum_{i=1}^{n} x_i^{\,2} - \left(n\bar{x}\right)^2} \qquad \left[\because \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}\right]$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - n^2 \bar{x}\,\bar{y}}{n\sum_{i=1}^{n} x_i^{\,2} - n^2 \bar{x}^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\,\bar{x}^2}$$

And

$$b = \frac{\sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}$$

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

$$= \frac{\bar{y} \sum_{i=1}^{n} x_i^{\,2} - \bar{x}\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^{\,2} - n\,\bar{x}^2}$$

Putting this values of a and b in the equation $y = ax + b$ we get the equation of the line best fitting the data as $y = ax + b$.

Note:

The equations $a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$ and $a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i$ are called the normal equations.

Dropping of the suffices above equation can be written as $a\sum x^2 + b\sum x = \sum xy$ and $a\sum x + bn = \sum y$

Mathematical problem on Linear curve fitting

Problem 01: Use the method of least squares to fit a straight line to the following data:

x	*0*	*5*	*10*	*15*	*20*
y	*7*	*11*	*16*	*20*	*26*

Estimate the value of y when x=25.

Solution:

Assume that the least square straight line to be fitted to the given data be $y = ax + b$.

Then we have the normal equations are

$$a\sum x^2 + b\sum x = \sum xy \;\cdots\cdots\cdots\cdots\cdots\cdots(i)$$

and

$$a\sum x + bn = \sum y \;\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(ii)$$

Here the number of data points n= 5.

Calculation for finding the coefficients a and b of the least square line.

x	y	xy	x^2
0	7	0	0
5	11	55	25
10	16	160	100
15	20	300	225
20	26	520	400
$\sum$ x = 50	$\sum$ y = 80	$\sum$ xy =1035	$\sum$ x^2=750

Now putting these values in the above equations (i) and (ii) we get

$750a + 50b = 1035$ and $50a + 5b = 80$

Solving above two equations by calculator, we get values of $a = 0.94$ and $b = 6.6$.

Putting these values in the equation $y = ax + b$ we get the required line as $y = 0.94x + 6.6$.

Expected value of $y = 0.94 \times 25 + 6.6 = 30.1$ as $x = 25$. **(As desired)**

Problem 02: Find the least square line $y = ax + b$ for the data points $(-1,10)$, $(0,9)$, $(1,7)$, $(2,5)$, $(3,4)$, $(4,3)$ $(5,0)$ and $(6,-1)$.

Solution:

Given that least square straight is $y = ax + b$ and number of data points n= 8.

Then we have the normal equations are

$$a\sum x^2 + b\sum x = \sum xy \;\cdots\cdots\cdots\cdots\cdots\cdots(i)$$

and

$$a\sum x + bn = \sum y \;\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(ii)$$

Calculation for finding the coefficients a and b of the least square line.

x	y	xy	x^2
-1	10	-10	1
0	9	0	0
1	7	7	1
2	5	10	4
3	4	12	9
4	3	12	16
5	0	0	25
6	-1	-6	36
$\sum x = 20$	$\sum y = 37$	$\sum xy = 25$	$\sum x^2 = 92$

Now putting these values in the above equations (i) and (ii) we get

$92a + 20b = 25$ and $20a + 8b = 37$

Solving above two equations by calculator, we get values of $a = -1.60714$ and $b = 8.64286$.

Putting these values in the equation $y = ax + b$ we get the required line as $y = -1.60714x + 8.64286$.

(As desired)

Another Method of finding Equation of Line:

Use the following steps to find the equation of line of best fit for a set of ordered pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n).

Step 1: Calculate the mean of the *x*-values and the mean of the *y*-values.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Step 2: The following formula gives the slope of the line of best fit:

$$m = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

Step 3: Compute the *y*-intercept of the line by using the formula:

$$b = \overline{Y} - m\overline{X}$$

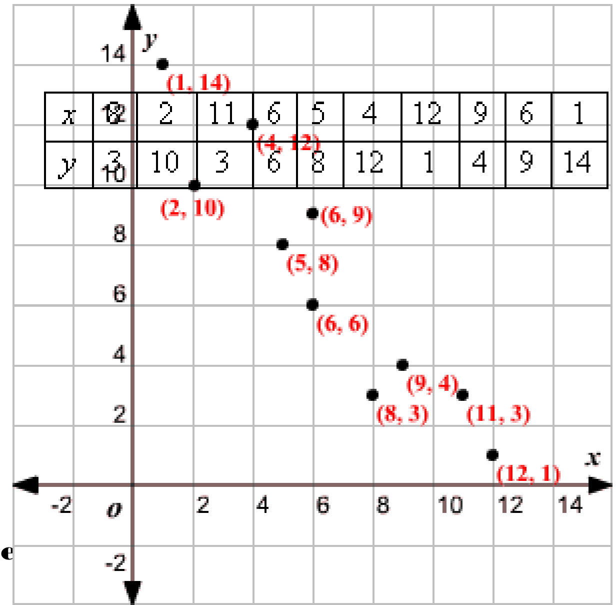
Step 4: Use the slope *m* and the *y*-intercept *b* to form the equation of the line.

