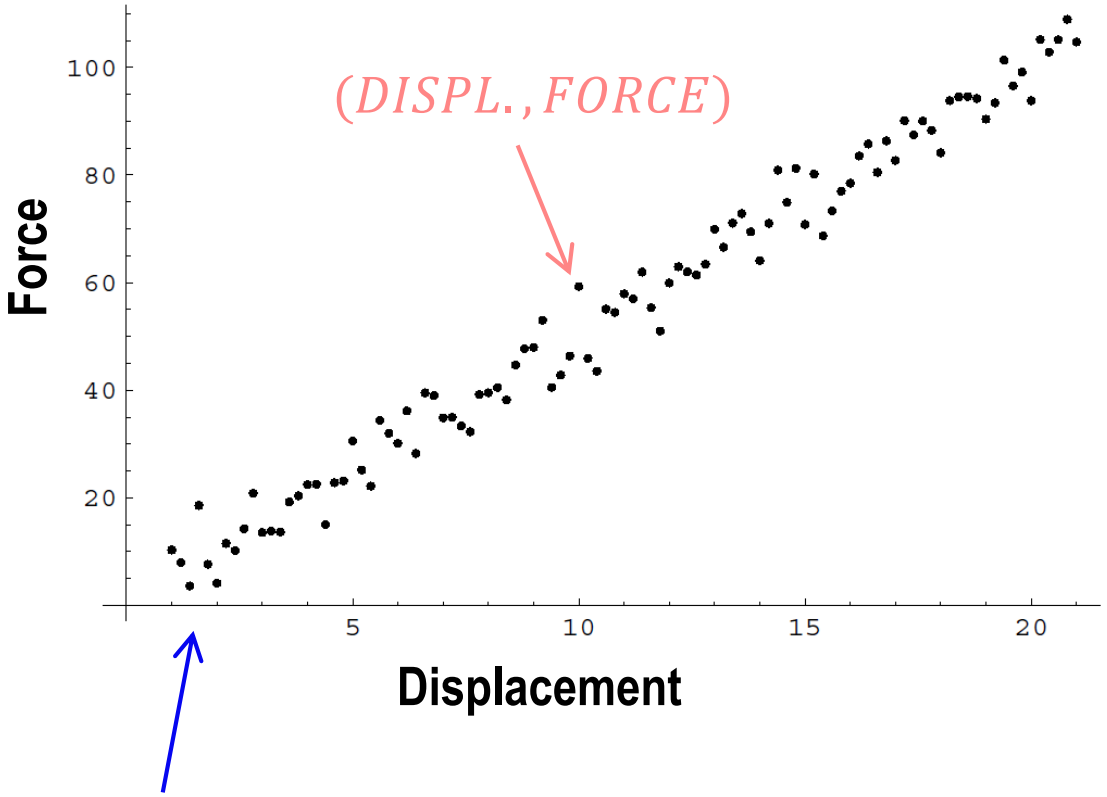
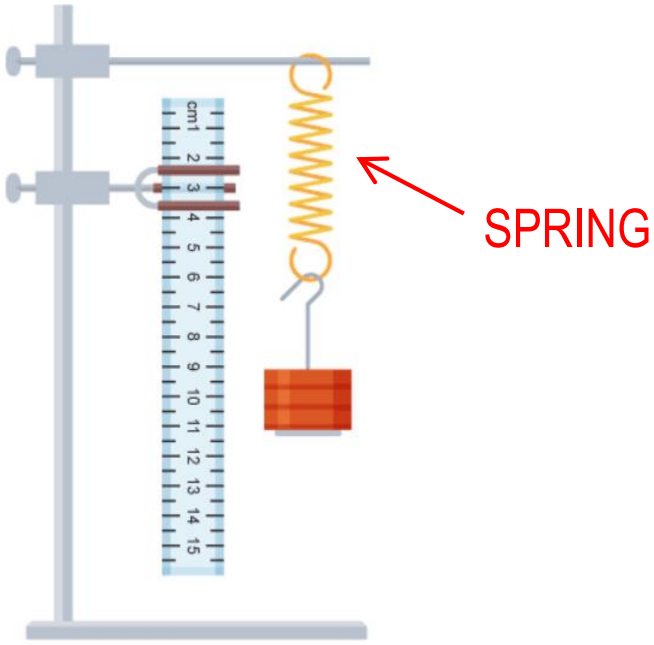
The Method of Least Squares

(application of stationary points for functions of two variables)

