Loughborough University

Introduction to Linear Regression

The Method of Least-squares

Ciprian D Coman

Aims of the session

Loughborough University

- Introduce the idea behind an important class of statistical methods (known as **regression methods**)
- Motivate the least-squares criterion used to define a **regression line** (in the simplest possible setting)

Motivation

Loughborough University

A warehouse manager of a company dealing in large quantities of steel cable needs to be able to estimate how much cable is left on his partially used drums. A random sample of twelve partially used drums is taken and each drum is weighed and the corresponding length of cable measured. The results are given in the table below:

Weight of drum and cable (x) kg.	Measured length of cable (y) m.
30	70
40	90
40	100
50	120
50	130
50	150
60	160
70	190
70	200
80	200
80	220
80	230

Motivation (cont'd)

Loughborough University

Weight of drum and cable (x) kg.	Measured length of cable (y) m.
30	70
40	90
40	100
50	120
50	130
50	150
60	160
70	190
70	200
80	200
80	220
80	230

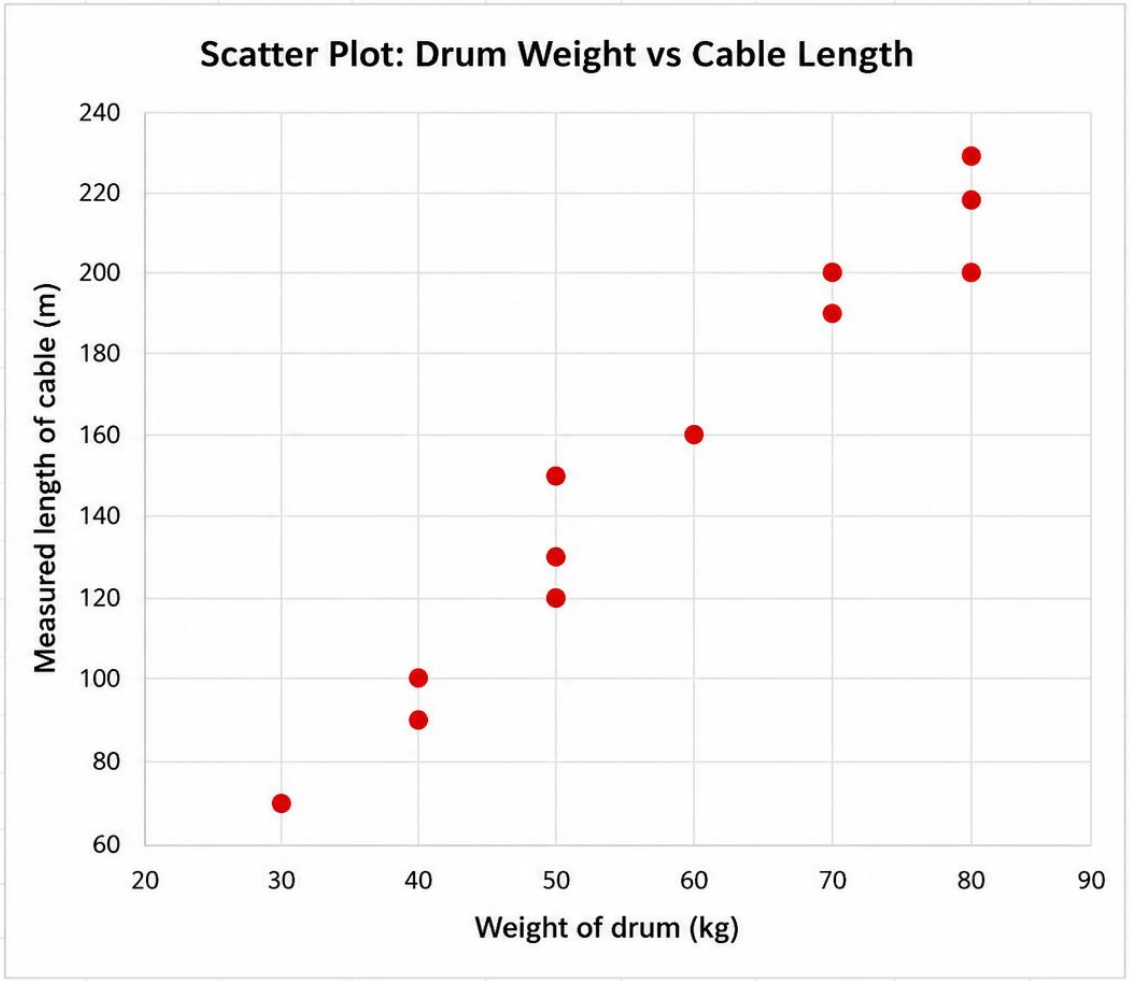
The manager wants **to predict** the lengths of cable left on drums whose weights are:

(i) 35 kg (ii) 85 kg (iii) 100 kg

Scatter plot

Loughborough University

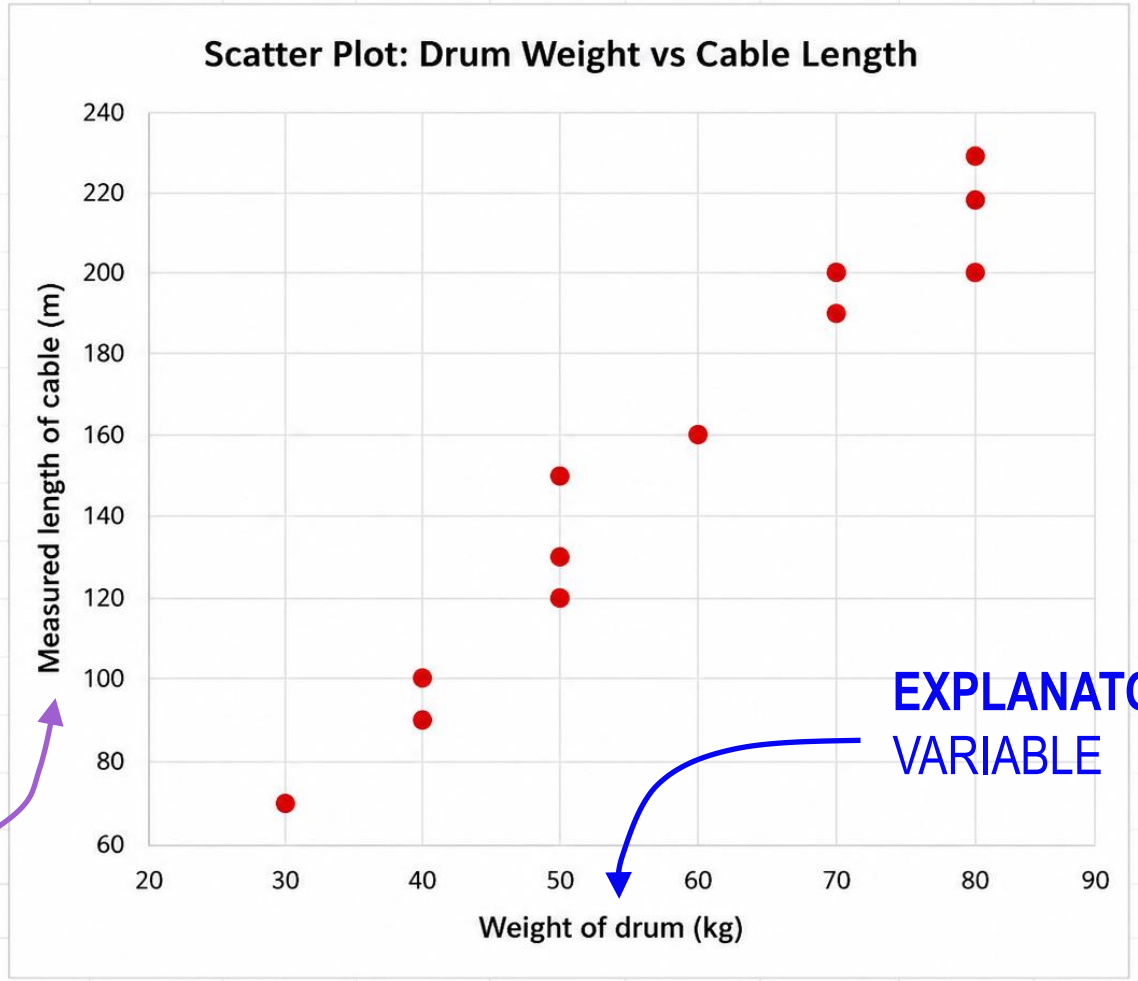
	A	B
1	Weight of drum (kg)	Measured length of cable (m)
2	30	70
3	40	90
4	40	100
5	50	120
6	50	130
7	50	150
8	60	160
9	70	190
10	70	200
11	80	200
12	80	220
13	80	230



