Least squares method.

Robert Mařík

March 3, 2006

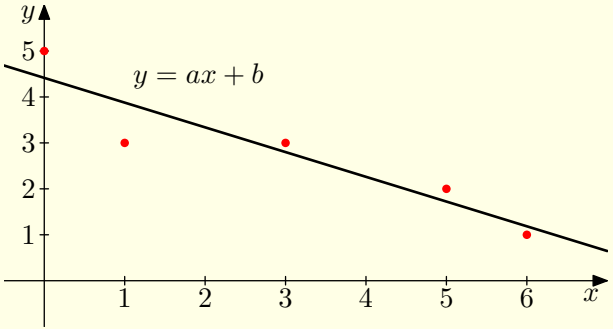
Contents

1 Motivation

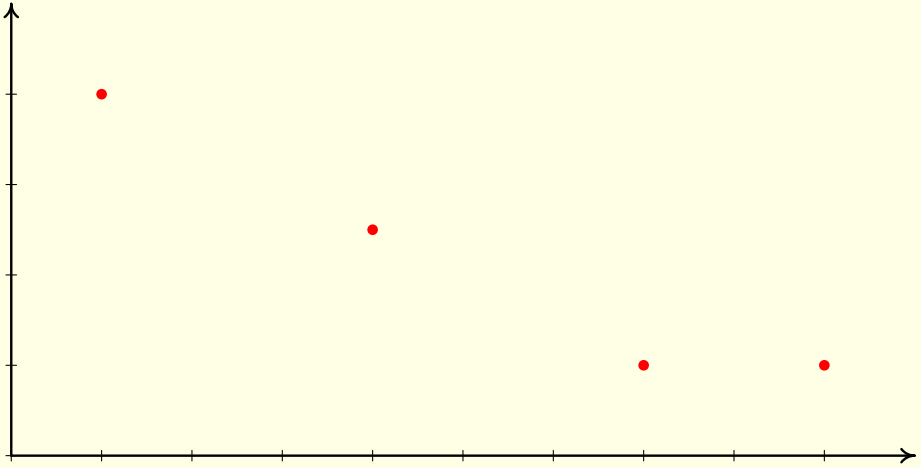
In several problems in science we are confronted with an analysis of experimental or statistical data where a linear relationship

$$y = ax + b$$

is theoretically predicted to exist between two variables (y and x).
The method of linear lest-squares analysis allows us to determine from an experimental measure of y and x what the best linear fit to the data is.



