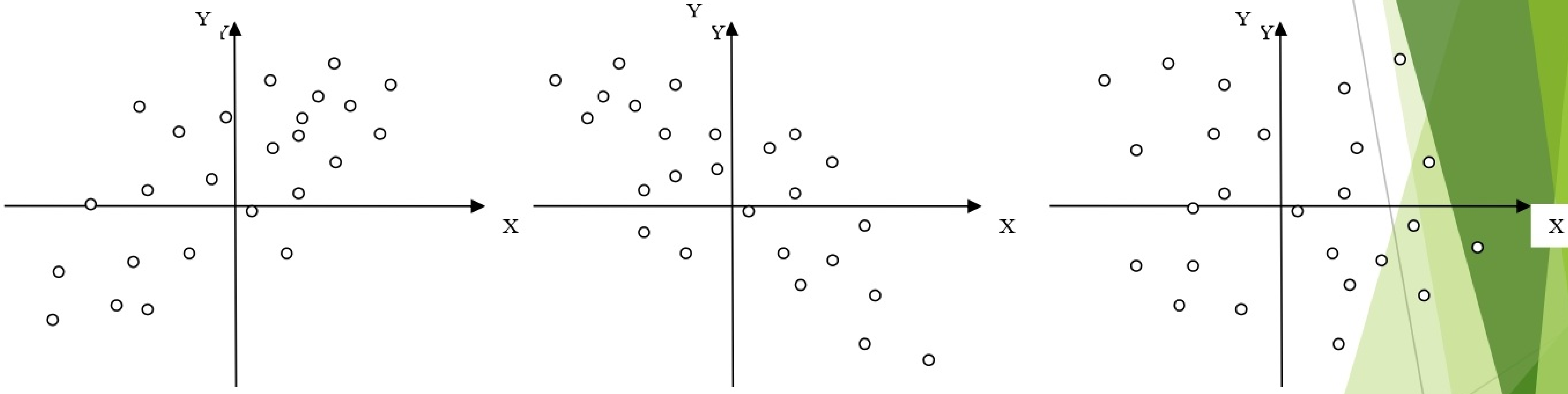
UNIT-V

CORRELATION AND REGRESSION

The relationship between *x* and *y*

- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict dependent variable?
- CORRELATION $\neq$ CAUSATION
  - In order to infer causality: manipulate independent variable and observe effect on dependent variable

Scattergrams



Positive correlation Negative correlation No correlation

Variance vs Covariance

- *First, a note on your sample:*
  - *If you're wishing to assume that your sample is representative of the general population (RANDOM EFFECTS MODEL), use the degrees of freedom (n – 1) in your calculations of variance or covariance.*
  - *But if you're simply wanting to assess your current sample (FIXED EFFECTS MODEL), substitute n for the degrees of freedom.*

Variance vs Covariance

- Do two variables change together?

Variance:

• Gives information on variability of a single variable.

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Covariance:

• Gives information on the degree to which two variables vary together.

• Note how similar the covariance is to variance: the equation simply multiplies x's error scores by y's error scores as opposed to squaring x's error scores.

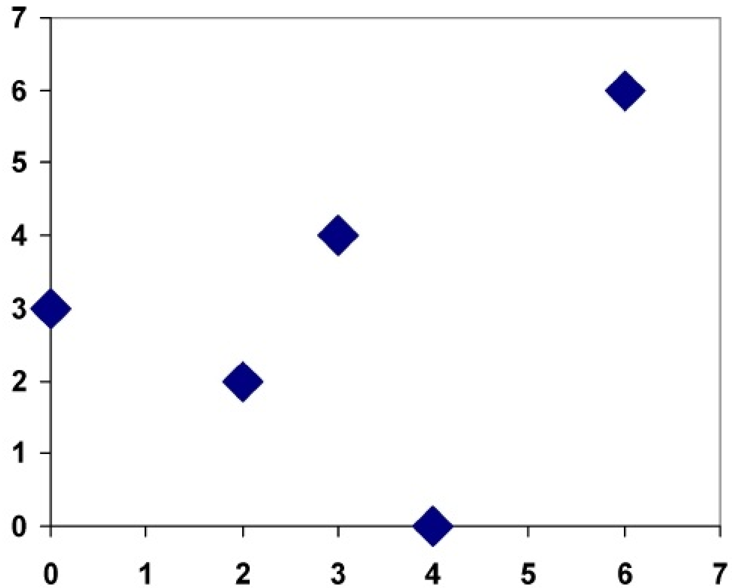
$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariance

$$\mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- When X ↑ and Y ↑: cov (x,y) = pos.
- When X ↓ and Y ↑: cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