Scatter plot

Loughborough University

	A	B
1	Weight of drum (kg)	Measured length of cable (m)
2	30	70
3	40	90
4	40	100
5	50	120
6	50	130
7	50	150
8	60	160
9	70	190
10	70	200
11	80	200
12	80	220
13	80	230

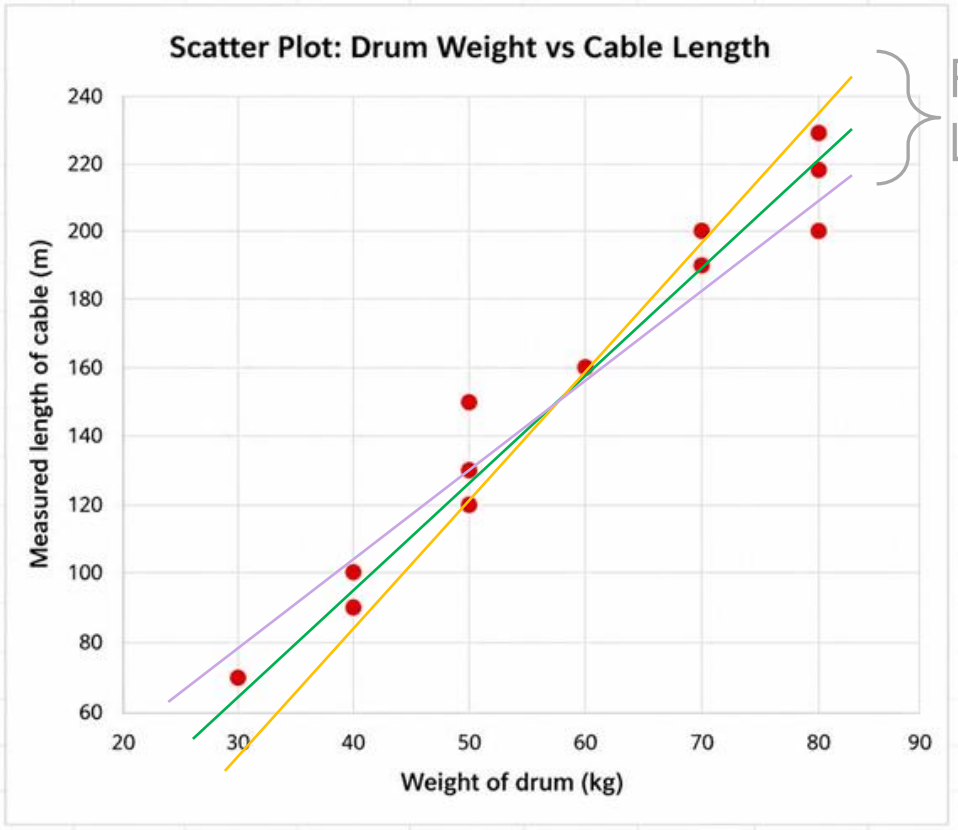


RESPONSE VARIABLE

EXPLANATORY VARIABLE

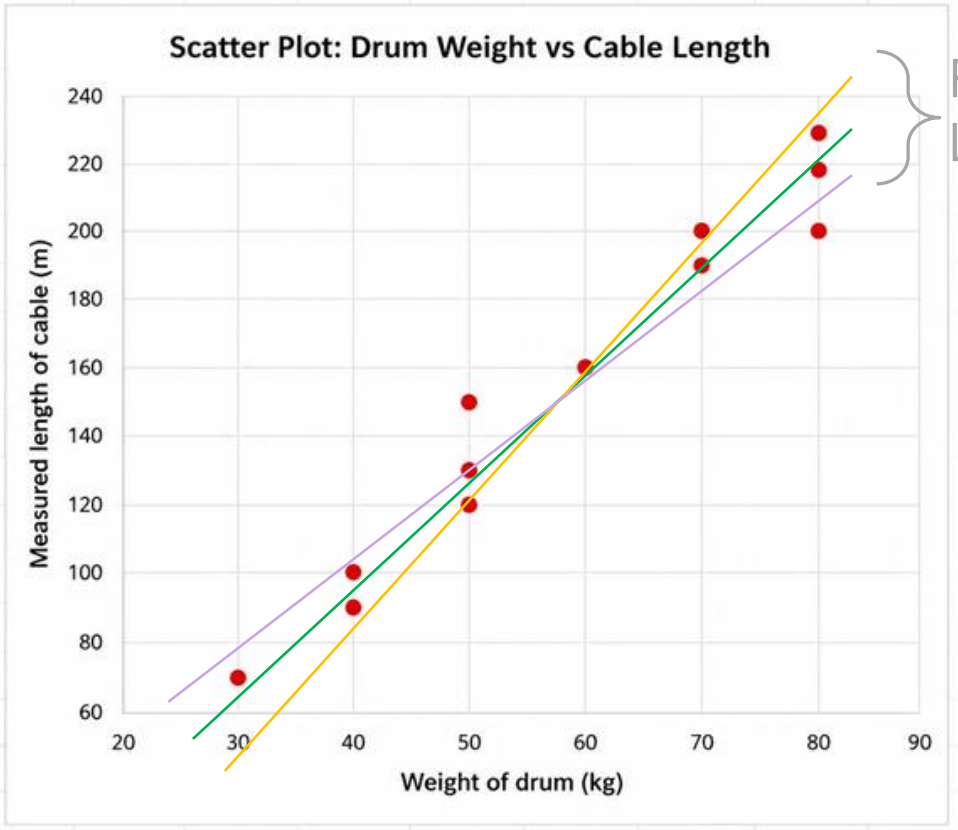
Regression line

