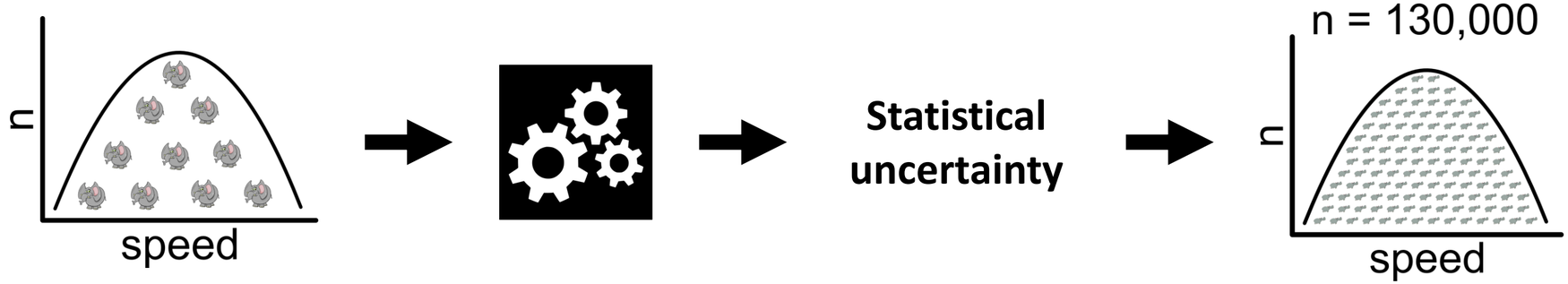
Regression

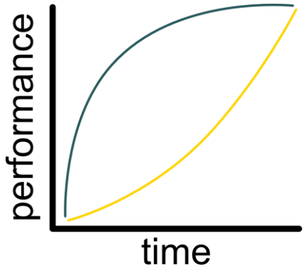
Pärt Prommik, PhD

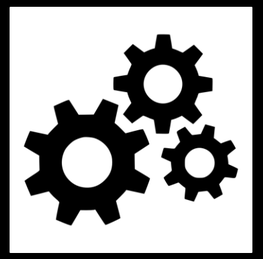
Ülo Maiväli, PhD

Reminder

id	y	x
a	0	2
b	1	1
c	2	0

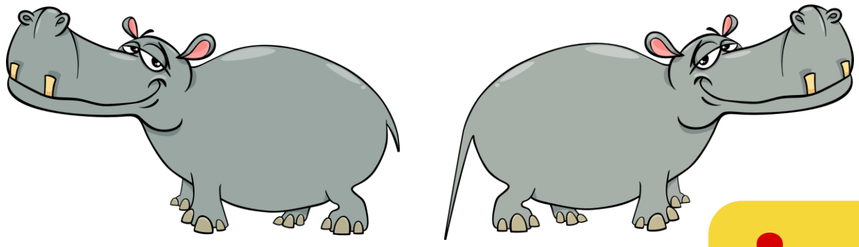






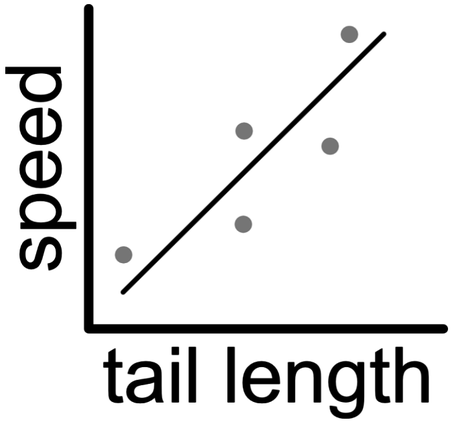

b~r+(m|s)