Example Covariance



x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0	3	-3	0	0
2	2	-1	-1	1
3	4	0	1	0
4	0	1	-3	-3
6	6	3	3	9
$\bar{x} = 3$	$\bar{y} = 3$			$\sum = 7$

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}))}{n-1} = \frac{7}{4} = 1.75$$

What does this number tell us?

Problem with Covariance:

- The value obtained by covariance is dependent on the size of the data's standard deviations: if large, the value will be greater than if small… *even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets*.

Example of how covariance value relies on variance

	High variance data				Low variance data		
Subject	**x**	**y**	**x error * y error**		**x**	**y**	**X error * y error**
1	101	100	2500		54	53	9
2	81	80	900		53	52	4
3	61	60	100		52	51	1
4	51	50	0		51	50	0
5	41	40	100		50	49	1
6	21	20	900		49	48	4
7	1	0	2500		48	47	9
Mean	51	50			51	50	
Sum of x error * y error :			7000		**Sum of x error * y error :**		28
			1166.67		**Covariance:** 200		**4.67**

Solution: Pearson's r

- **Covariance does not really tell us anything**
    - ***Solution: standardise this measure***
- **Pearson's R: standardises the covariance value.**
- **Divides the covariance by the multiplied standard deviations of X and Y:**

$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}$$

Pearson's R continued

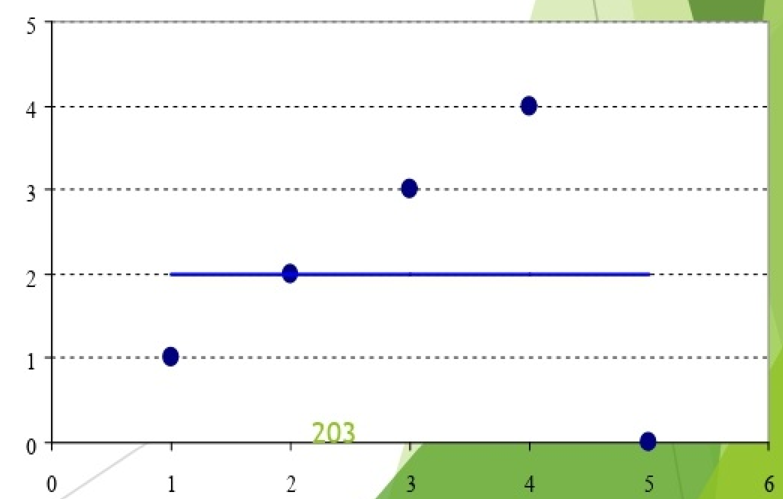
$$\operatorname{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1} \longrightarrow r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{(n-1)s_x s_y}$$

$$\downarrow$$

$$r_{xy}=\frac{\sum_{i=1}^{n}Z_{x_i} * Z_{y_i}}{n-1}$$

Limitations of r

- When r = 1 or r = -1:
  - We can predict y from x with certainty
  - all data points are on a straight line: y = ax + b
- r is actually $\hat{r}$
  - r = true r of whole population
  - $\hat{r}$ = estimate of r based on data
- r is very sensitive to extreme values:



