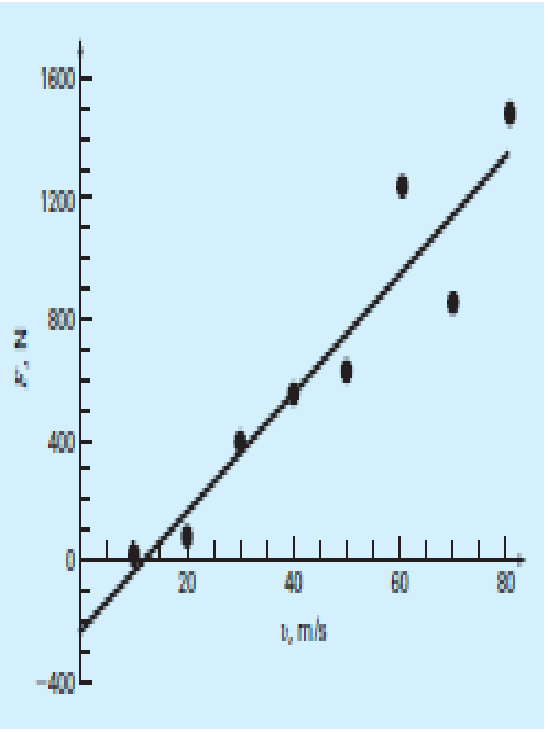
LINEAR LEAST-SQUARES REGRESSION

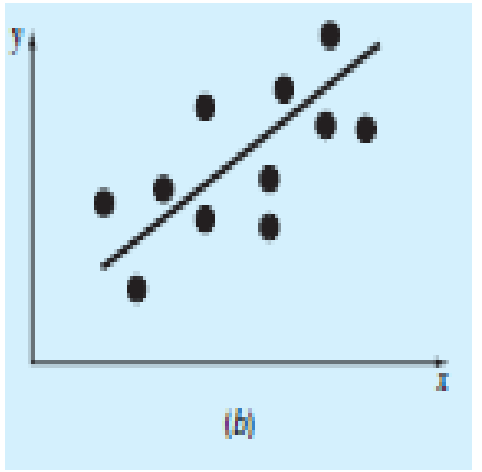
Where substantial error is associated with data, the best curve-fitting strategy is to Derive an approximating function that fits the shape or general trend of the data without necessarily matching the individual points. One approach to do this is to visually inspect the plotted data and then sketch a "best" line through the points.

Although such "eyeball" approaches have commonsense appeal and are valid for "back-of-the-envelope" calculations, they are deficient because they are arbitrary. That is, unless the points define a perfect straight line (in which case, interpolation would be appropriate), different analysts would draw different lines. To remove this subjectivity, some criterion must be devised to establish a basis for the fit. One way to do this is to derive a curve that minimizes the discrepancy between the data points and the curve. To do this, we must first quantify the discrepancy. The simplest example is fitting a straight line to a set of paired observations:

$(x_1, y_1), (x_2, y_2), , \ldots, (x_n, y_n)$.

The mathematical expression for the straight line is





$$y = ax + b + e \qquad (1)$$

where *a* and *b* are coefficients representing the intercept and the slope, respectively, and *e* is the error, or *residual,* between the model and the observations, which can be represented by rearranging Eq. (1) as

$$e = y - ax - b$$

Thus, the residual is the discrepancy between the true value of *y* and the approximate value, $ax + b$, predicted by the linear equation.

A strategy that overcomes the shortcomings of the aforementioned approaches is to minimize the sum of the squares of the residuals:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y - ax - b)^2 \qquad (2)$$

This criterion, which is called *least squares,* has a number of advantages, including that it yields a unique line for a given set of data. Before discussing these properties, we will present a technique for determining the values of a and b that minimize Eq. (2).

Least-Squares Fit of a Straight Line

To determine values for a and b, Eq. (2) is differentiated with respect to each unknown coefficient:

$$\frac{\delta S_r}{\delta a} = -2\sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0$$

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

Now, realizing that $\sum_{i=1}^{n} b = n\,b$, we can express the equations as a set of two simultaneous linear equations with two unknowns (a and b):

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb \qquad (3)$$

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

Example: Fit a straight line to the values in the following Table

x	10	20	30	40	50	60	70	80
y	25	70	380	550	610	1220	830	1450

Solution:

x	y	x^2	xy
10	25	100	250
20	70	400	1400
30	380	900	11400
40	550	1600	22000
50	610	2500	30500
60	1220	3600	73200
70	830	4900	58100
80	1450	6400	116000
Sum=360	Sum=5135	Sum= 20400	Sum= 312850

Substituting from the value of the above table in equations (3) we have

$5135 = 360\,a + 8\,b$

$312850 = 20400\,a + 360\,b$

Then $a = 19.4702$ & $b = -234.2857$

Hence $y = 19.4702\,x - 234.2857$

Application of linear regression

Linearization of nonlinear relation

1]The exponential model y=be^{ax}

Ln y = ln b+ a x, let Y=ln y and B=ln b then we have the following linear equation Y=ax + B which is similar Eq.(3)