Hooke's Law



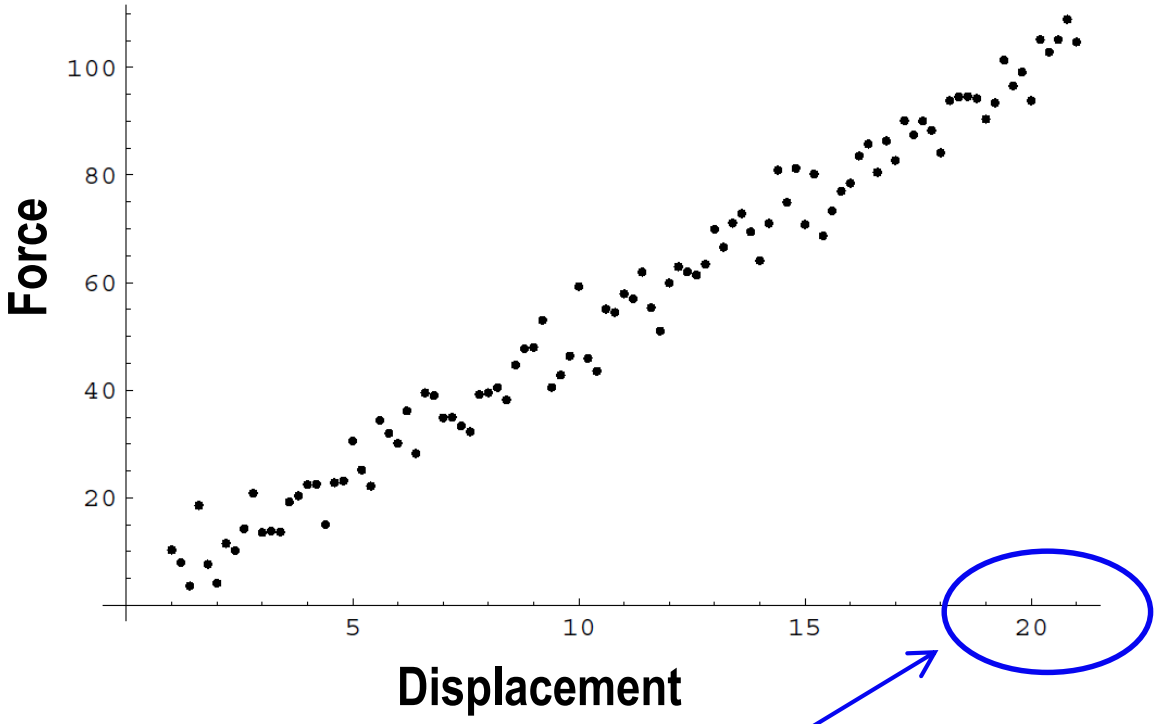
scatterplot

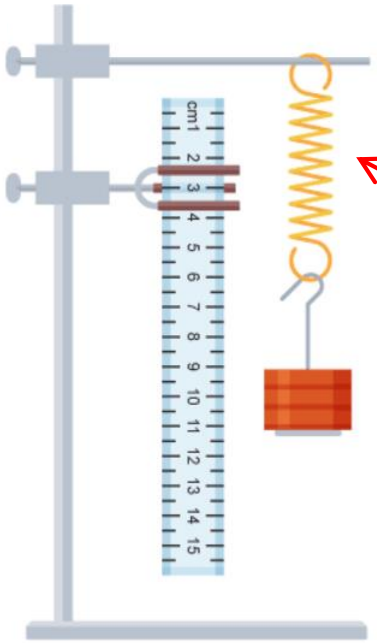


$F = kx$

(the **force** F applied to extend or compress an elastic spring by some **distance** x
Is directly proportional to that distance)

Hooke's Law





$F = kx$

displacement = 30, 35, 40, etc

what are the corresponding forces?

Regression line

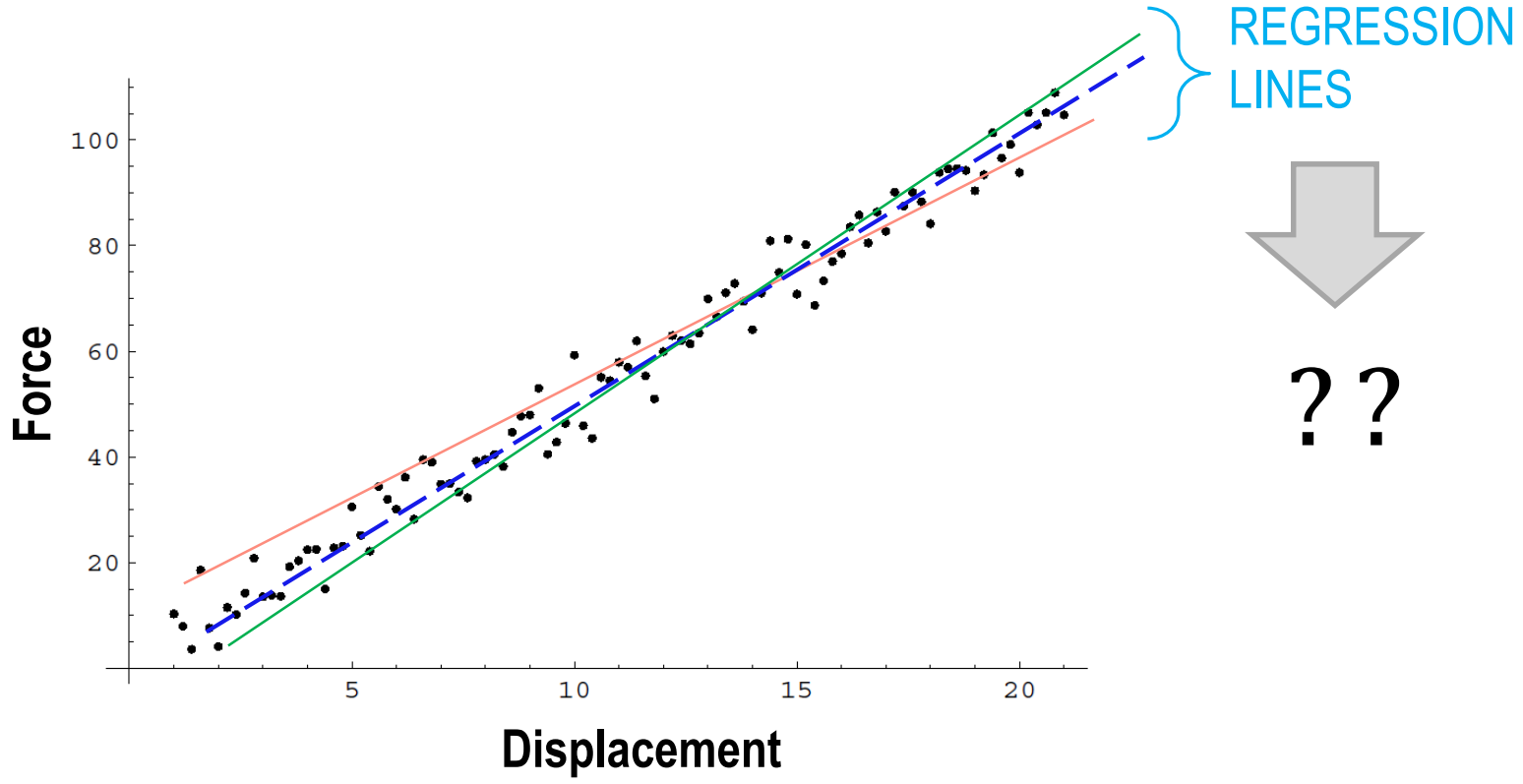
The University of Nottingham

Any straight line that describes how a **response** variable y changes as an **explanatory** variable x changes. It is often used to predict the value of y for a given value of x