Regression

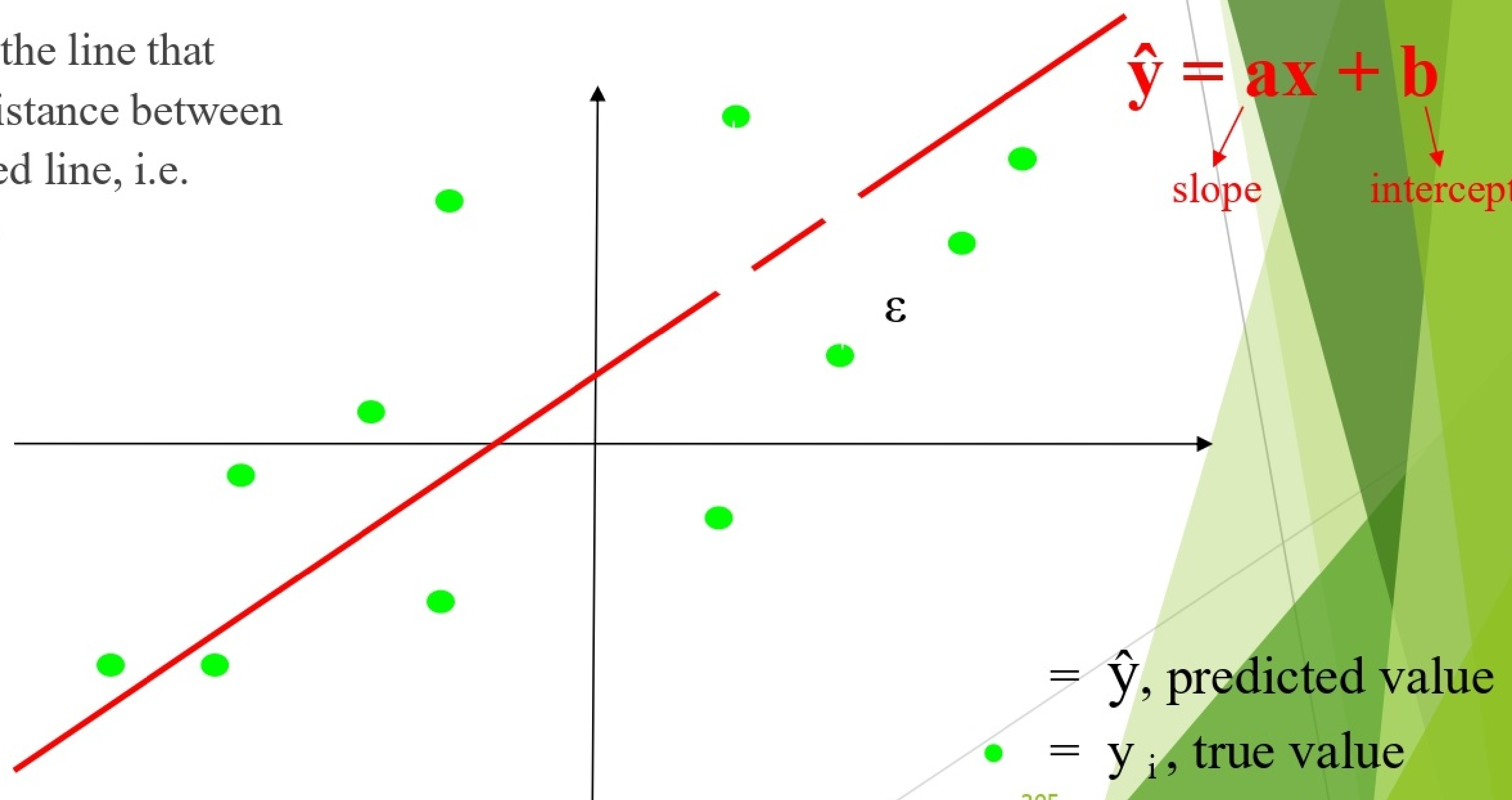
- Correlation tells you if there is an association between x and y but it doesn't describe the relationship or allow you to predict one variable from the other.

- To do this we need REGRESSION!

Best-fit Line

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x

- This will be the line that minimises distance between data and fitted line, i.e. the residuals