The two equations to determine a & B are

$$\sum_{i=1}^{n} Y_i = nB + a\sum_{i=1}^{n} x_i \qquad (4)$$

$$\sum_{i=1}^{n} Y_i x_i = B\sum_{i=1}^{n} x_i + a\sum_{i=1}^{n} x_i^2$$

Example: Fit an exponential model to

X	0.2	0.4	0.8	1.2	1.6	2.4
y	550	750	1000	1400	2000	3750

Solution

x	y	x^2	Y=ln y	xY
0.2	550	0.04	6.3099	1.2620
0.4	750	0.1600	6.6201	2.6480
0.8	1000	0.6400	6.9078	5.5262
1.2	1400	1.4400	7.2442	8.6931
1.6	2000	2.5600	7.6009	12.1614
2.4	3750	5.7600	8.2295	19.7508
6.6	9450	10.6	42.9124	50.0415

Substituting from the value in the above table in Eq.(4) we have

$$6.6\,a+6B=42.9124$$
$$10.6\,a+6.6B=50.0415$$

a= 0.8497 & B=6.2174 , since B=lnb, then b=$e^{6.2174}$ = 501.3979

Hence y= 501.3979 $e^{0.8497x}$

2] Power model $y=bx^{a}$

Then log y=log b + a log x . This equation can be written as follows Y=B + a X

$$\sum_{i=1}^{n} Y_i = nB + a\sum_{i=1}^{n} X_i \qquad (5)$$

$$\sum_{i=1}^{n} Y_i X_i = B\sum_{i=1}^{n} X_i + a\sum_{i=1}^{n} X_i^2$$

Example: Fit a power model to

X	1	2	3	4	5
y	0.5	1.7	3.4	5.7	8.4

x	y	X=log_{10}x	Y=log_{10}y	X^2	X Y
1	0.5	0	-0.3010	0	0
2	1.7	0.3010	0.2304	0.0906	0.0694
3	3.4	0.4771	0.5315	0.2276	0.2536
4	5.7	0.6021	0.7559	0.3625	0.4551
5	8.4	0.6990	0.9243	0.4886	0.6460
		2.0792	2.1411	1.1693	1.4241

Substituting from the value in the above table in Eq.(5) we have

$$2.1411 = 5B + 2.0792\,a$$
$$1.4241 = 2.0792B + 1.1693\,a$$

Then B=-0.3002 & a = 1.7518. Hence b=$10^{-0.3001}$=0.50107

Then y= 0.50107 $x^{1.7518}$

3] Growth rate model $y = \frac{ax}{b+x}$

Then $\frac{1}{y} = \frac{b+x}{ax} = \frac{b}{a}\,\frac{1}{x} + \frac{1}{a} \rightarrow \quad Y = A\,X + B$

Where $Y = \frac{1}{y}$, $X = \frac{1}{x}$, $A = \frac{b}{a}$, and $B = \frac{1}{a}$

Example: Fit a growth rate model for the following

X	2.5	3.5	5	6	7.5	10
y	5	3.4	2	1.6	1.2	0.8

Solution: Substituting from the value of the blow table in Eq.(5) we have

3.7025 = 6B + 1.2857 A

0.6043 = 1.2857 B + 0.3372 A

Hence A= -3.0648 B=1.2738

Thus a= $\frac{1}{B}=\frac{1}{1.2738}=0.7851$, A = $\frac{b}{a}=\frac{b}{0.7851} \rightarrow b=-3.0648\times 0.7851=-2.4062$

Then $y=\frac{0.7851\,x}{-2.4062+x}$

x	y	$X = \frac{1}{x}$	$Y = \frac{1}{y}$	X^2	X Y
2.5	5	0.4000	0.2000	0.1600	0.0800
3.5	3.4	0.2857	0.2941	0.0816	0.0840
5	2	0.2000	0.5000	0.0400	0.1000
6	1.6	0.1667	0.6250	0.0278	0.1042
7.5	1.2	0.1333	0.8333	0.0178	0.1111
10	0.8	0.1000	1.2500	0.0100	0.1250
		1.2857	3.7025	0.3372	0.6043

Some other forms which can be linearized

1] $y = a\,x^n + b \quad \rightarrow \quad X = x^n \quad \rightarrow \quad y = a\,X + b$

2] $x\,y = ax + b \quad \rightarrow \quad y = a + b\,\frac{1}{x} \quad \rightarrow \quad y = a + b\,X$

3] $y = \frac{1}{ax+b} \quad \rightarrow \quad \frac{1}{y} = a\,x + b \quad \rightarrow \quad Y = ax + b$

