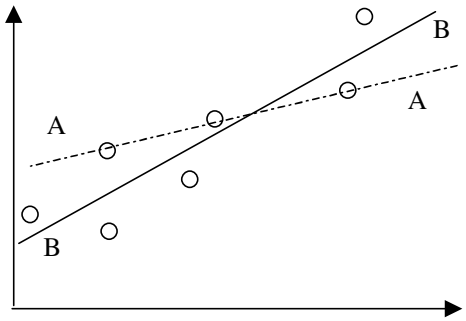
Least Squares Linear Regression

Least Squares Curve Fitting

What we are doing here is finding a simple mathematical function that describes data that shows a trend. We want to draw a smooth line (a fit) through the data and calculate the parameters of the line. If there is a theory (model) for how the trend should go, then the parameters measured are the fit to the model. Note that unlike Lagrange Interpolation, the curve is not constrained to go exactly through each data point – it simply goes "through the middle".

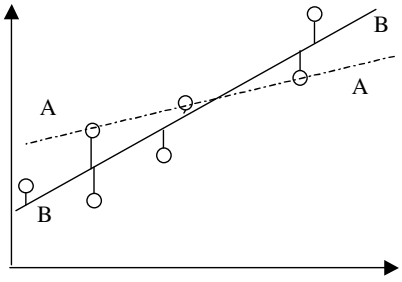
As in Lagrange Interpolation, it can be used to predict values within the range of the data that were not actually measured and, more important, predict values outside the range of the data to see where it is heading or where it came from – this is called **extrapolation**.

Look at the following data and two straight lines that are possible fits:



The line AA has the advantage that it passes closely through three data points, but line BB looks like a better fit even though it passes through none. BB looks better because it passes through the "middle" of the data. The method of least squares provides a quantitative and statistically valid method of judging when a function passes through the "middle" of data.

Before we can use least squares we need a way of measuring the spread of the data about the line. The spread is called the **scatter**. Scatter is shown below by the vertical lines, and each vertical line is called a **residual**.



Origin of the Method - Connection between Residuals, Standard Deviation and Mean value

Back in the Statistics topic we introduced the mean value as an independent parameter. In fact, there is a connection between the mean value and the standard deviation. Here's the formula we used for the population:

Standard deviation σ = $\sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$ = root mean square residual

The quantity $x_i - \mu$ is called a ***residual***. It is the difference of the i^{th} data value from the mean μ.

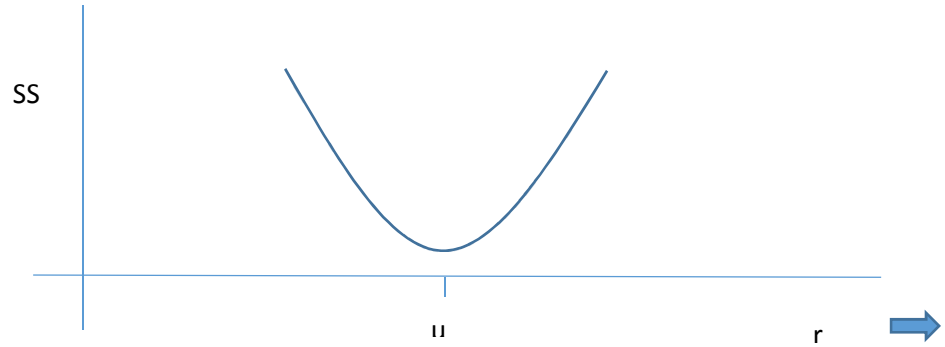
Imagine that the parameter μ in the formula for the standard deviation is an adjustable parameter, unrelated to the mean. To emphasize this let's write a formula with r not μ:

$\sigma_r = \sqrt{\frac{\Sigma(x_i-r)^2}{N}}$

The quantity we focus on is the sum of squares of residuals inside the square root:

SS = sum of squares of squares of residuals = $\sum_1^N (x_i - r)^2$ = $(x_1-r)^2 + (x_2-r)^2 + \ldots + (x_N-r)^2$

Looking at the expression for SS, because the square of the residual appears, SS is a minimum when r equals the true mean and is larger when r is greater or less than the mean:



We can prove that SS is a minimum when r = μ by using calculus. A minimum in SS occurs as r is varied when $\frac{dSS}{dr}$ = 0

when $\frac{dSS}{dr}$ = $-2(x_1 - r) - 2(x_2 - r) - 2(x_3 - r) - 2(x_4 - r) - \ldots\ldots\ldots\ldots\ldots\ldots - 2(x_N - r) = 0$

Cancelling out the -2 and moving the r's to the right-hand side gives

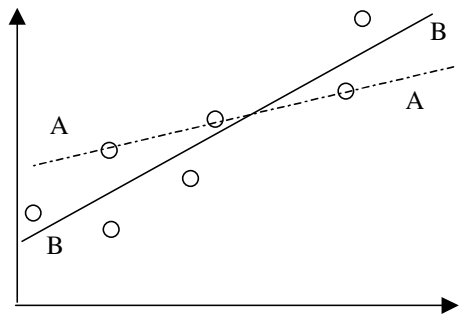
$x_1 + x_2 + x_3 + \ldots\ldots\ldots x_N = Nr$ or $r = \frac{\sum_1^N (x_i)}{N}$

which is exactly the definition of the mean value.