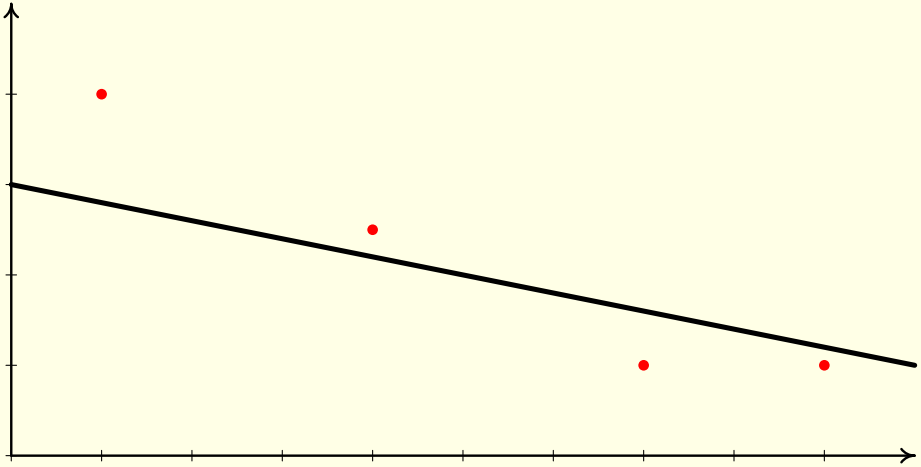
Least squares method



Data file

x	1	4	7	9
y	4	2.5	1	1

Least squares method

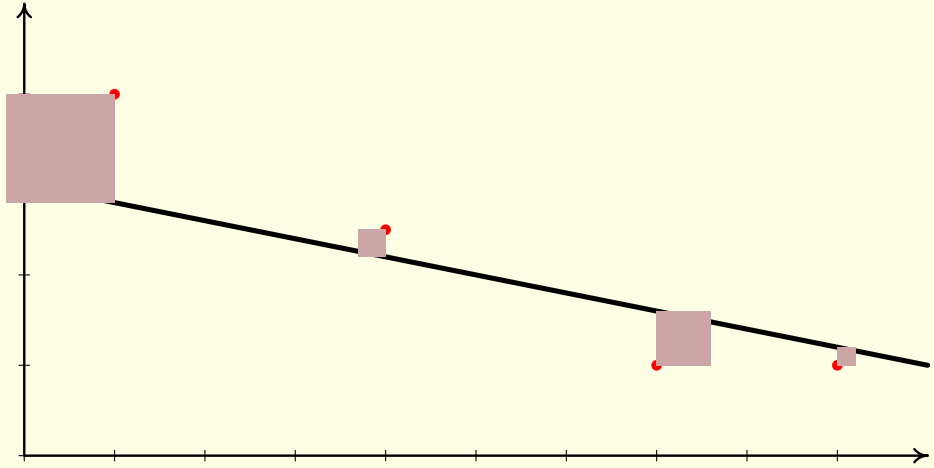


Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.

Least squares method

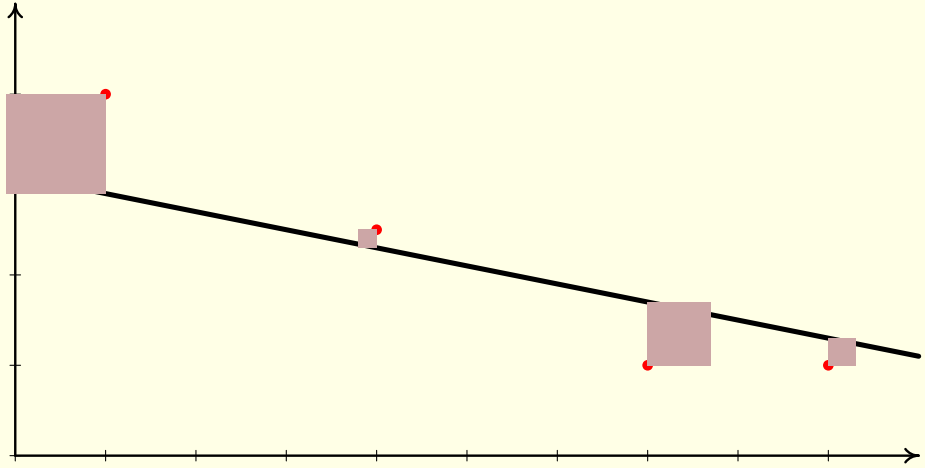


Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.
For the best linear fit is the total area of all red squares minimal.

Least squares method

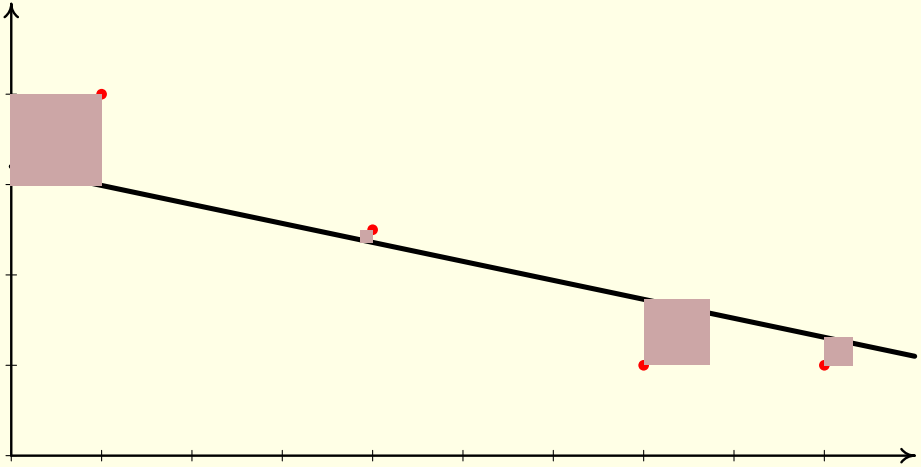


Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.
For the best linear fit is the total area of all red squares minimal.