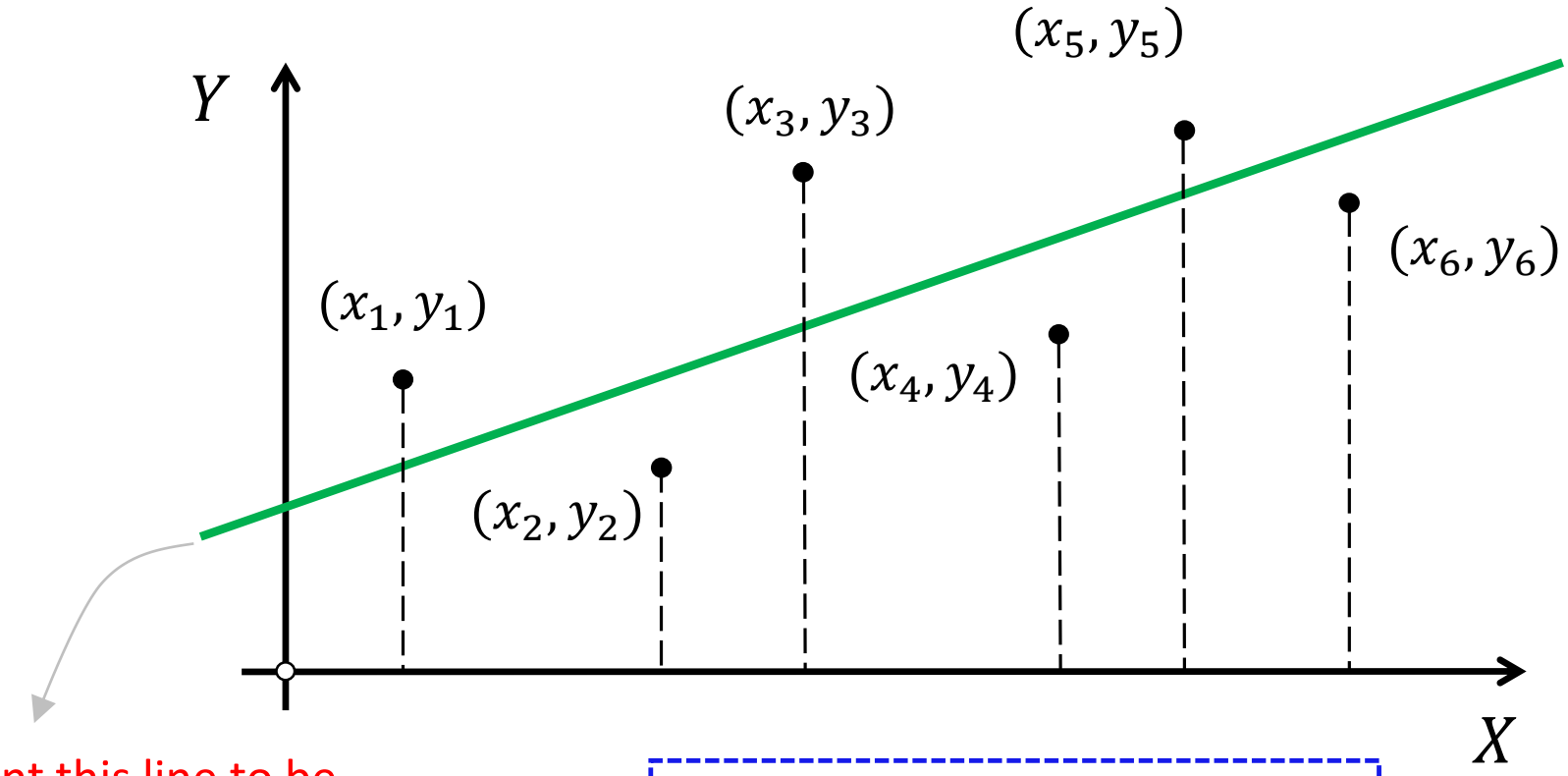
Different people will draw different lines on a scatterplot….

We need a way to draw the regression line that does **not** depend on our guess as to where the line should go.