A regression line is a straight line used to model how a **response** variable y changes as an **explanatory** variable x changes.



Regression line

A regression line is a straight line used to model how a **response** variable y changes as an **explanatory** variable x changes.

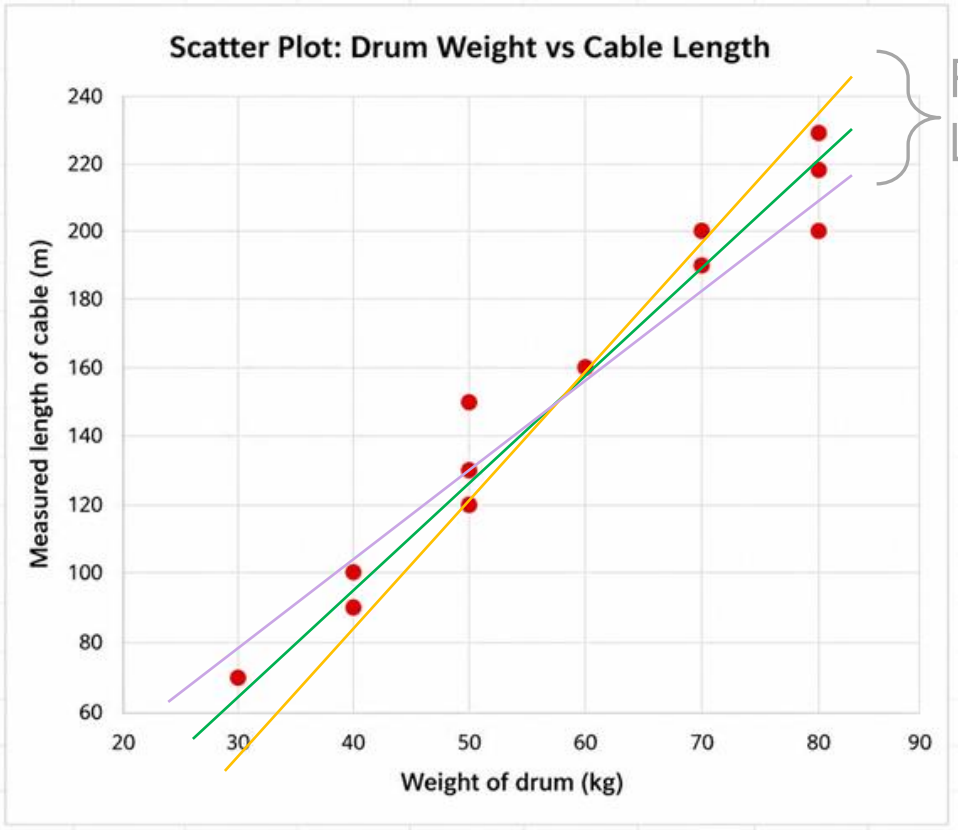


REGRESSION LINES

Different people will draw different lines on a scatterplot....