Least squares method

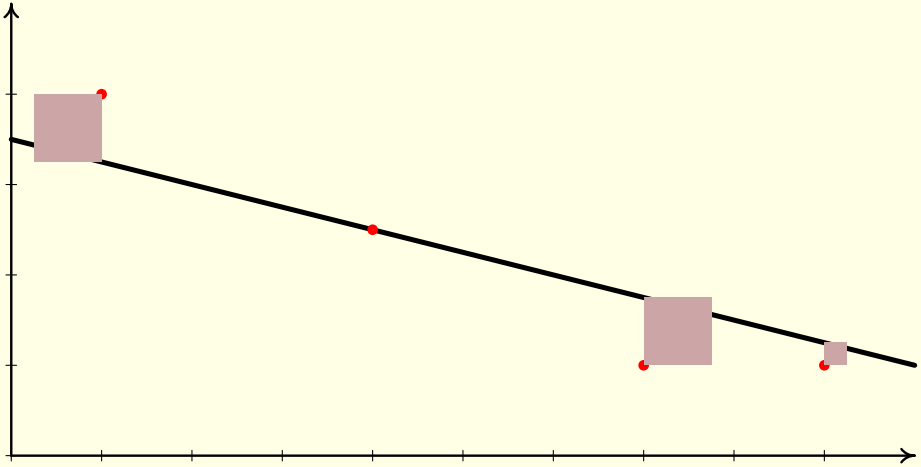


Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.
For the best linear fit is the total area of all red squares minimal.

Least squares method

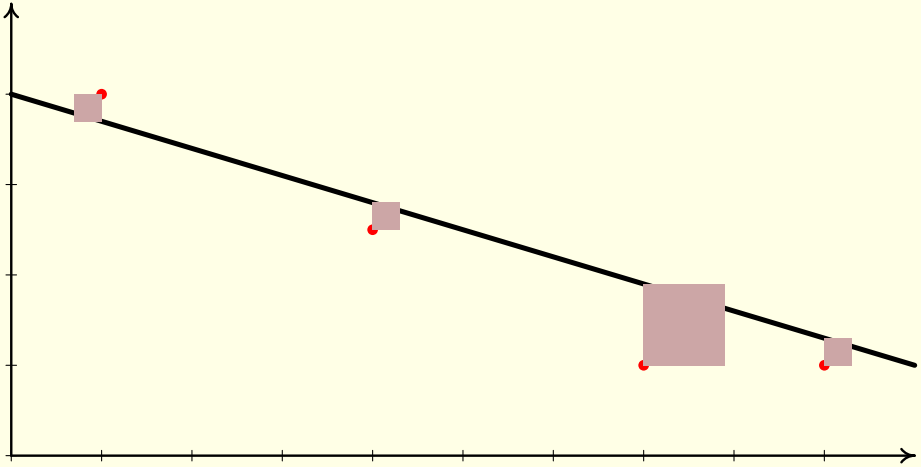


Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.
For the best linear fit is the total area of all red squares minimal.

Least squares method

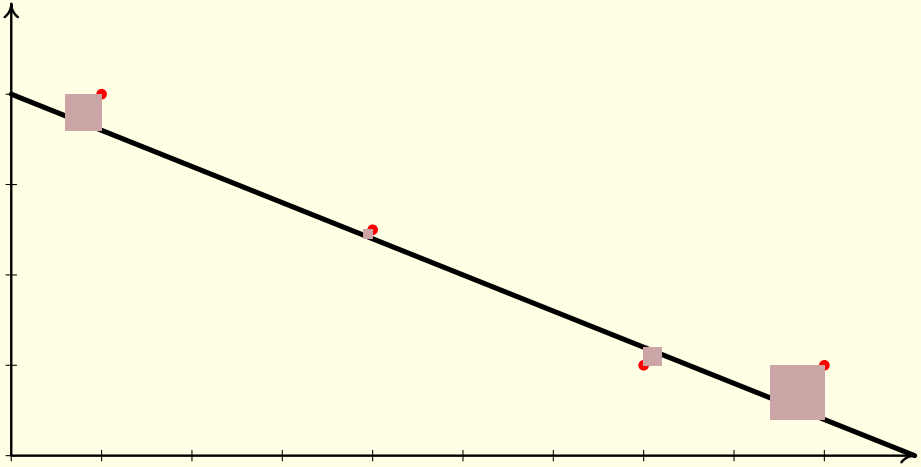


Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.
For the best linear fit is the total area of all red squares minimal.