Regression line



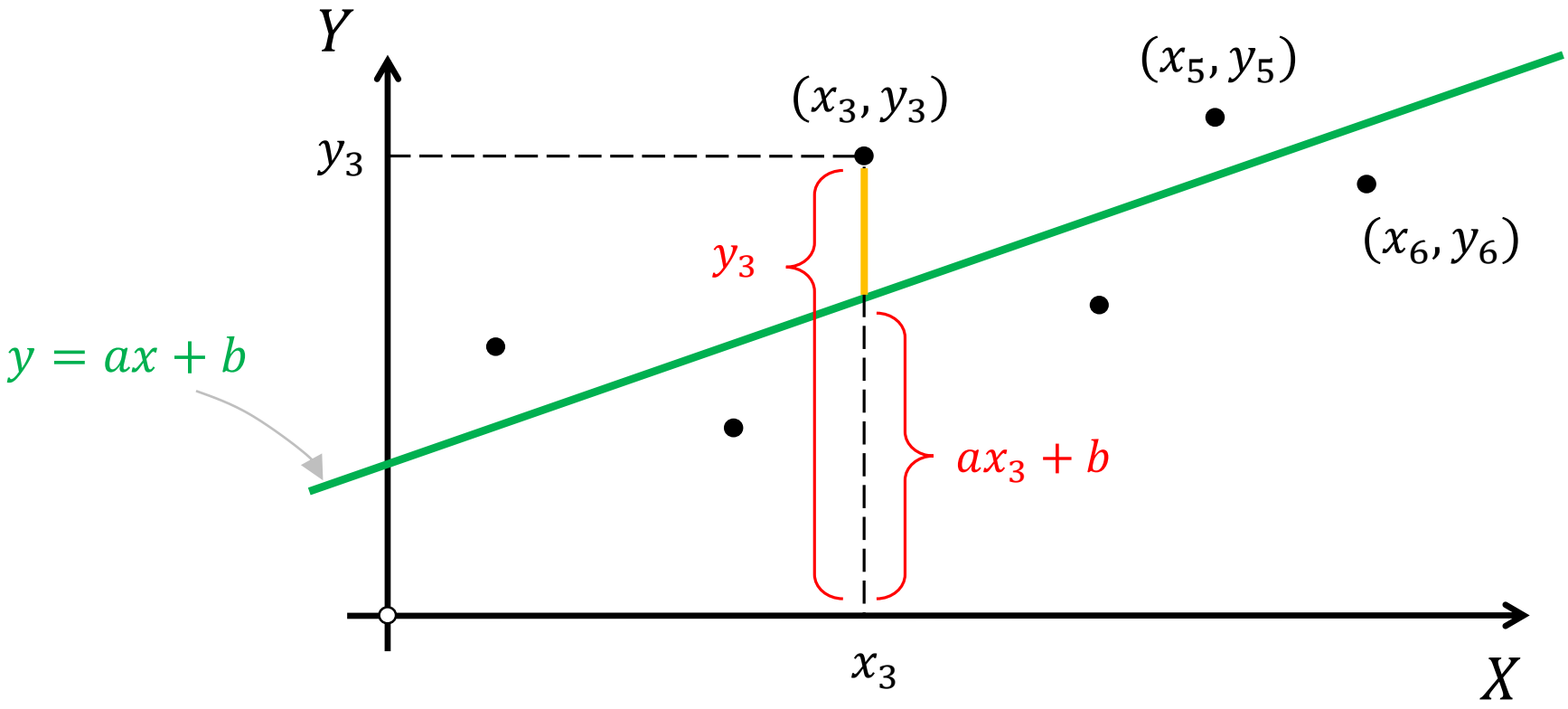
LEAST-SQUARES REGRESSION



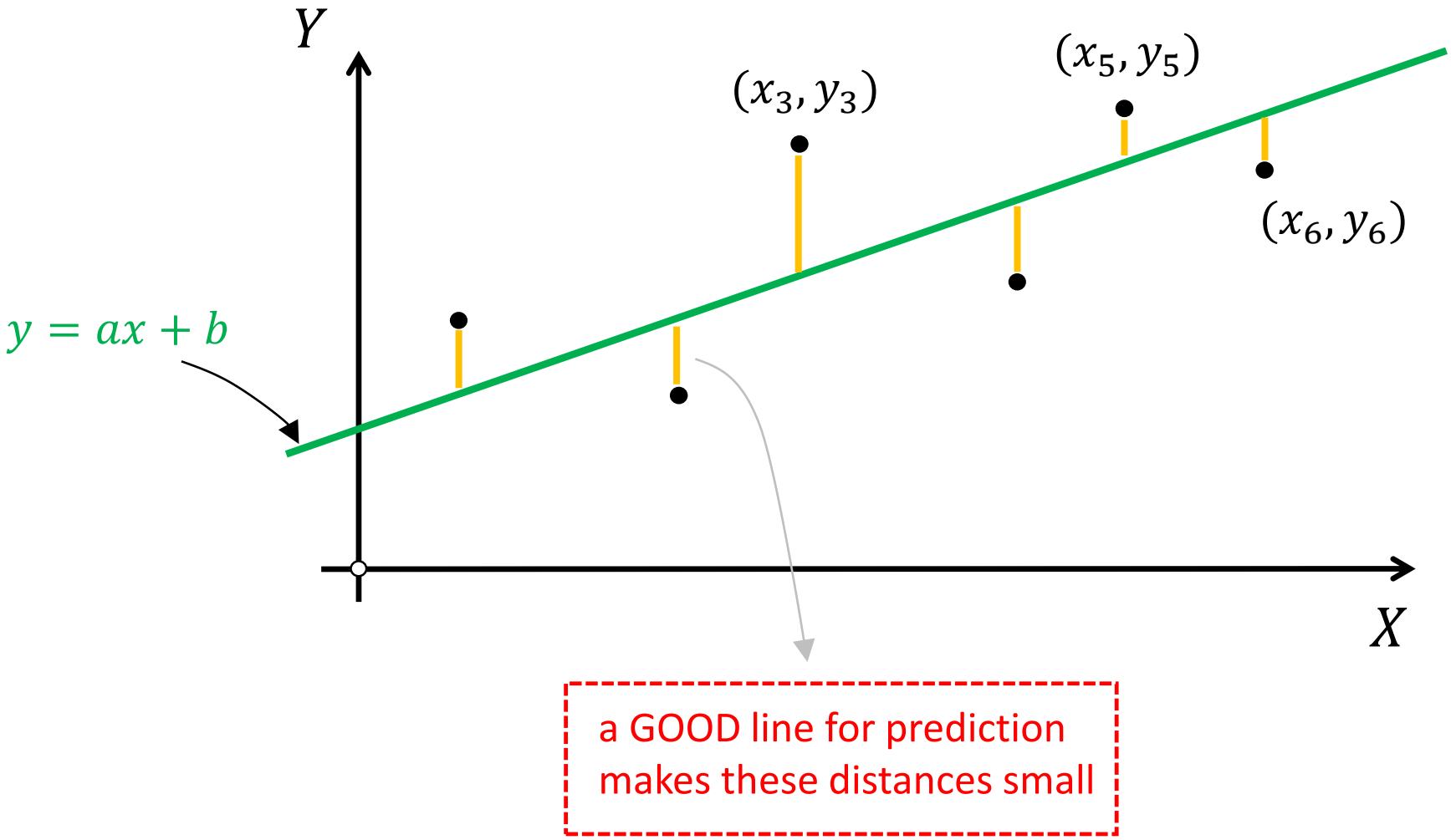
We want this line to be as close as possible to all of our data points