Regression line

A regression line is a straight line used to model how a **response** variable y changes as an **explanatory** variable x changes.

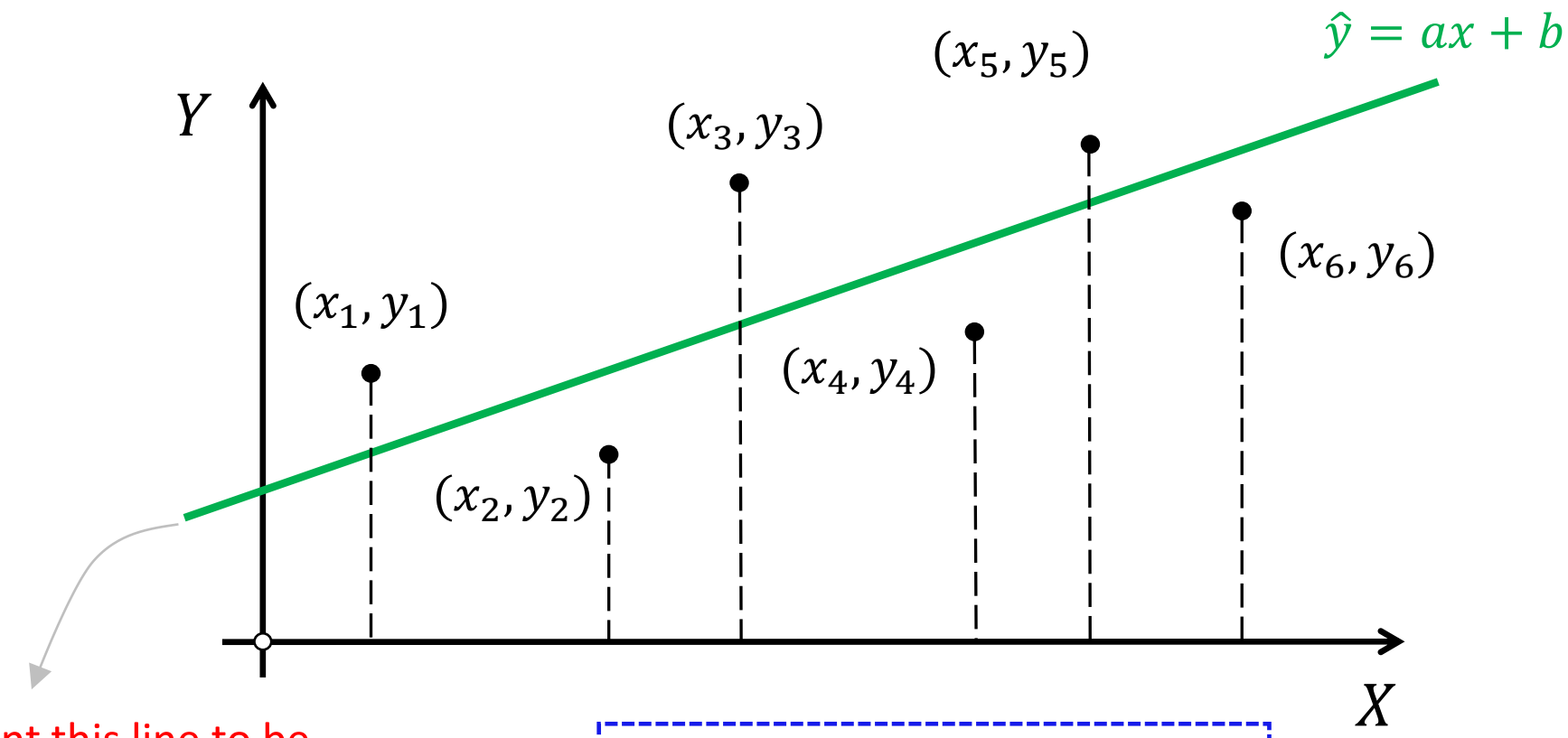


REGRESSION LINES

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.