Least squares method

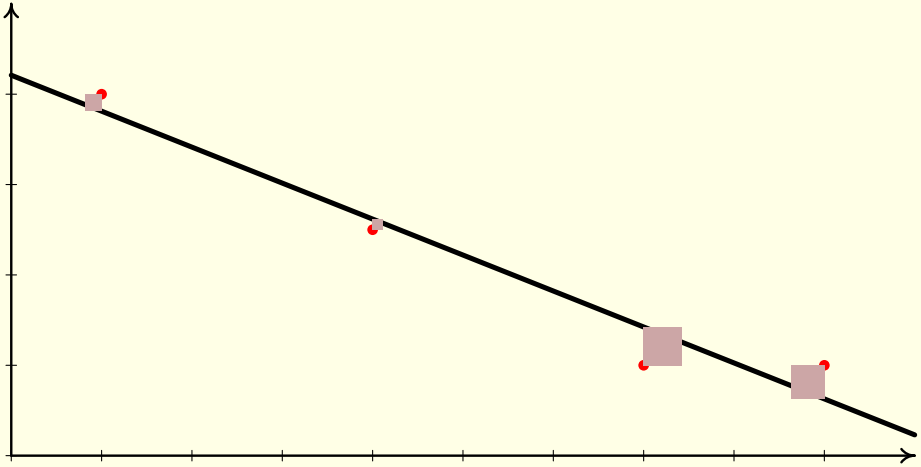


Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.
For the best linear fit is the total area of all red squares minimal.

Least squares method



Data file

x	1	4	7	9
y	4	2.5	1	1

We look for the best linear fit for the data file on the picture.
For the best linear fit is the total area of all red squares minimal.
This is the best linear fit. How to find the line?

2 Formula

Consider three points $[x_1, y_1]$, $[x_2, y_2]$ and $[x_3, y_3]$. The vertical distances between these ponits and the line $y = ax + b$ are $s_1 = |ax_1 + b - y_1|$, $s_2 = |ax_2 + b - y_2|$, $s_3 = |ax_3 + b - y_3|$ and we have ti find a minimum of the function

$$S(a, b) = (ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + (ax_3 + b - y_3)^2.$$

Local extremum appears in the point where all partial derivatives vanish.

$$\begin{aligned}\frac{\partial S}{\partial a} &= 2(ax_1 + b - y_1)x_1 + 2(ax_2 + b - y_2)x_2 + 2(ax_3 + b - y_3)x_3 \\ &= 2a(x_1^2 + x_2^2 + x_3^2) + 2b(x_1 + x_2 + x_3) - 2(x_1y_1 + x_2y_2 + x_3y_3)\end{aligned}$$

a

$$\begin{aligned}\frac{\partial S}{\partial b} &= 2(ax_1 + b - y_1) + 2(ax_2 + b - y_2) + 2(ax_3 + b - y_3) \\ &= 2a(x_1 + x_2 + x_3) + 2 \cdot 3b - 2(y_1 + y_2 + y_3).\end{aligned}$$

Putting these derivatives equal to zero and simplifying we get

$$\begin{aligned} a(x_1^2 + x_2^2 + x_3^2) &+ b(x_1 + x_2 + x_3) &&= x_1y_1 + x_2y_2 + x_3y_3, \\ a(x_1 + x_2 + x_3) &+ b \cdot 3 &&= y_1 + y_2 + y_3. \end{aligned}$$

2 Formula

Consider three points $[x_1, y_1]$, $[x_2, y_2]$ and $[x_3, y_3]$. The vertical distances between these ponits and the line $y = ax + b$ are $s_1 = |ax_1 + b - y_1|$, $s_2 = |ax_2 + b - y_2|$, $s_3 = |ax_3 + b - y_3|$ and we have ti find a minimum of the function

$$S(a, b) = (ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + (ax_3 + b - y_3)^2.$$

Local extremum appears in the point where all partial derivatives vanish.

$$\begin{aligned}\frac{\partial S}{\partial a} &= 2(ax_1 + b - y_1)x_1 + 2(ax_2 + b - y_2)x_2 + 2(ax_3 + b - y_3)x_3 \\ &= 2a(x_1^2 + x_2^2 + x_3^2) + 2b(x_1 + x_2 + x_3) - 2(x_1y_1 + x_2y_2 + x_3y_3)\end{aligned}$$

a

$$\begin{aligned}\frac{\partial S}{\partial b} &= 2(ax_1 + b - y_1) + 2(ax_2 + b - y_2) + 2(ax_3 + b - y_3) \\ &= 2a(x_1 + x_2 + x_3) + 2 \cdot 3b - 2(y_1 + y_2 + y_3).\end{aligned}$$

Putting these derivatives equal to zero and simplifying we get

$$\begin{aligned}a(x_1^2 + x_2^2 + x_3^2) &+ b(x_1 + x_2 + x_3) &&= x_1y_1 + x_2y_2 + x_3y_3, \\ a(x_1 + x_2 + x_3) &+ b \cdot 3 &&= y_1 + y_2 + y_3.\end{aligned}$$