(REGRESSION LINE)

no line will pass exactly through all the points in the scatterplot

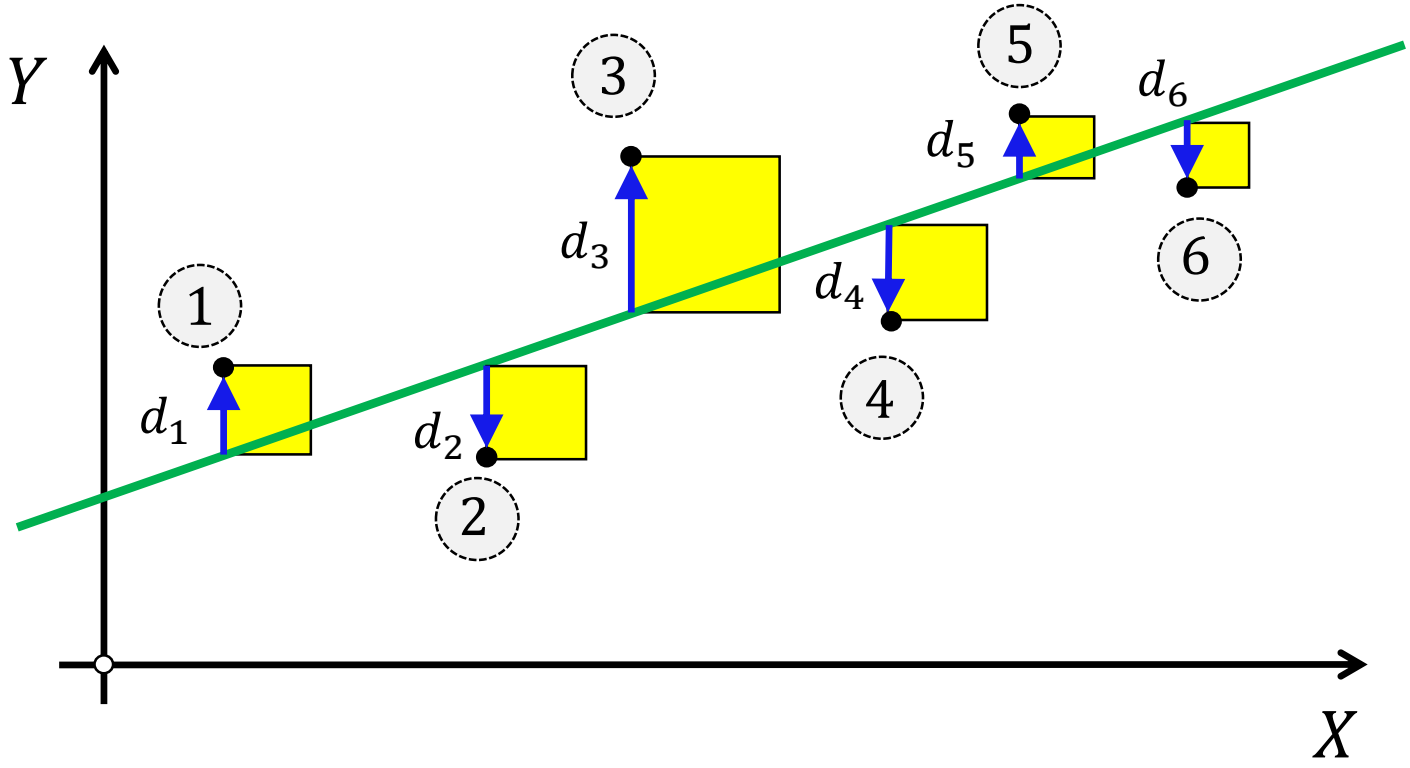


$y_3 - (ax_3 + b) =$ **vertical distance** from the **green line** to the point (x_3, y_3)


Y
(x_3, y_3)
(x_5, y_5)
(x_6, y_6)
$y = ax + b$
X
a GOOD line for prediction
makes these distances small

"BEST FIT" = MINIMIZE THE TOTAL AREAS OF THE **SQUARES**

regression line



$d_1, \ldots, d_6 =$ vertical distances from the data points to the **regression line**

(the sum of the vertical **blue vectors** is **zero**
if the green line is the least-squares regression line)

Least-squares regression line

Always passes through the point $(\bar{x}, \bar{y})$

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

average of the x-values

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$$

average of the y-values

Least Squares Fit (LSF)

The University of Nottingham

Suppose we have a set of data

$$X = \{x_1, x_2, x_3, \dots, x_n\} \qquad \text{(INPUTS)}$$

with corresponding measurements

$$Y = \{y_1, y_2, y_3, \dots, y_n\} \qquad \text{(OUTPUTS)}$$

We suspect that there is an approximate linear relationship between X and Y, i.e.

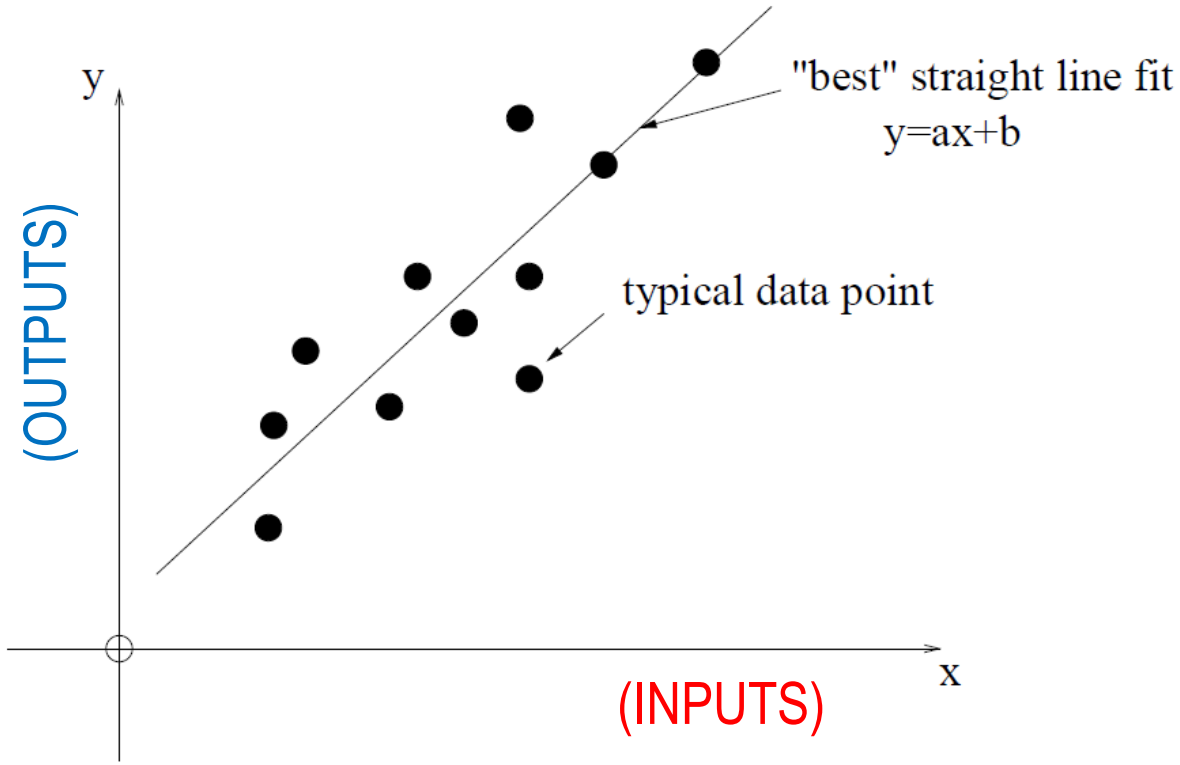
$$y_i \approx ax_i + b, \qquad i = 1,2, \dots, n$$