Problem 03: Use the least square method to determine the equation of line of best fit for the data. Then plot the line.

x	8	2	11	6	5	4	12	9	6	1
y	3	10	3	6	8	12	1	4	9	14

Solution:

Plot the points on a coordinate plane.



Calculate the means of the *x*-values and the *y*-values.

$$\overline{X} = \frac{8+2+11+6+5+4+12+9+6+1}{10} = 6.4$$

$$\overline{Y} = \frac{3+10+3+6+8+12+1+4+9+14}{10} = 7$$

Now calculate $x_i - \overline{X}$, $y_i - \overline{Y}$, $(x_i - \overline{X})(y_i - \overline{Y})$ and $(x_i - \overline{X})^2$ for each *i*.

i	x_i	y_i	$x_i - \overline{X}$	$y_i - \overline{Y}$	$\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$	$\left(x_i - \overline{X}\right)^2$
1	8	3	1.6	−4	−6.4	2.56
2	2	10	−4.4	3	−13.2	19.36
3	11	3	4.6	−4	−18.4	21.16
4	6	6	−0.4	−1	0.4	0.16
5	5	8	−1.4	1	−1.4	1.96
6	4	12	−2.4	5	−12	5.76
7	12	1	5.6	−6	−33.6	31.36
8	9	4	2.6	−3	−7.8	6.76
9	6	9	−0.4	2	−0.8	0.16
10	1	14	−5.4	7	−37.8	29.16
					$\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) = -131$	$\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 = 118.4$

Calculate the slope.

$$m = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} = \frac{-131}{118.4} \approx -1.1$$

Calculate the *y*-intercept.

Use the formula to compute the *y*-intercept.

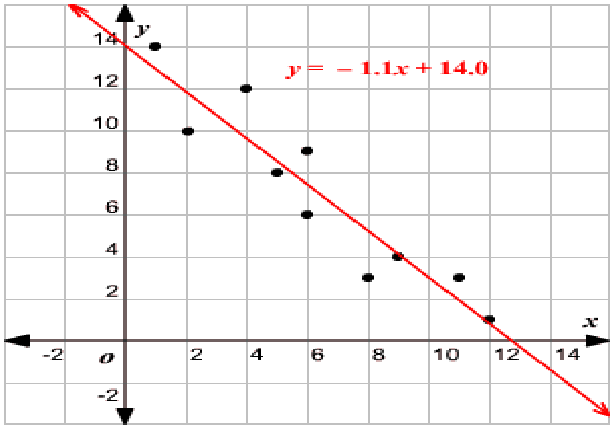
$$\begin{aligned} b &= \overline{Y} - m\overline{X} \\ &= 7 - (-1.1 \cdot 6.4) \\ &= 7 + 7.04 \\ &\approx 14.0 \end{aligned}$$

Use the slope and *y*-intercept to form the equation of the line of best fit.

The slope of the line is –1.1 and the *y* -intercept is 14.0.

Therefore, the equation is *y* = –1.1 *x* + 14.0.

Draw the line on the scatter plot.



Problems for practicing:

1.Find the least square line $y = ax + b$ for the data

x	-2	-1	0	1	2
y	1	2	3	3	4

2. Find the values of a_0 and a_1 so that $y = a_0 + a_1 x$ fits the data given in the table:

x	0	1	2	3	4
y	1	2.9	4.8	6.7	8.6

3.Fit a straight line of the form $y = a_0 + a_1 x$ to the data:

x	1	2	3	4	6	8
y	2.4	3.1	3.5	4.2	5	6

4.The table below gives the temperature T (in 0^0C) and length l (in mm) of a heated rod. If $l = a_0 + a_1 T$ find the values of a_0 and a_1 using linear least squares

T	40	50	60	70	80
l	600.5	600.6	600.8	600.9	601

5.Find the least square line $y = ax + b$ for the data

x	-4	-2	0	2	4
y	1.2	2.8	6.2	7.8	13.2

6.Fit a straight line to the following data regarding x as the independent variable

x	0	1	2	3	4

y	1	1.8	3.3	4.5	6.3

7.Find the least square fit straight line of the form $y = ax + b$ for the data of fertilize application and yield of a plant

fertilizer	0	10	20	30	40	50
Yield (kg)	.8	.8	1.3	1.6	1.7	1.8