2 Formula

Consider three points $[x_1, y_1]$, $[x_2, y_2]$ and $[x_3, y_3]$. The vertical distances between these ponits and the line $y = ax + b$ are $s_1 = |ax_1 + b - y_1|$, $s_2 = |ax_2 + b - y_2|$, $s_3 = |ax_3 + b - y_3|$ and we have ti find a minimum of the function

$$S(a, b) = (ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + (ax_3 + b - y_3)^2.$$

Local extremum appears in the point where all partial derivatives vanish.

$$\begin{aligned}\frac{\partial S}{\partial a} &= 2(ax_1 + b - y_1)x_1 + 2(ax_2 + b - y_2)x_2 + 2(ax_3 + b - y_3)x_3 \\ &= 2a(x_1^2 + x_2^2 + x_3^2) + 2b(x_1 + x_2 + x_3) - 2(x_1y_1 + x_2y_2 + x_3y_3)\end{aligned}$$

a

$$\begin{aligned}\frac{\partial S}{\partial b} &= 2(ax_1 + b - y_1) + 2(ax_2 + b - y_2) + 2(ax_3 + b - y_3) \\ &= 2a(x_1 + x_2 + x_3) + 2 \cdot 3b - 2(y_1 + y_2 + y_3).\end{aligned}$$

Putting these derivatives equal to zero and simplifying we get

$$\begin{aligned} a(x_1^2 + x_2^2 + x_3^2) &+ b(x_1 + x_2 + x_3) &&= x_1y_1 + x_2y_2 + x_3y_3, \\ a(x_1 + x_2 + x_3) &+ b \cdot 3 &&= y_1 + y_2 + y_3. \end{aligned}$$

Consider n points $[x_1, y_1]$, ..., $[x_n, y_n]$. The cost function is $S(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$ and partial derivatives are

$$\begin{aligned} \frac{\partial S}{\partial a} &= 2\sum_{i=1}^{n}(ax_i + b - y_i)x_i &&= 2\left(a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i\right), \\ \frac{\partial S}{\partial b} &= 2\sum_{i=1}^{n}(ax_i + b - y_i) &&= 2\left(a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} 1 - \sum_{i=1}^{n} y_i\right). \end{aligned}$$

Since $\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + \cdots + 1}_{n\text{-times}} = n$, the equations for stationary points are

$$\begin{aligned} a\sum_{i=1}^{n} x_i^2 \;+\; b\sum_{i=1}^{n} x_i \;&=\; \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i \;+\; b\,n \;&=\; \sum_{i=1}^{n} y_i. \end{aligned} \tag{1}$$

This is a system of linear equations in unknowns a and b. Solving this system we get the slope a and the y-intercept b of the best linear fit $y = ax + b$.

Consider n points $[x_1, y_1]$, $\ldots$, $[x_n, y_n]$. The cost function is $S(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$ and partial derivatives are

$$\begin{aligned} \frac{\partial S}{\partial a} &= 2\sum_{i=1}^{n}(ax_i + b - y_i)x_i &&= 2\left(a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i\right), \\ \frac{\partial S}{\partial b} &= 2\sum_{i=1}^{n}(ax_i + b - y_i) &&= 2\left(a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} 1 - \sum_{i=1}^{n} y_i\right). \end{aligned}$$

Since $\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + \cdots + 1}_{n\text{-times}} = n$, the equations for stationary points are

$$\begin{aligned} a\sum_{i=1}^{n} x_i^2 \;+\; b\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i \;+\; b\,n &= \sum_{i=1}^{n} y_i. \end{aligned} \tag{1}$$

This is a system of linear equations in unknowns a and b. Solving this system we get the slope a and the y-intercept b of the best linear fit $y = ax + b$.

3 Summary for the best linear fit

Theorem The line $y = ax + b$ is the best linear fit for the data file $[x_1, y_1]$, $[x_2, y_2]$, ..., $[x_n, y_n]$ iff the coefficients a, b satisfy

$$\begin{aligned} a\sum x_i^2 \;+\; b\sum x_i \;&= \sum x_i y_i \\ a\sum x_i \;+\; b\,n \;&= \sum y_i \end{aligned} \tag{2}$$

Problem: Find the best linear fit for the following data file.