and want find the **best fit** of the data to the models' parameters *a* and *b*?

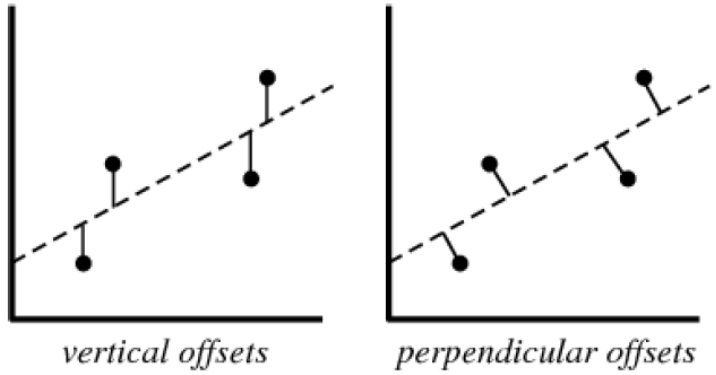
Ciprian D. Coman

LSF

"**Best fit**" means that the data points (x_1, y_1), (x_2, y_2), (x_3, y_3).... are close (in some sense) to the line $y = ax + b$:



LSF



We wish to minimize
the "distance" between
each point (x_i, y_i) and
the line of best fit
(the one we are looking for).

- At the data point (x_i, y_i), the difference between the data y_i, and the calculated value on the straight line $ax_i + b$, is

$$d_i = y_i - (ax_i + b).$$

- A measure of the overall error is then

$$S(a,b) = \sum_{i=1}^{n} d_i^2 \quad = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 .$$

LSF

- The least error (i.e. best fit) occurs when S is a minimum as a function of a and b.

$$\text{i.e.} \quad \frac{\partial S}{\partial a} = 0 \quad \text{and} \quad \frac{\partial S}{\partial b} = 0.$$

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

- These can be re-arranged to give

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

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

  which are two simultaneous equations to be solved for a and b.
- These are sometimes called the 'normal equations'.

Example 1

Find the least squares best fit to the data

x_i	0	10	20	30	40	50
y_i	1.2	3.1	4.8	6.8	9.2	10.9

Solution: We calculate

x_i^2	0	100	400	900	1600	2500
$x_i y_i$	0	31	96	204	368	545

so that $\sum_{i=1}^{6} x_i = 150, \quad \sum_{i=1}^{6} y_i = 36$

$$\sum_{i=1}^{6} x_i^2 = 5500, \quad \sum_{i=1}^{6} x_i y_i = 1244, \quad \sum_{i=1}^{6} 1 = 6.$$

Example 1

The normal equations give

$$150a + 6b = 36$$

$$5500a + 150b = 1244.$$

Solving these simultaneous equations gives $a = 0.1966$ and $b = 1.0857$, and so the best straight line fit has equation

$$y = 0.1966x + 1.0857. \qquad \square$$

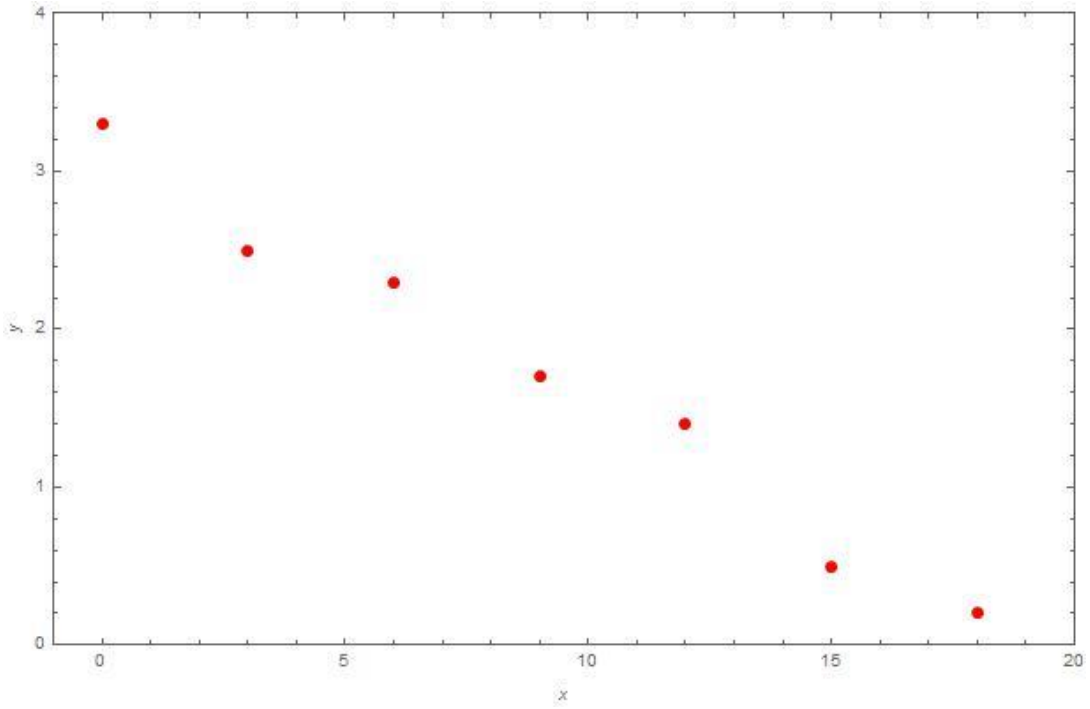
We can calculate the value of y, y_{calc}, at the data points to compare the result we get with the 'measured' value y_i:

