

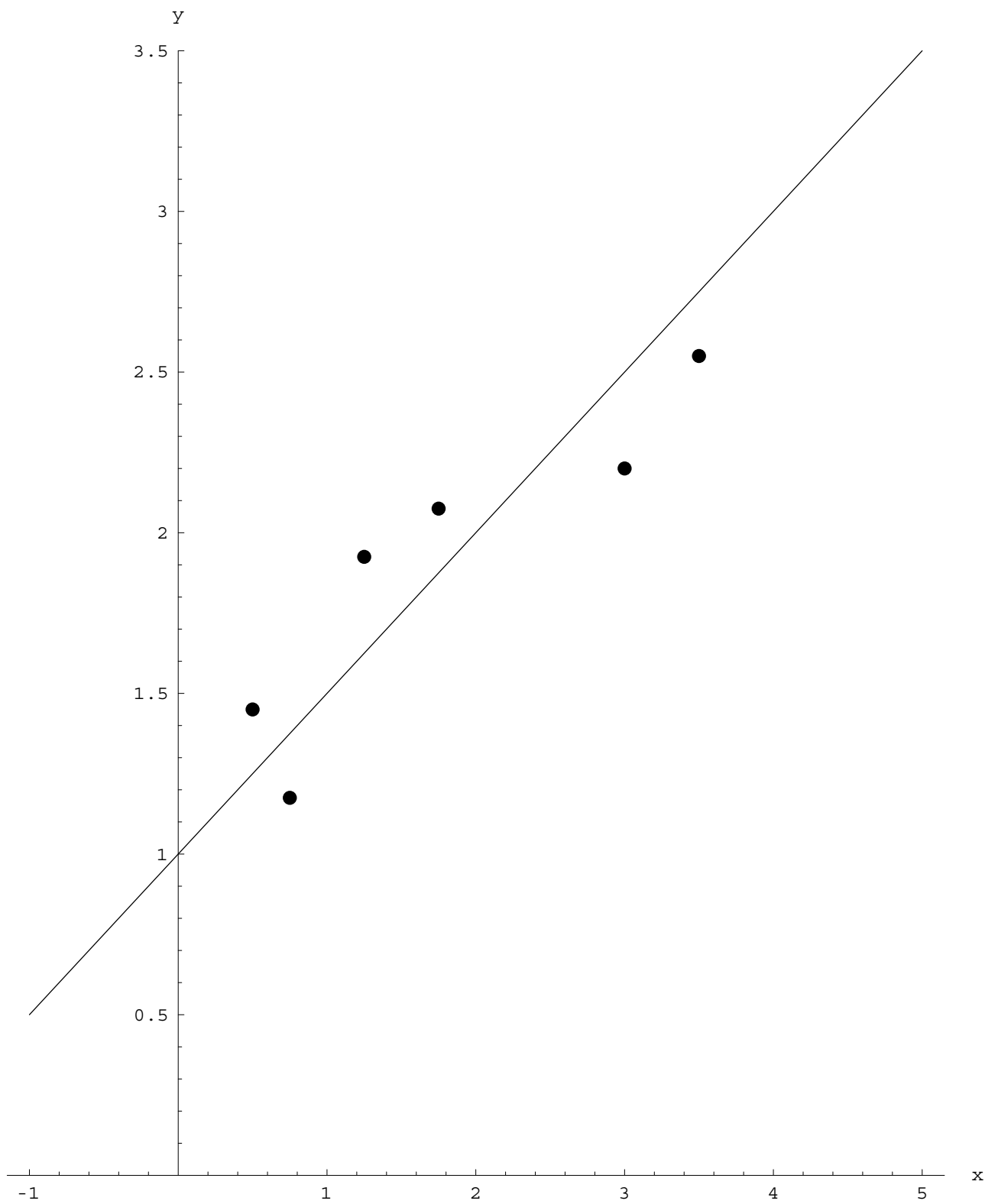
The Method of Least Squares¹

In this major section of the course the theme is optimisation; we initially presented three situations, of which we are up to the third.