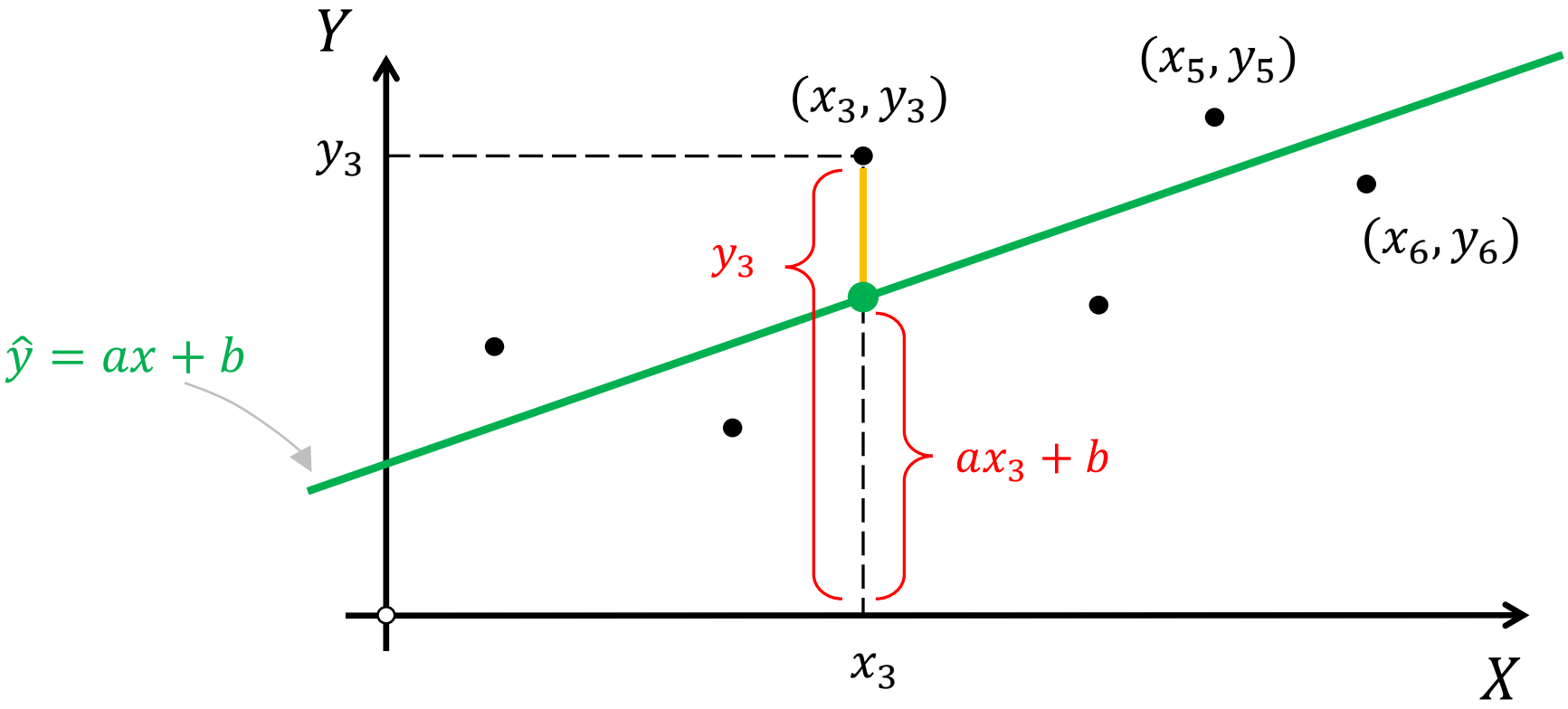
LEAST-SQUARES REGRESSION