= $\hat{y}$, predicted value

= y_i, true value

ε = residual error

Least Squares Regression

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

Model line: $\hat{y} = ax + b$ a = slope, b = intercept

Residual (ε) = $y - \hat{y}$

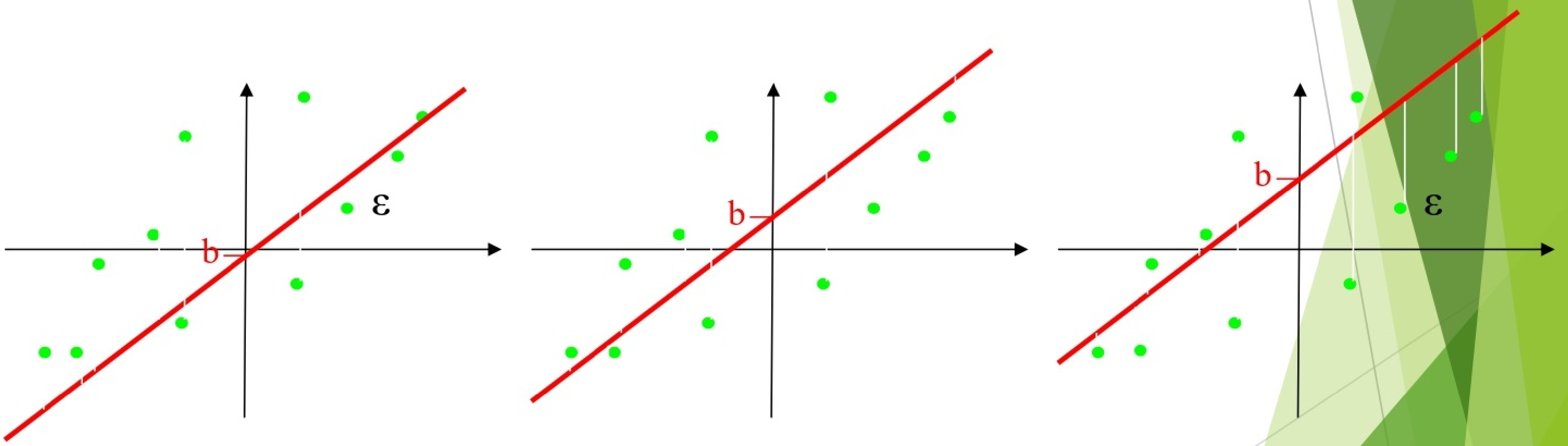
Sum of squares of residuals = $\Sigma (y - \hat{y})^2$

- we must find values of a and b that minimise

$$\Sigma (y - \hat{y})^2$$

Finding b

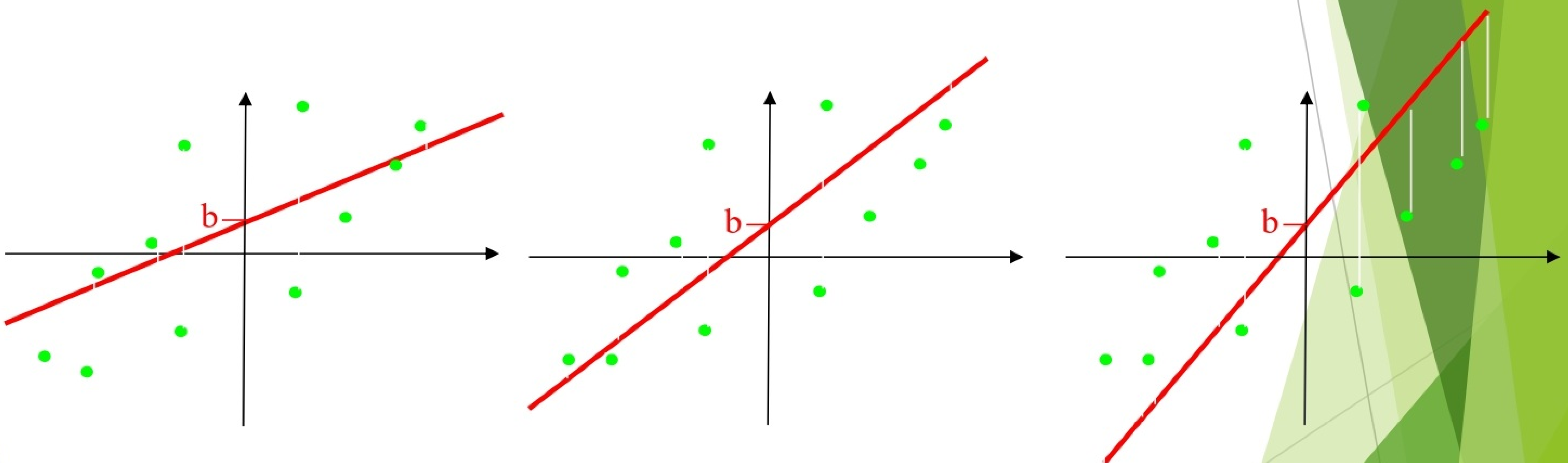
- First we find the value of b that gives the min sum of squares



- Trying different values of b is equivalent to shifting the line up and down the scatter plot

Finding a

- Now we find the value of a that gives the min sum of squares



- Trying out different values of a is equivalent to changing the slope of the line, while b stays constant