In particular we shall, today, look at the problem of finding the “line which best fits a set of data points.”

This is an optimisation problem: we want the *optimal* fit of the line through the data points!

¹There is no coverage of this little topic in either *Stewart* or the *Notes*! However you may find some useful info in statistics textbooks under “*Linear Regression*” e.g. in [pp28-29 of ‘*Primer of Statistics*’ (4th Ed.) by Phipps and Quine].



Earlier in this unit we needed to find the straight line which best fits a collection of data points. We put our clear plastic ruler down on the graph paper and moved it until it “looked good”. Now that we have dealt with optimisation of functions of two variables we can be more precise and obtain the “best” fit!

Derivation

Let the data points be

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

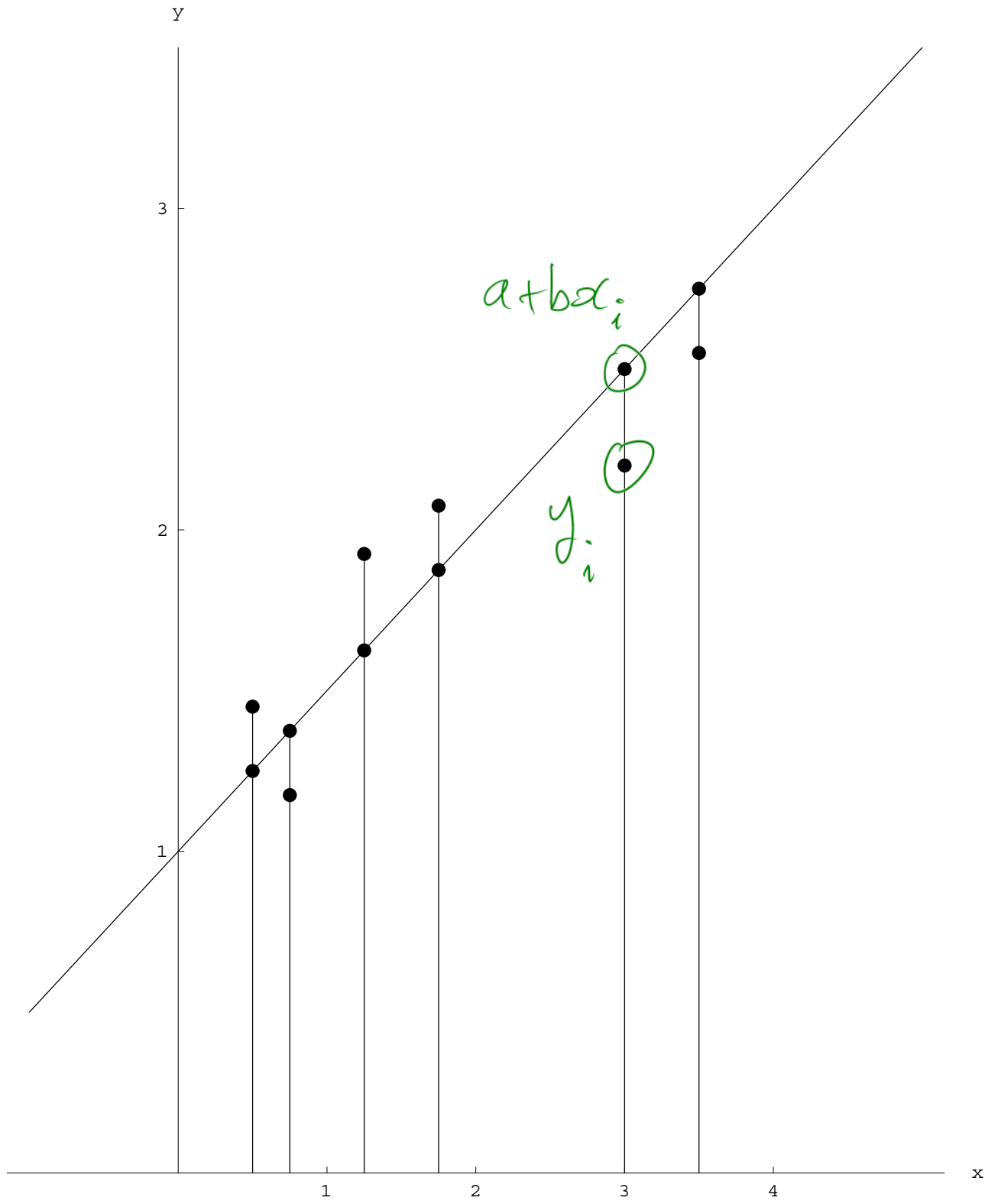
Let the line of “best fit” be

$$y = a + bx.$$

This line gives us “predicted” y -values:

$$a + bx_1, a + bx_2, \dots, a + bx_n$$

Let us make the measure of “goodness of fit” the sum of the squares of the deviations from predicted y -values.