x_i	0	10	20	30	40	50
y_i	1.2	3.1	4.8	6.8	9.2	10.9
y_{calc}	1.09	3.05	5.02	6.98	8.95	10.91

Example 2

Find the least squares fit to the following data

x	0	3	6	9	12	15	18
y	3.3	2.5	2.3	1.7	1.4	0.5	0.2



Example 2

Solution: There are 7 data points, so

$$\sum_{i=1}^{7} 1 = 7$$

and

$$\sum_{i=1}^{7} x_i = 63, \quad \sum_{i=1}^{7} x_i^2 = 819, \quad \sum_{i=1}^{7} = 11.9, \quad \sum_{i=1}^{7} x_i y_i = 64.5$$

The normal equations are

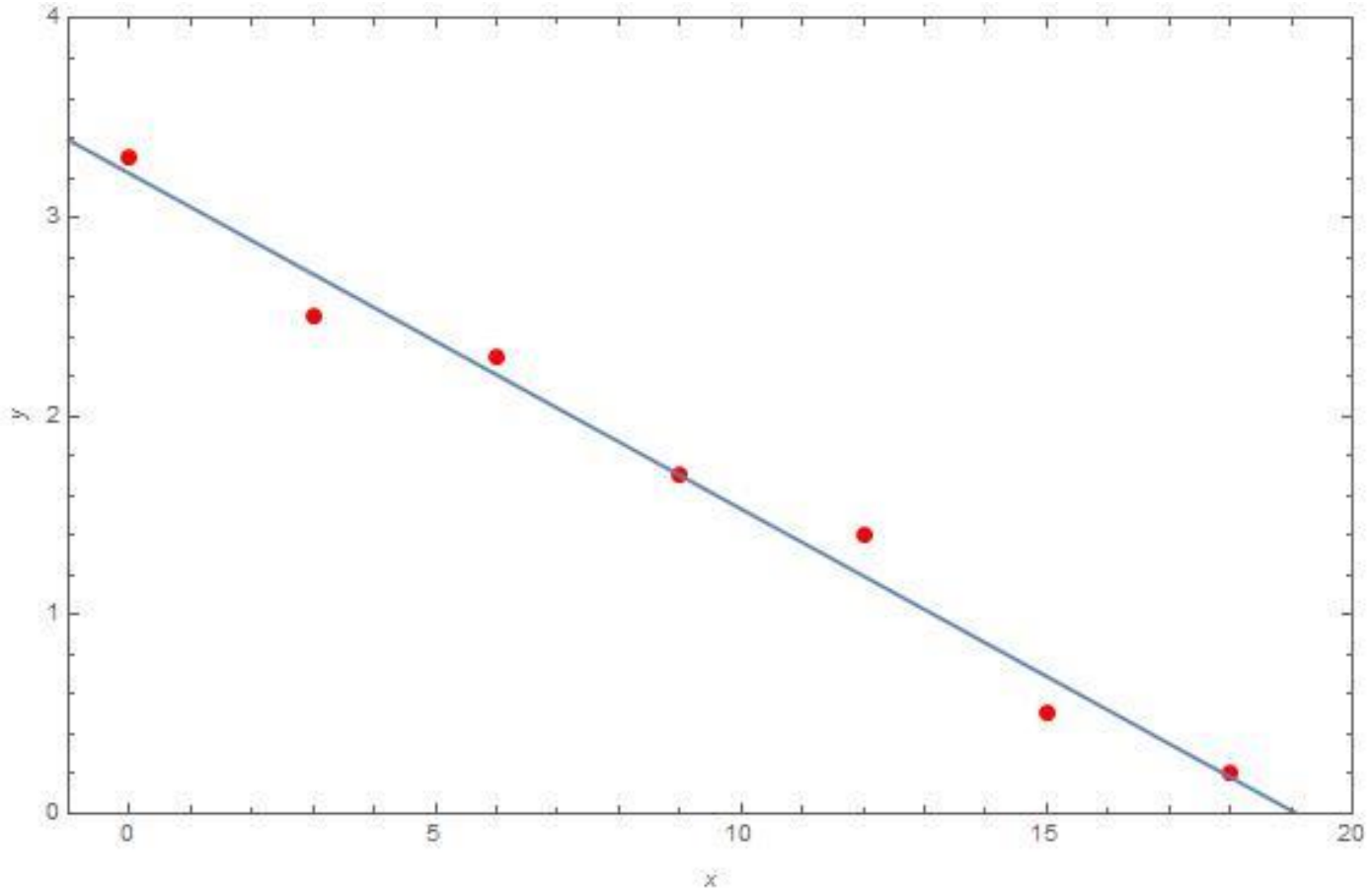
$$64.5 - 819a - 63b = 0 \qquad \text{and} \qquad 11.9 - 63a - 7b = 0\,.$$

Solving these gives $a = -0.169$ and $b = 3.221$, so the least squares fit is

$$y = -0.169x + 3.221$$

(see next page for comparison)

Example 2



Observations

By solving the system for a and b it can be shown that

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

(SLOPE) (INTERCEPT)