x_i	0	1	3	5	6
y_i	5	3	3	2	1

Solution: the points are $[0, 5]$, $[1, 3]$, $[3, 3]$, $[5, 2]$ and $[6, 1]$. We have five points and hence $n = 5$. We do all computations in the following table

i	x_i	y_i		
1	0	5		
2	1	3		
3	3	3		
4	5	2		
5	6	1		
$\sum$				

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

i	x_i	y_i	x_i^2	
1	0	5	0	
2	1	3	1	
3	3	3	9	
4	5	2	25	
5	6	1	36	
$\sum$				

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

i	x_i	y_i	x_i^2	$x_i y_i$
1	0	5	0	0
2	1	3	1	3
3	3	3	9	9
4	5	2	25	10
5	6	1	36	6
$\sum$				

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

i	x_i	y_i	x_i^2	$x_i y_i$
1	0	5	0	0
2	1	3	1	3
3	3	3	9	9
4	5	2	25	10
5	6	1	36	6
$\sum$	15	14	71	28

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

i	x_i	y_i	x_i^2	$x_i y_i$
1	0	5	0	0
2	1	3	1	3
3	3	3	9	9
4	5	2	25	10
5	6	1	36	6
$\sum$	15	14	71	28

We write the system for coefficients

$$71a + 15b = 28,$$
$$15a + 5b = 14.$$

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

i	x_i	y_i	x_i^2	$x_i y_i$
1	0	5	0	0
2	1	3	1	3
3	3	3	9	9
4	5	2	25	10
5	6	1	36	6
$\sum$	15	14	71	28

We write the system for coefficients

$$\begin{aligned} 71a + 15b &= 28, \\ 15a + 5b &= 14. \end{aligned}$$

The solution of this system is $a = -\dfrac{7}{13} \doteq -0.538$ and $b = \dfrac{287}{65} \doteq 4.415$. The best linear fit is the line

$$\boxed{y = -0.538x + 4.415}.$$

Data file and the best linear fit are graphed on the picture.

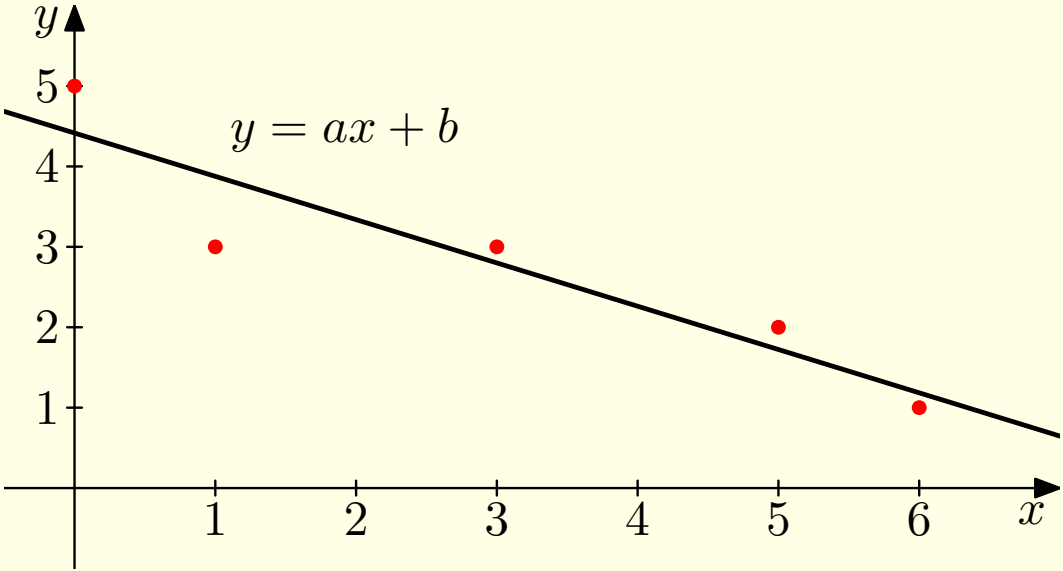


Figure 1: Least squares method.

4 Nonlinear fit

Common nonlinear functions can be simplified into linear functions by an algebraic modification and after a suitable substitution. Several important examples are shown in Table 1.