Then E , the total “error” between the linear model and the data (also a measure of “goodness of fit”) is

$$E(a, b) := \overset{\text{actual} - \text{predicted}}{(y_1 - (a + bx_1))^2} + (y_2 - (a + bx_2))^2 + \dots + (y_n - (a + bx_n))^2.$$

E is a function of a and b , so we have an optimisation problem of a familiar type.

We differentiate

$$\begin{aligned}\frac{\partial E}{\partial a} &= 2(y_1 - a - bx_1)(-1) \\ &\quad + 2(y_2 - a - bx_2)(-1) + \\ &\quad \cdots + 2(y_n - a - bx_n)(-1)\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial b} &= 2(y_1 - a - bx_1)(-x_1) \\ &\quad + 2(y_2 - a - bx_2)(-x_2) + \\ &\quad \cdots + 2(y_n - a - bx_n)(-x_n).\end{aligned}$$

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

$$\begin{aligned}y_1 - a - bx_1 + y_2 - a - bx_2 + \\ \cdots + y_n - a - bx_n = 0\end{aligned}$$

$$\begin{aligned}x_1(y_1 - a - bx_1) + x_2(y_2 - a - bx_2) + \\ \cdots + x_n(y_n - a - bx_n) = 0.\end{aligned}$$

This leads to

$$(y_1 + y_2 + \cdots + y_n) - na \\ - b(x_1 + x_2 + \cdots + x_n) = 0$$

$$(x_1y_1 + x_2y_2 + \cdots + x_ny_n) \\ - a(x_1 + x_2 + \cdots + x_n) \\ - b(x_1^2 + x_2^2 + \cdots + x_n^2) = 0.$$