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

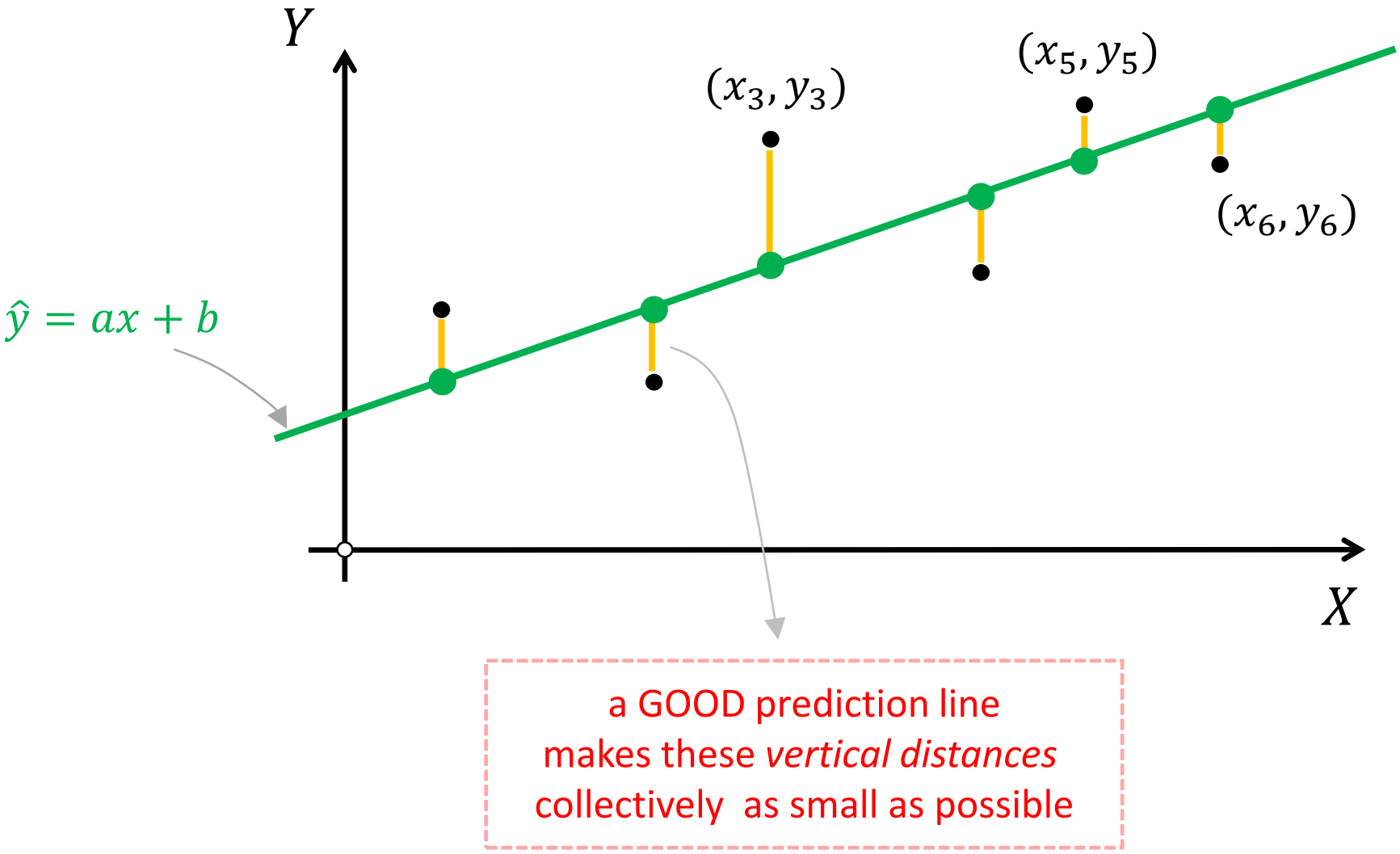
(REGRESSION LINE)