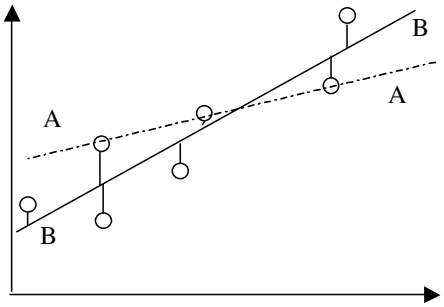
The mean is the value of the adjustable parameter that minimizes the sum of squares of residuals – least squares. This idea is the basis of the Least Squares method of line fitting.

Least Squares

Back to the data again:



The best fit straight line is one that goes through the middle of the data. Is it line A or line B? From what precedes you know it will be whatever minimizes the sum of squares of residuals. The residuals in a 2D plot of data pairs (x,y) are the vertical lines from the data to the fit:



For the data points (x_1,y_1), (x_2,y_2), ….(x_n,y_n) the residuals are $(y_1 - f(x_1))$, $(y_2 - f(x_2))$, …. , $(y_n - f(x_n))$.
The sum of squares of residuals is

$$SS = (y_1-f(x_1))^2 + (y_2-f(x_2))^2 + ….. + (y_N-f(x_N))^2$$

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

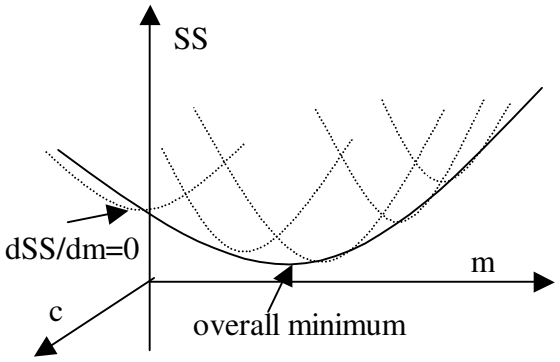
where f(x) = mx +c for a straight line, in which case it is called **linear regression**.
The function f(x) need not be a straight line but that is more complicated – see later.

Linear Regression

In this case the function is a straight line: $y = mx + c$

$$SS= (y_1-mx_1-c)^2 + (y_2-mx_2-c)^2 + ….. + (y_N-mx_N-c)^2$$

Since we can vary both the slope, m, or the intercept, c, independently to get the best fit, there are now two derivatives to put to zero:



These derivative equations are two simultaneous linear equations:

1. Minimize with respect to m:

$$\frac{\partial SS}{\partial m} = 2(y_1 - mx_1 - c)(-x_1) + 2(y_2 - mx_2 - c)(-x_2) + \ldots = \sum_{i=1}^{i=n} 2(y_i - mx_i - c)(-x_i) = 0$$

2. Minimize with respect to c

$$\frac{\partial SS}{\partial c} = 2(y_1 - mx_1 - c)(-1) + 2(y_2 - mx_2 - c)(-1) + \ldots = \sum_{i=1}^{i=n} 2(y_i - mx_i - c)(-1) = 0$$

Dividing each equation by –2 and separating the summations gives the **normal equations** :

$$m\sum x_i^2 + c\sum x_i = \sum x_i y_i$$
$$m\sum x_i + cN = \sum y_i$$

The equations look simpler with the following substitutions:

$$S = \sum_{i=1}^{N} = N\ ; \quad S_x = \sum x_i\ ; \quad S_y = \sum y_i\ ; \quad S_{xy} = \sum x_i y_i\ ; \quad S_{xx} = \sum x_i^2\ ; \quad \Delta = SS_{xx} - (S_x)^2\ ;$$

so the normal equations become

$$mS_{xx} + cS_x = S_{xy}$$
$$mS_x + cS = S_y$$

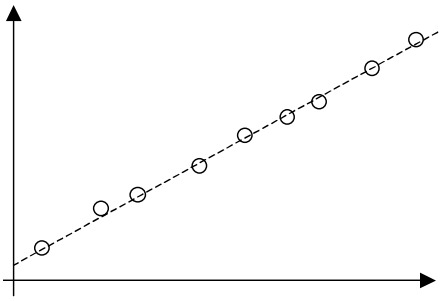
Then the solutions for m and c are:

$$m = \frac{SS_{xy} - S_x S_y}{\Delta} \qquad c = \frac{S_{xx}S_y - S_x S_{xy}}{\Delta} \qquad 1.$$