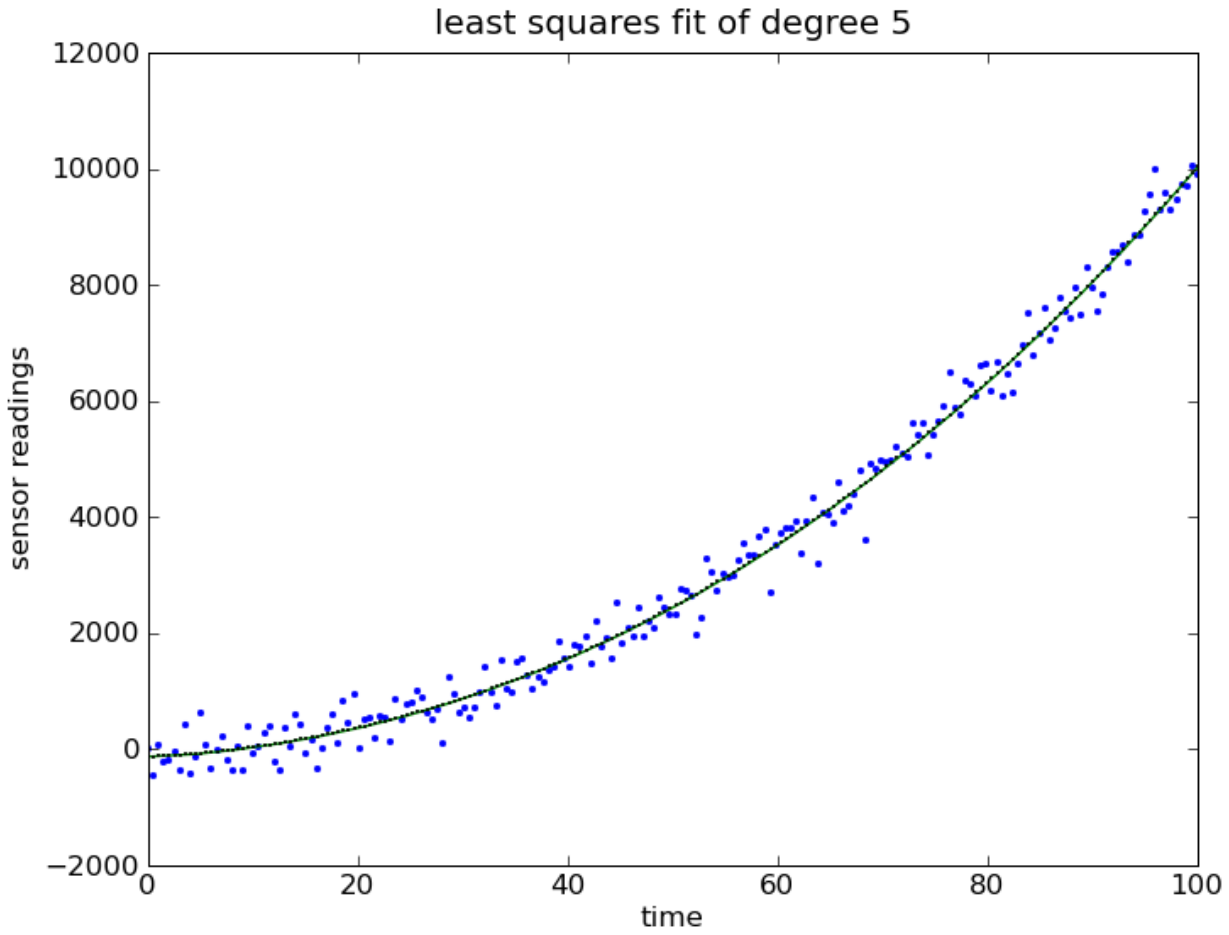
where

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

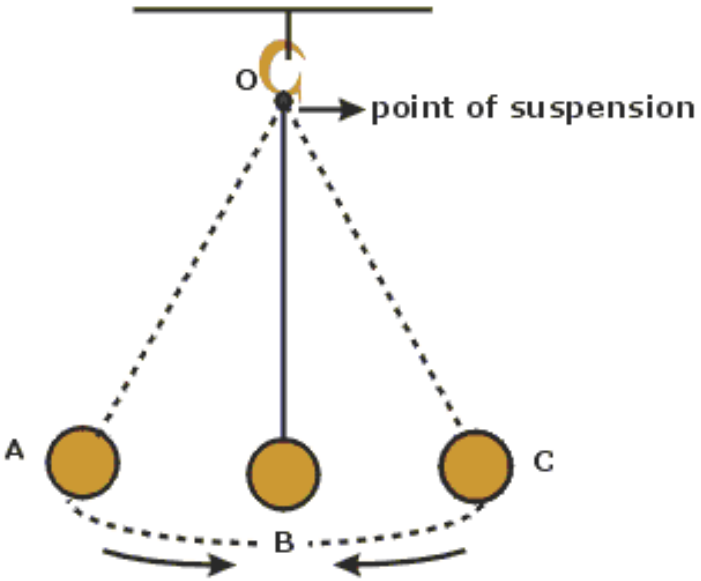
$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$$

Observations

Many other functions can be fitted if the linear model is not adequate



Simple pendulum



$$T = \frac{2\pi}{\sqrt{g}} L^{1/2}$$

PERIOD

LENGTH

choose: $L_1, L_2, \ldots, L_n$

measure: $T_1, T_2, \ldots, T_n$

$$T_i \approx A L_i^{\alpha}$$

$$A, \alpha = ?$$

General case

The University of Nottingham

This sort of thing is also useful if we suspect that the two sets of data obey a **scaling law** of the form

$$y_i \approx K(x_i)^{\alpha} \qquad\qquad i = 1,2,\dots,n$$

The question is how can we find the two unknown parameters K and α.
(least squares fit is the answer)

Take the 'log' of the above relation $\Longrightarrow$ $\ln y_i = \alpha \ln x_i + \ln K$ $\qquad i = 1,2,\dots,n$

my new 'a' my new 'b' (see previous slide)

So we have reduced the problem to fitting a straight line $y = ax + b$ to

$$\{\ln x_1, \ln x_2, \dots, \ln x_n\} \quad \& \quad \{\ln y_1, \ln y_2, \dots, \ln y_n\}$$