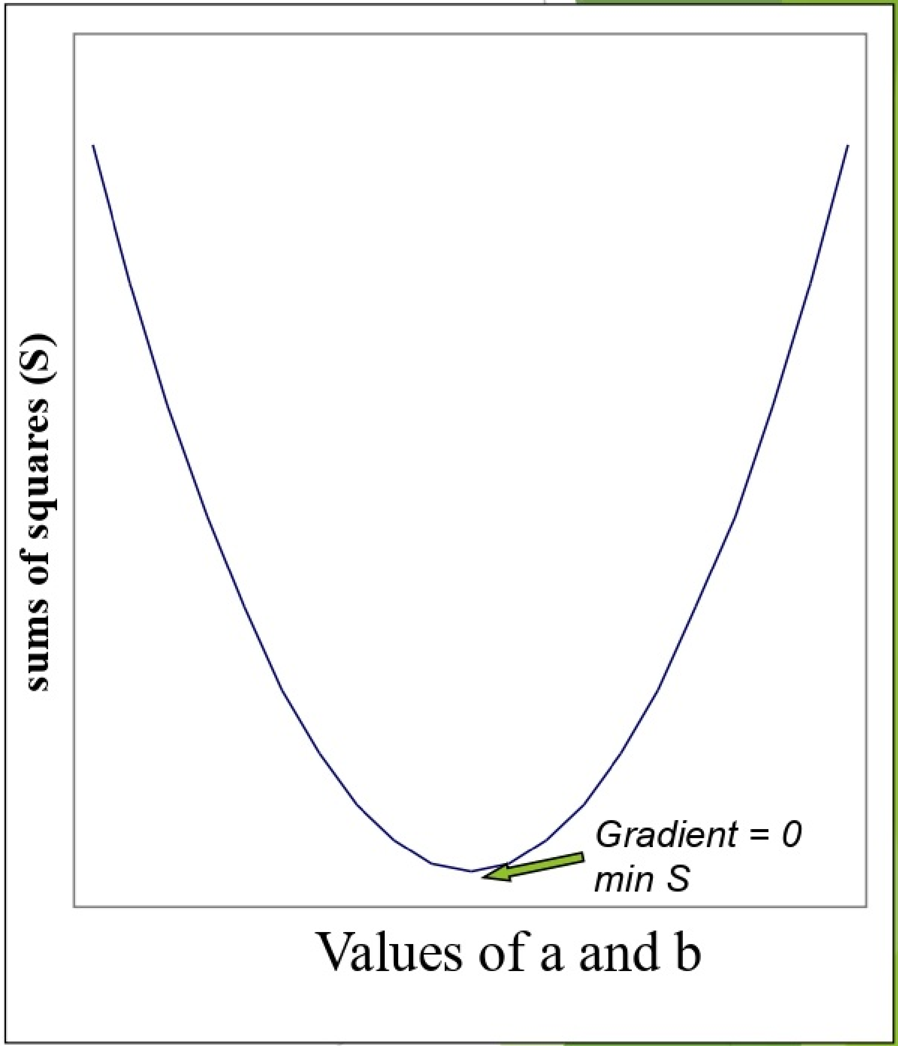
Minimising sums of squares

- Need to minimise $\Sigma(y - \hat{y})^2$
- $\hat{y} = ax + b$
- so need to minimise:

$$\Sigma(y - ax - b)^2$$

- If we plot the sums of squares for all different values of a and b we get a parabola, because it is a squared term

- So the min sum of squares is at the bottom of the curve, where the gradient is zero.



The maths bit

- The min sum of squares is at the bottom of the curve where the gradient = 0

- So we can find a and b that give min sum of squares by taking partial derivatives of $\Sigma(y - ax - b)^2$ with respect to a and b separately

- Then we solve these for 0 to give us the values of a and b that give the min sum of squares

The solution

- Doing this gives the following equations for a and b:

$$a = \frac{r\, s_y}{s_x}$$

r = correlation coefficient of x and y
s_y = standard deviation of y
s_x = standard deviation of x

- From you can see that:
  - A low correlation coefficient gives a flatter slope (small value of a)
  - Large spread of y, i.e. high standard deviation, results in a steeper slope (high value of a)
  - Large spread of x, i.e. high standard deviation, results in a flatter slope (high value of a)