i.e. the line of best fit is given by $y = a + bx$ with a and b given by

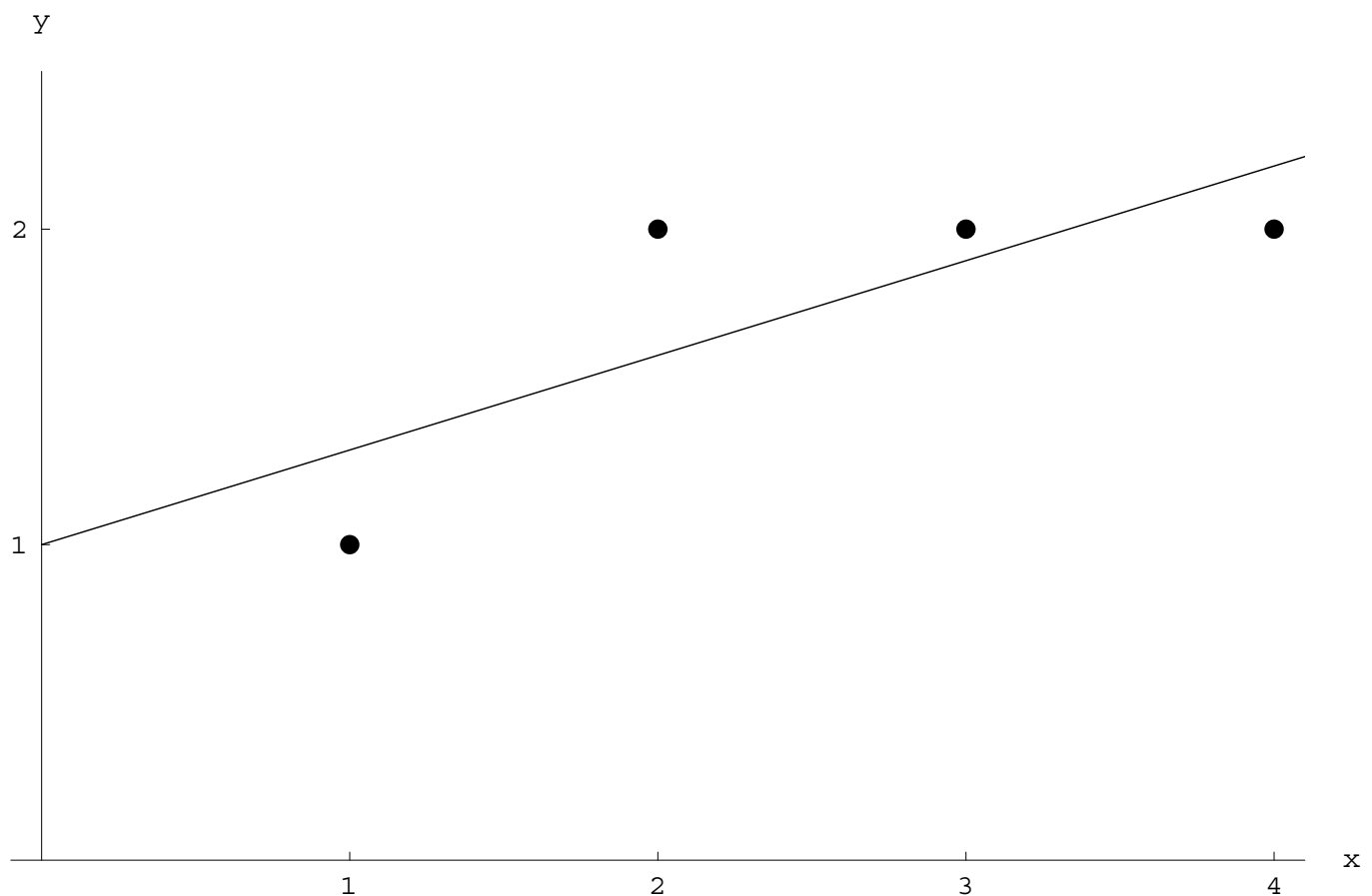
$$na + \left(\sum x_i\right) b = \left(\sum_{i=1}^n y_i\right) \\ \left(\sum x_i\right) a + \left(\sum x_i^2\right) b = \left(\sum x_i y_i\right)$$

It is easy to calculate the second order partial derivatives and check that the sole solution to these equations is a local minimum. Since the derivatives exist everywhere this the sole solution must be an absolute minimum.

Example

Find the least squares line of best fit for the following data.

x	1	2	3	4
y	1	2	2	2



We expand our table:

$n = 4$

	1	x	x^2	y	xy
	1	1	1	1	1
	1	2	4	2	4
	1	3	9	2	6
	1	4	16	2	8
Σ	4	10	30	7	19

This means that the equations for a and b are

$$4a + 10b = 7 \quad (1)$$

$$10a + 30b = 19 \quad (2)$$

(1) $\times 3$ yields

$$12a + 30b = 21 \quad (3)$$

(3)-(2) gives

$$2a = 2 \text{ so that } a = 1$$

$$10 + 30b = 19 \text{ so that } b = 0.3$$

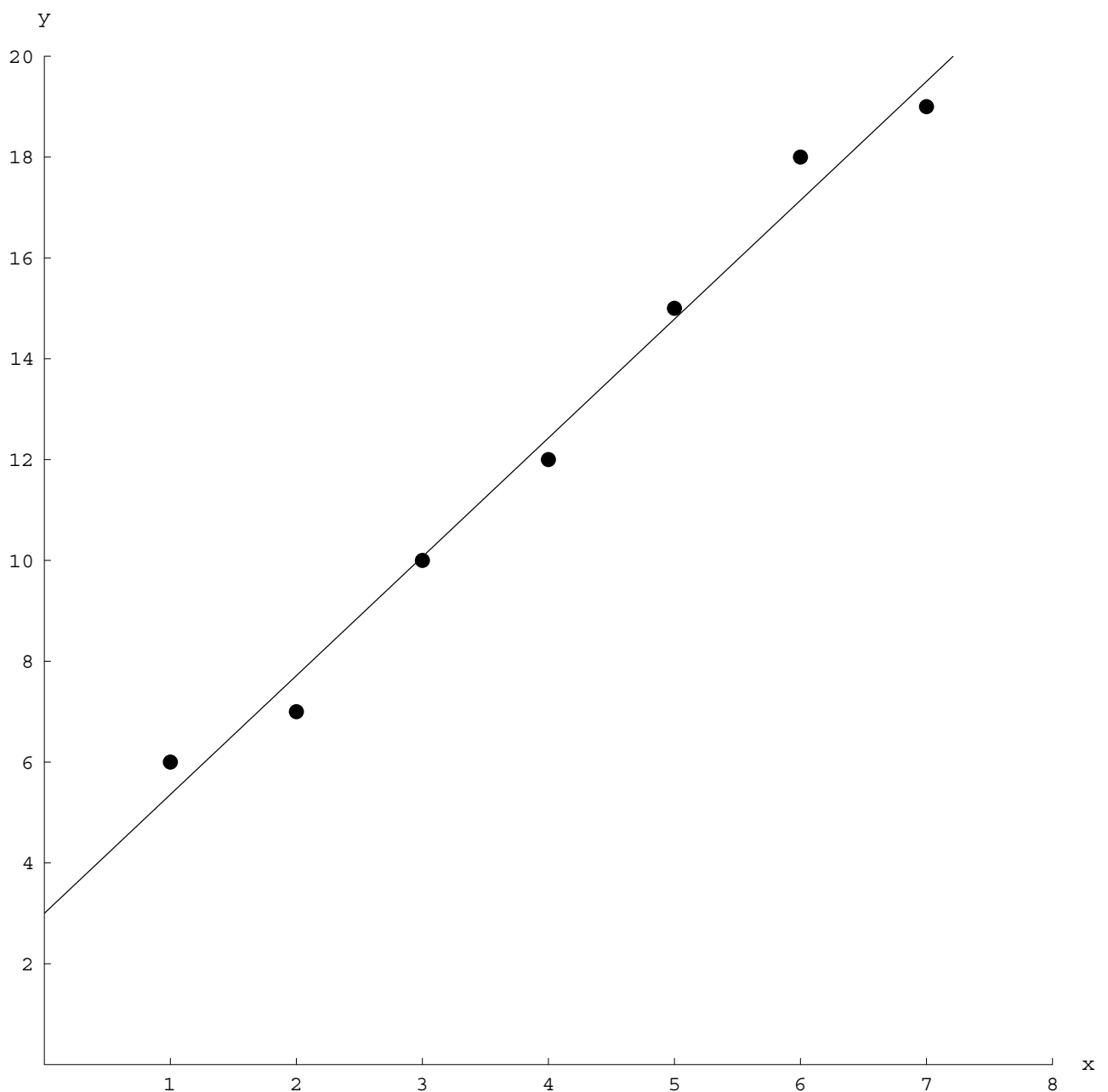
It follows that the least squares line of best fit is

$$y = 1 + 0.3x$$

Example

Find the least squares line of best fit for the following data.

x	1	2	3	4	5	6	7
y	6	7	10	12	15	18	19



We expand our table.