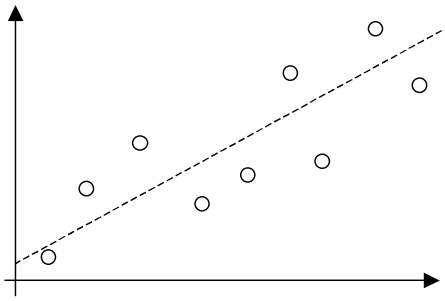
From these results, the m and c parameters for the best-fit straight line can easily be determined.

Standard Deviation and Errors

Look at the following two sets of data and the least-squares lines that fit them:

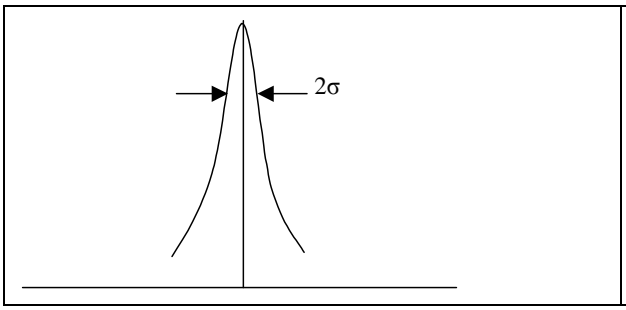


graph 1

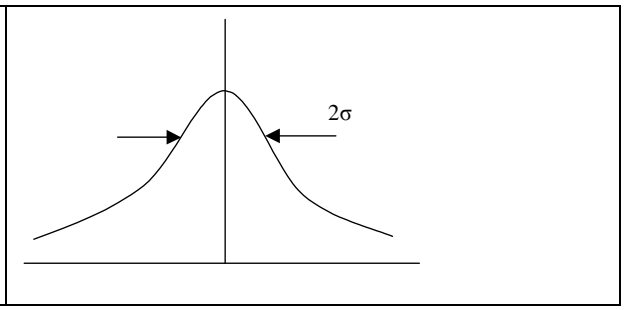


graph 2

It's clear that although the lines in graphs 1 and 2 have the same parameters (m and c), the data that produced them have greatly different residuals. If we were to add in a lot more data points and then plot a histogram of the number of residuals in a range against their value we would get **Gaussian** curves (also called Normal or Bell curves) like:



graph 1



graph 2

The width of the curves is characterized by the standard deviation σ.
σ is far worse for graph 2 that graph 1.
The standard deviation σ for the sample is calculated from:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{N-2}}$$

2.

It is the root mean sum of the squares of the residuals. The denominator N-2 (you might expect it to be N) is the number of degrees of freedom. It is 2 short because f(x) contains two coefficients (m and c) derived from the data and overall therefore has lost 2 degrees of freedom.

Worked Example 1

In a certain experiment, the resistance *R* of a certain resistor was measured as a function of the temperature *T*. The data found are shown in the following table. Find the least-squares straight line, expressing *R* as a function of *T*.

T (°c)	0.0	6.0	10.0
R(Ω)	2.5	3.1	3.5

$\sum x$ = 16, $(\sum x)^2$ = 256, $\sum y$ = 9.1, $\sum xx$ = 136, $\sum xy$ = 53.6, n = 3
m = (3*53.6 – 16*9.1)/(3*136 - 256) = 15.2/152 = 0.10
b = (136*9.1 – 53.6*16)/(3*136 - 256) = 380/152 = 2.5

The linear relation is R = 0.1*T + 2.5

Worked Example 2

The altitude h (in km) of a rocket was measured at three positions at a horizontal distance x (in km) from the launch site, shown in the table.

x (km)	0	1.0	2.0
h (km)	0	2.25	4.5

Find the least-squares straight line for h as a function of x.
n = 3; $\sum x$ = 3; $\sum y$ = 6.75; $(\sum x)(\sum y)$ = 20.25; $\sum xy$ = 11.25; $\sum x^2$ = 5; $(\sum x)^2$ = 9
m = $\frac{3(11.25)-20.25}{15-9}$ = 2.25 b = $\frac{5(6.75)-(11.25)3}{1-9} = \frac{33.75-33.75}{6}$ = 0

The linear relation is h = 2.25x

Least Squares Fitting of Nonlinear Data

Linearization of Nonlinear Data

Many equations in engineering and science describe data that do not lie on the straight line

$$y = ax + b$$

and so would appear to be unsuitable for linear regression. However it is sometimes possible to reform the equation so that it does have the linear form for which our standard linear regression results can then be used.