To fit data files by ...	substitute ...	and use a linear fit ....
$y = \frac{a}{x} + b$	$X = \frac{1}{x}$	$y = aX + b$
$y = ax^2 + b, \quad x > 0$	$X = x^2$	$y = aX + b$
$y = \sqrt{ax + b}$	$Y = y^2$	$Y = ax + b$
$y = be^{ax}, \quad y, b > 0$	$Y = \ln y,\ B = \ln b$	$Y = ax + B$
$y = bx^a, \quad x, y, b > 0$	$Y = \ln y,\ B = \ln b,\ X = \ln x$	$Y = aX + B$
$y = ax + \frac{b}{x}, \quad x > 0$	$X = x^2,\ Y = xy$	$Y = aX + b$

Table 1: Nonlinear fits

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

$$y = be^{ax}$$

The data file presents a rapidly decreasing relationship between x and y. The graph of this data file shows that the linear fitting cannot produce a good result. From this reason it seems to be better use the exponential function $y = be^{ax}$ as a mathematical model for this data file. We start with the exponential function

$$y = be^{ax}$$

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

$$
\begin{aligned}
y &= be^{ax} \\
\ln y &= \ln\left(be^{ax}\right) \\
\ln y &= \ln b + \ln e^{ax} \\
\ln y &= \ln b + ax.
\end{aligned}
$$

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

$$
\begin{aligned}
y &= be^{ax} \\
\ln y &= \ln\left(be^{ax}\right) \\
\ln y &= \ln b + \ln e^{ax} \\
\ln y &= \ln b + ax.
\end{aligned}
$$

Substituting $Y = \ln y$ and $B = \ln b$ we obtain a linear equation

$$Y = ax + B.$$

We fit the data file $[x_i, Y_i]$ by the linear function $Y = ax + B$.

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

i	x_i	y_i	$Y_i = \ln y_i$	x_i^2	$x_i Y_i$
1	0	45	3.807	0	0
2	1	20	2.996	1	2.996
3	3	5	1.609	9	4.828
4	5	2	0.693	25	3.466
5	6	1	0	36	0
$\sum$	15	73	9.105	71	11.290

$$y = be^{ax}$$
$$\ln y = \ln (be^{ax})$$
$$\ln y = \ln b + \ln e^{ax}$$
$$\ln y = \ln b + ax.$$

Substituting $Y = \ln y$ and $B = \ln b$ we obtain a linear equation

$$Y = ax + B.$$

$Y = \ln y$, $B = \ln b$, $Y = ax + B$

$$a \sum x_i^2 + B \sum x_i = \sum x_i Y_i$$
$$a \sum x_i + Bn = \sum Y_i$$

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

$$y = be^{ax}$$
$$\ln y = \ln\left(be^{ax}\right)$$
$$\ln y = \ln b + \ln e^{ax}$$
$$\ln y = \ln b + ax.$$

Substituting $Y = \ln y$ and $B = \ln b$ we obtain a linear equation

$$Y = ax + B.$$

i	x_i	y_i	$Y_i = \ln y_i$	x_i^2	$x_i Y_i$
1	0	45	3.807	0	0
2	1	20	2.996	1	2.996
3	3	5	1.609	9	4.828
4	5	2	0.693	25	3.466
5	6	1	0	36	0
$\sum$	15	73	9.105	71	11.290

$$\left.\begin{aligned} 71a + 15B &= 11.290 \\ 15a + 5B &= 9.105 \end{aligned}\right\} \Rightarrow$$

$Y = \ln y$, $B = \ln b$, $Y = ax + B$

$$a\sum x_i^2 + B\sum x_i = \sum x_i Y_i$$
$$a\sum x_i + Bn = \sum Y_i$$

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

$$y = be^{ax}$$
$$\ln y = \ln\left(be^{ax}\right)$$
$$\ln y = \ln b + \ln e^{ax}$$
$$\ln y = \ln b + ax.$$

Substituting $Y = \ln y$ and $B = \ln b$ we obtain a linear equation

$$Y = ax + B.$$

i	x_i	y_i	$Y_i = \ln y_i$	x_i^2	$x_i Y_i$
1	0	45	3.807	0	0
2	1	20	2.996	1	2.996
3	3	5	1.609	9	4.828
4	5	2	0.693	25	3.466
5	6	1	0	36	0
$\sum$	15	73	9.105	71	11.290

$$\left.\begin{aligned} 71a + 15B &= 11.290 \\ 15a + 5B &= 9.105 \end{aligned}\right\} \Rightarrow a = -0.616 \text{ and } B = 3.670$$

$Y = \ln y$, $B = \ln b$, $Y = ax + B$

$$a\sum x_i^2 + B\sum x_i = \sum x_i Y_i$$
$$a\sum x_i + Bn = \sum Y_i$$

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

$$y = be^{ax}$$
$$\ln y = \ln(be^{ax})$$
$$\ln y = \ln b + \ln e^{ax}$$
$$\ln y = \ln b + ax.$$