4] $x\,y = a\,x + b\,y \quad \rightarrow \quad 1 = \frac{a}{y} + \frac{b}{x} \quad \rightarrow \quad 1 = a\,Y + b\,X \rightarrow \quad Y = \frac{1}{a} - \frac{b}{a}\,X \rightarrow \quad Y = B + AX$

5] $y = a \log x + b \quad \rightarrow \quad y = aX + b$

Exercises:

1] Use the least square regression to fit : (a) a straight line ,(b)a power (exponential) equation, (c) growth rate equation , and (d) a parabola for the following data

1-

X	1	2	3	4	5	6
y	3.6	4.7	5.5	7.5	8.7	9.9

2-

x	1	2	2.5	4	6	8	8.5
y	0.4	0.7	0.8	1	1.2	1.3	1.4

3-

x	0.05	0.4	0.8	1.2	1.6	2	2.4
y	550	750	1000	1400	2000	2700	3750

2] Use the least square regression to fit a curve on the form $y = a + bx^2$ suitable for this data

x	0	2	4	6	8	10
y	7.76	11.8	24.4	43.6	71.2	107

3] Use the least square regression to fit a curve on the form $y = a + b/x^3$ suitable for this data

X	1	2	3	4	5	6
y	66	22	14	11	9.4	8.6

4] Use the relation x y = ax + b to find the best value for a and b to fit the following data

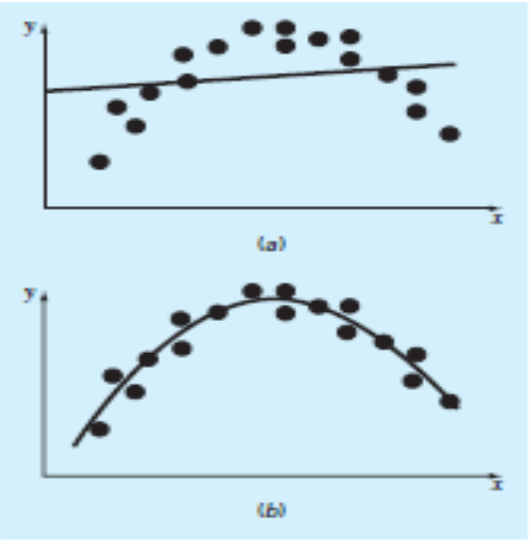
x	36.8	51.5	25.3	21	15.8	12.6
y	12.5	12.9	13.1	13.3	14.1	14.5

POLYNOMIAL REGRESSION

In the first section, a procedure was developed to derive the equation of a straight line using the least-squares criterion. Some data, although exhibiting a marked pattern such as seen in This figure, are poorly represented by a straight line. For these cases, a curve would be better suited to fit the data. As discussed in the first section, one method to accomplish this objective is to use transformations. Another alternative is to fit polynomials to the data using *polynomial regression.*

The least-squares procedure can be readily extended to fit the data to a higher-order polynomial. For example, suppose that we fit a second-order polynomial or quadratic

$$y = a_0 + a_1 x + a_2 x^2 + e$$



For this case the sum of the squares of the residuals is

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{2} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

To generate the least-squares fit, we take the derivative of the above equation with respect to each of the unknown coefficients of the polynomial, as in

$$\frac{\delta S_r}{\delta a_0} = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0 \qquad \sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2$$

$$\frac{\delta S_r}{\delta a_1} = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i - a_2 x_i^2) x_i = 0 \rightarrow \sum_{i=1}^{n} y_i x_i = a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3$$

$$\frac{\delta S_r}{\delta a_2} = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i - a_2 x_i^2) x_i^2 = 0 \qquad \sum_{i=1}^{n} y_i x_i^2 = a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4$$

Example: Fit a second-order polynomial to the data in the first two column in the following table

Solution:

x	y	x^2	x^3	x^4	X y	x^2y
0	2.1	0	0	0	0	0
1	7.7	1	1	1	7.7000	7.7
2	13.6	4	8	16	27.2000	54.4
3	27.2	9	27	81	81.6000	244.8
4	40.9	16	64	256	163.6000	654.4
5	61.1	25	125	625	305.5000	1527.5
Sum=15	152.6	55	225	979	585.6	2488.8

Substituting in the last system we have

$$6a_0 + 15a_1 + 55a_2 = 152.6$$
$$15a_0 + 55a_1 + 225a_2 = 585.6$$
$$55a_0 + 225a_1 + 979a_2 = 2488.8$$

Using Gauss elimination we have $a_0 = 2.4786$ & $a_1 = 2.3593$ & $a_2 = 1.8607$

Then the equation is

$$y = 2.4786 + 2.3593x + 1.8607x^2$$

Exercise: Fit a second-order polynomial to the data in the first two column in the following tables

(1)

x	1	2	2.5	4	6	8	8.5
y	0.4	0.7	0.8	1	1.2	1.3	1.4

(2)

X	1	2	3	4	5	6
y	66	22	14	11	9.4	8.6

(3)

x	0	2	4	6	8	10
y	7.76	11.8	24.4	43.6	71.2	107