	1	x	x^2	y	xy
	1	1	1	6	6
	1	2	4	7	14
	1	3	9	10	30
	1	4	16	12	48
	1	5	25	15	75
	1	6	36	18	108
	1	7	49	19	133
Σ	7	28	140	87	414

The equations for a and b are

$$7a + 28b = 87 \quad (4)$$

$$28a + 140b = 414 \quad (5)$$

(4) $\times 4$ yields

$$28a + 112b = 348 \quad (6)$$

(5)-(6) gives

$$28b = 66 \text{ so that } b = \frac{33}{14}$$

$$7a + 66 = 87 \text{ so that } a = 3$$

It follows that the least squares line of best fit is

$$y = 3 + \frac{33}{14}x.$$

Example

Find the least squares line of best fit when plotting $Y = \ln y$ against $X = \ln x$ for the following data.

x	5	10	15	20	25
y	8	25	50	80	120

We expand our table. ($X = \ln x$ and $Y = \ln y$.)

	1	x	y	X	Y	X^2	XY
	1	5	8	1.6	2.1	2.56	3.36
	1	10	25	2.3	3.2	5.29	7.36
	1	15	50	2.7	3.9	7.29	10.53
	1	20	80	3.0	4.4	9.00	13.20
	1	25	120	3.2	4.8	10.25	15.36
Σ	5			12.8	18.8	34.38	49.81