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

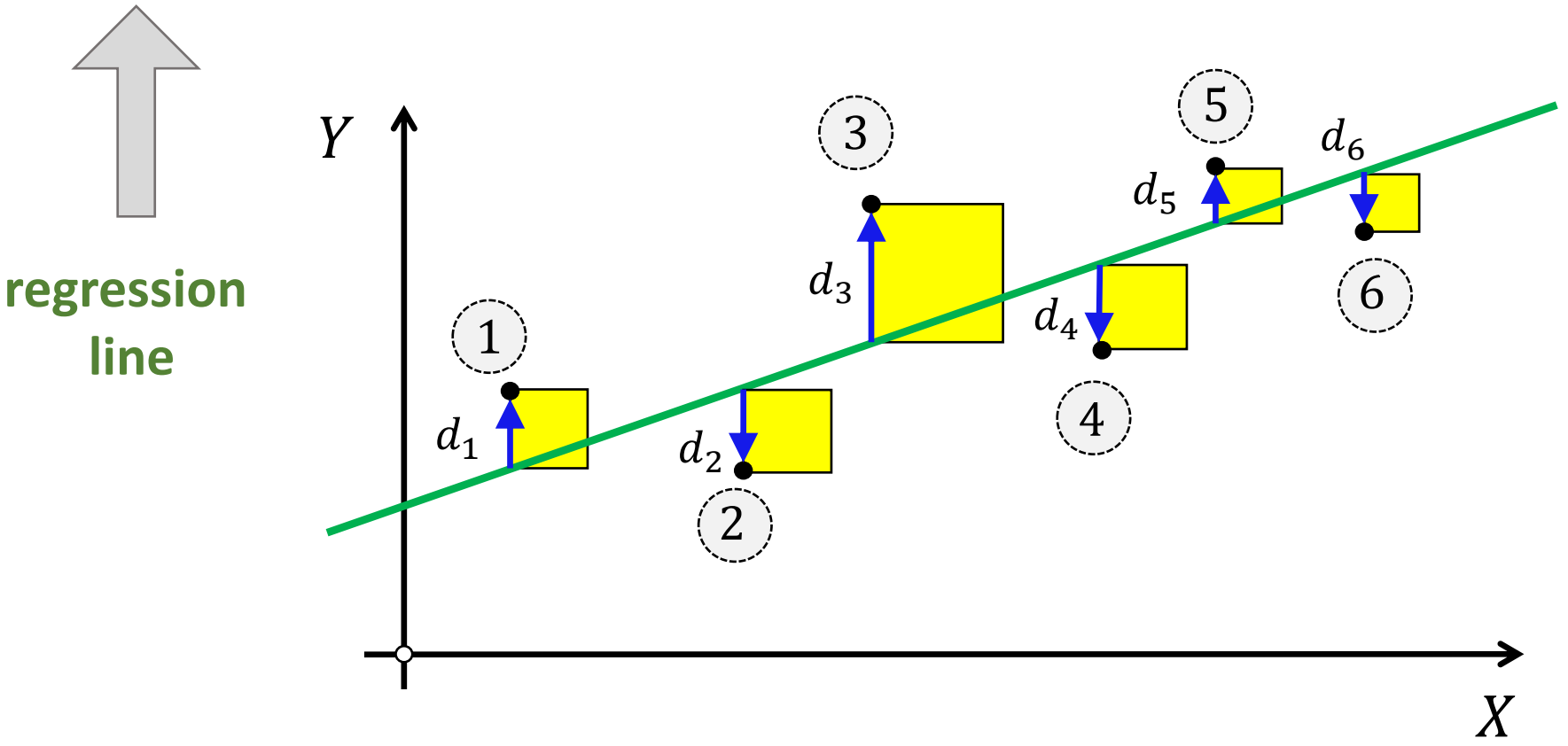


PREDICTION ERROR

$y_3 - (ax_3 + b) =$ *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)
$\hat{y} = ax + b$
X
a GOOD prediction line
makes these *vertical distances*
collectively as small as possible

"BEST FIT" = MINIMIZE THE SUM OF THE SQUARED DISTANCES



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

total area of the YELLOW squares $= (d_1)^2 + (d_2)^2 + \cdots + (d_6)^2$

Least-squares regression line

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

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

average of the x-values

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

average of the y-values

Key formulae

It can be shown that a and b are given by the following formulae:

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

(SLOPE)

$$b = \bar{y} - a\bar{x}$$

(INTERCEPT)

where

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

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

Solution for the example at the start....

$a = 3, b = -20$

$\hat{y} = 3x - 20$

regression line
for the given data

The manager wants **to predict** the lengths of cable left on drums whose weights are:

(i) 35 kg $\quad$ (ii) 85 kg $\quad$ (iii) 100 kg

Weight of drum and cable (x) kg.	Measured length of cable (y) m.
30	70
40	90
40	100
50	120
50	130
50	150
60	160
70	190
70	200
80	200
80	220
80	230

Solution for the example at the start....

Loughborough University

$a = 3, b = -20$

$\hat{y} = 3x - 20$

regression line
for the given data

$x = 35$: $\hat{y}_{35} = 85$

$x = 85$: $\hat{y}_{85} = 235$

$x = 100$: $\hat{y}_{100} = 280$

The manager wants **to predict** the lengths of cable left on drums whose weights are:

(i) 35 kg (ii) 85 kg (iii) 100 kg

Weight of drum and cable (x) kg.	Measured length of cable (y) m.
30	70
40	90
40	100
50	120
50	130
50	150
60	160
70	190
70	200
80	200
80	220
80	230