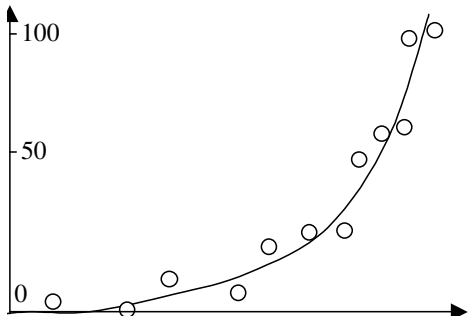
Exponentials

Exponential functions are common in science, engineering and biology. For example: radioactive decay, population growth, etc.

The following experimental data is expected to lie on an exponential function

$$y = ae^{bx} \qquad 1.$$

An exponential of the correct form has been drawn through it.



Apart from not lying on a straight line, the data looks quite normal, with scatter (residuals) independent of the value of x and therefore the same standard deviation for all the data points. That is an important point to remember. In the linear regression section we made that assumption.

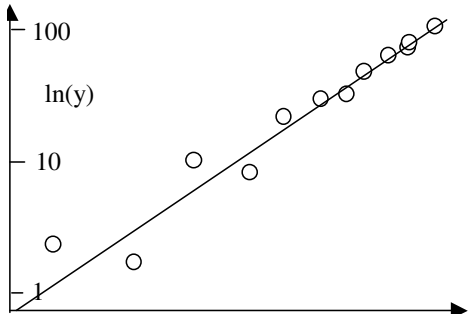
Now we will attempt to reform the exponential into a linear form so we can use our linear regression formulas. Take the natural log of both sides of 1:

$$\ln y = \ln a + bx \ln e = \ln a + bx$$

Now we have a linear result by simply relabeling: Y = ln(y); C = ln(a) giving

$$Y = bx + C$$

Plotting the data now gives a straight line but with a new ordinate scale (the Y axis):



But, you can see what has happened. The scatter in the data is now distorted, appearing to be less at high values of x. The greater scatter at the low end will make the calculated coefficients more sensitive to data at the low end. Another way to look at this is to observe that the logarithm has compresses the range of the data at high values making it contribute less to the extent of the line.

Worked Example

The following data shows approximately exponential decay. It is the temperature of a car engine after it is turned off in cold weather. Use linear regression to derive the coefficients *a* and *b* in the formula: $T = ae^{-bt}$

time t (hours)	T centigrade
0	90
.1	76
.3	63
.9	22

Solution. Take natural logs: $\ln(T) = \ln(a) - bt = c + mx$

t = x	T	ln(T) = y
0	90	4.50
.1	76	4.33
.3	63	4.14
.9	22	3.09

Now take x as t and y as ln(T) in the formulas from earlier linear least squares:

$$S = \sum_{i=1}^{N} = N\,;\quad S_x = \sum x_i\,;\qquad S_y = \sum y_i\,;\qquad S_{xy} = \sum x_i y_i\,;\quad S_{xx} = \sum x_i^2\qquad ;$$

$$\Delta = SS_{xx} - (S_x)^2\,;$$

$$m = \frac{SS_{xy} - S_x S_y}{\Delta} \qquad\qquad c = \frac{S_{xx}S_y - S_x S_{xy}}{\Delta}$$

Calculate the individual terms

t =x	x^2	ln(T) = y	xy
0	0	4.50	0
.1	.01	4.33	.433
.3	.09	4.14	1.242
.9	.81	3.09	2.871

Do the summations and use the formulas

S	Sx	Sy	Sxx	Sxy	Δ	m	c
4	1.3	16.06	.91	4.456	1.95	-1.566	4.52

$c = \ln(a)$ gives $a = \ln^{-1}(c) = 91.9$. Therefore the exponential formula is: $T = 91.9e^{-1.56t}$

Simple power law

Simple power laws are common: wireless signal dropoff with distance, gravitational force from a planet, etc. The form of the equation is

$$y = ax^b$$

This can be linearized by also taking the logs of both sides:

$$\ln(y) = \ln(a) + b\ln(x)$$

Now defining the new variables Y = ln(y), A = ln(a) and X = ln(x) we get back the linear relation:

$$Y = A + bX$$

for which we can now use the standard formulas to find A and b.
To get a from A use the result

$$a = \exp(A).$$

Saturation Growth Rate

Such equations appear in population growth rate under a limit, such as the population growth of rabbits where the food supply is limited :

$$y = a\frac{x}{x+b}$$

This can be linearized by rearranging:

$$\frac{1}{y} = \frac{b}{a}\frac{1}{x} + \frac{1}{a}$$

Now defining the new variables Y = 1/y , X = 1/x , b/a = A and 1/a = B we get back the linear relation:

$$Y = AX + B$$

Both a and b can be then calculated from:

$$a = 1/B,\ b = A/B$$