This means that the equations for a and b are

$$5a + 12.8b = 18.4$$

$$12.8a + 34.38b = 49.81$$

$$a \approx -0.62$$

$$b \approx 1.68$$

It follows that the least squares line of best fit is

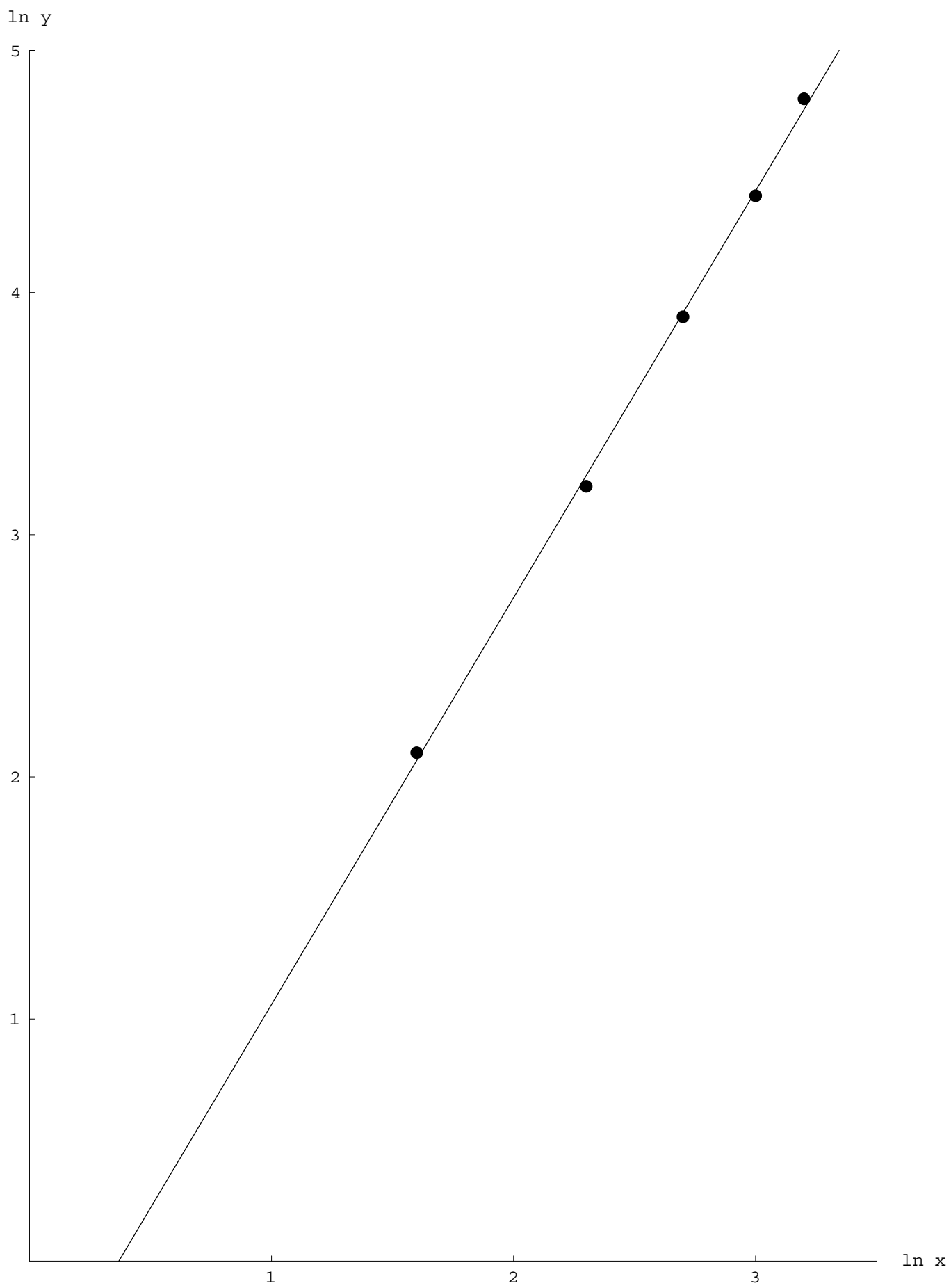
$$Y = -0.62 + 1.68X.$$

scaled line of best fit

$$y = e^{-0.62 + 1.68 \ln x}$$

$$\underline{\approx 0.5x^{1.7}} \text{ (1 d.p.)}$$

"curve of best fit"



Extension: polynomials of best fit (non-examinable)

The method of least squares can be generalized to give a method for finding the n th degree polynomial function of least squares best fit. For example, to find the least squares best fitting quadratic,

$$y = a + bx + cx^2,$$

we need to find the values of a , b and c which minimize the function

$$\begin{aligned} E(a, b, c) = & (a + bx_1 + cx_1^2 - y_1)^2 \\ & + (a + bx_2 + cx_2^2 - y_2)^2 \\ & + \cdots + (a + bx_n + cx_n^2 - y_n)^2. \end{aligned}$$

Here there are three independent variables a , b and c , but the procedure remains the same. The critical points are the points at which

$$\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = \frac{\partial E}{\partial c} = 0.$$

We get the following equations:

$$\begin{aligned} na + (\sum x_i) b + (\sum x_i^2) c &= \sum y_i, \\ (\sum x_i) a + (\sum x_i^2) b + (\sum x_i^3) c &= \sum x_i y_i, \\ (\sum x_i^2) a + (\sum x_i^3) b + (\sum x_i^4) c &= \sum x_i^2 y_i. \end{aligned}$$

Example

Find the least squares quadratic of best fit for the following data.

x	1	2	3	4
y	1	2	2	2

We expand our table.

	1	x	x^2	x^3	x^4	y	xy	xy^2
	1	1	1	1	1	1	1	1
	1	2	4	8	16	2	4	8
	1	3	9	27	81	2	6	18
	1	4	16	64	256	2	8	32
Σ	4	10	30	100	354	7	19	59

This means that the equations for a and b are

$$\begin{aligned}4a + 10b + 30c &= 7 \\10a + 30b + 100c &= 19 \\30a + 100b + 354c &= 59.\end{aligned}$$

The solution is

$$a = -0.25, \quad b = 1.55, \quad c = -0.25.$$

So the least squares best fitting quadratic is given by

$$y = -0.25 + 1.55x - 0.25x^2.$$