Substituting $Y = \ln y$ and $B = \ln b$ we obtain a linear equation

$$Y = ax + B.$$

i	x_i	y_i	$Y_i = \ln y_i$	x_i^2	$x_i Y_i$
1	0	45	3.807	0	0
2	1	20	2.996	1	2.996
3	3	5	1.609	9	4.828
4	5	2	0.693	25	3.466
5	6	1	0	36	0
$\sum$	15	73	9.105	71	11.290

$$\left.\begin{aligned} 71a + 15B &= 11.290 \\ 15a + 5B &= 9.105 \end{aligned}\right\} \Rightarrow a = -0.616 \text{ and } B = 3.670$$

$$\ln b = B = 3.670$$
$$b = e^B = e^{3.670} \doteq 39.253$$

$Y = \ln y$, $B = \ln b$, $Y = ax + B$

$$a\sum x_i^2 + B\sum x_i = \sum x_i Y_i$$
$$a\sum x_i + Bn = \sum Y_i$$

Fit the data file by an exponential function.

x_i	0	1	3	5	6
y_i	45	20	5	2	1

$$y = be^{ax}$$
$$\ln y = \ln\left(be^{ax}\right)$$
$$\ln y = \ln b + \ln e^{ax}$$
$$\ln y = \ln b + ax.$$

Substituting $Y = \ln y$ and $B = \ln b$ we obtain a linear equation

$$Y = ax + B.$$

We fit the data file $[x_i, Y_i]$ by the linear function $Y = ax + B$.

i	x_i	y_i	$Y_i = \ln y_i$	x_i^2	$x_i Y_i$
1	0	45	3.807	0	0
2	1	20	2.996	1	2.996
3	3	5	1.609	9	4.828
4	5	2	0.693	25	3.466
5	6	1	0	36	0
$\sum$	15	73	9.105	71	11.290

$$\left.\begin{aligned} 71a + 15B &= 11.290 \\ 15a + 5B &= 9.105 \end{aligned}\right\} \Rightarrow a = -0.616 \text{ and } B = 3.670$$

$$\ln b = B = 3.670$$
$$b = e^B = e^{3.670} \doteq 39.253$$

The best exponential fit is

$$y = 39.253 \cdot e^{-0.616x}.$$

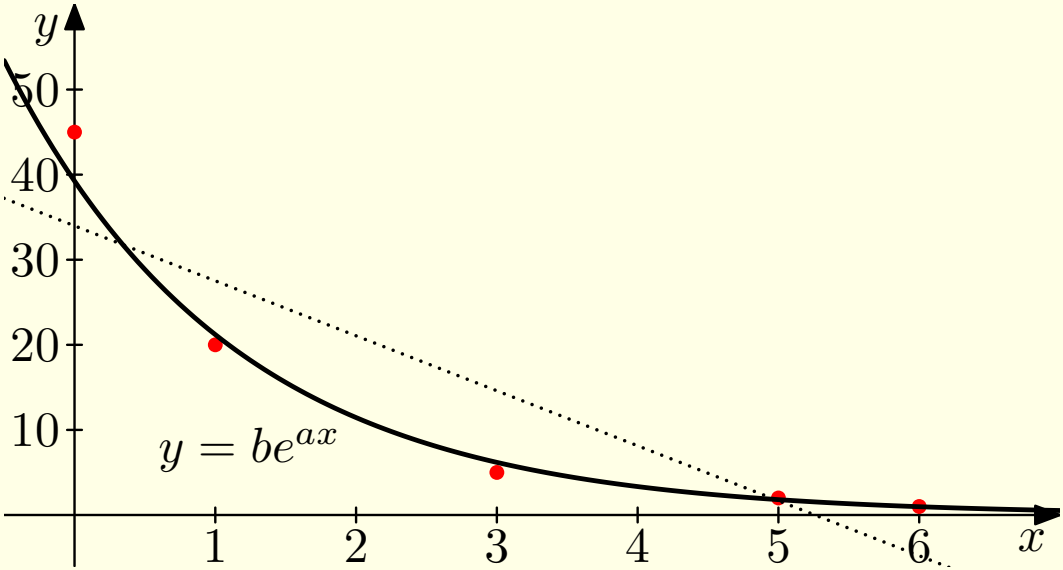


Figure 2: Exponential data fit

The graph of this exponential fit and the data file. You can also see the best linear fit, as a dotted line. It is clear that this linear fit is much worse mathematical model for the data file, comparing with the exponential fit.