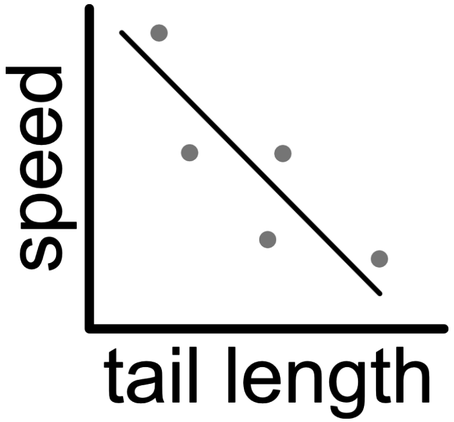
Analysing relationships

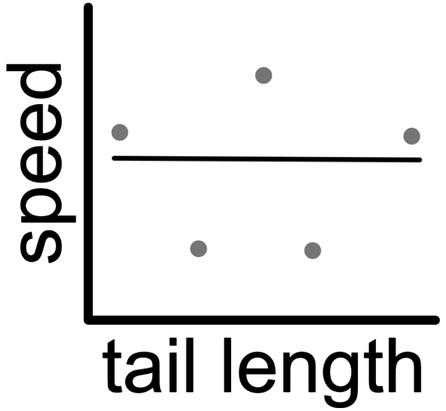


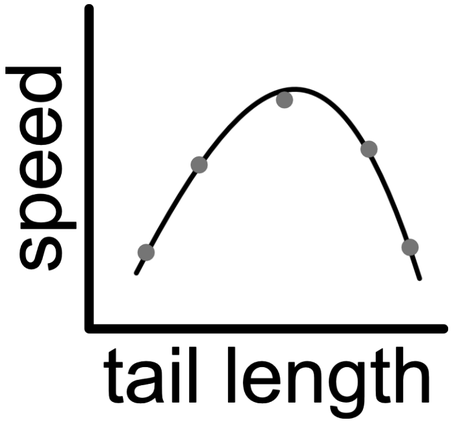
speed	tail length
20	10
31	32
50	17
28	23
44	39

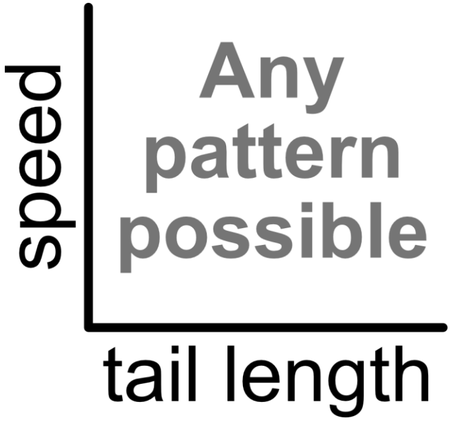
! Lines are good for expressing relationships











Line drawing task

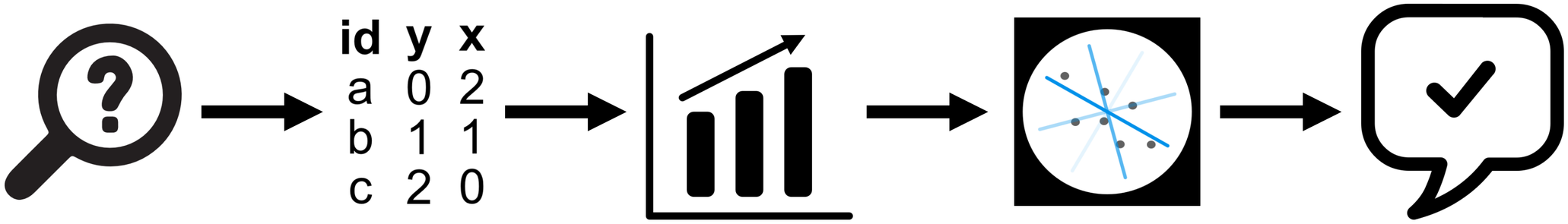
I’ll draw few data points.

You come and draw a **STRAIGHT LINE** that fits the data points:

it must be **AS CLOSE AS POSSIBLE** to all points.

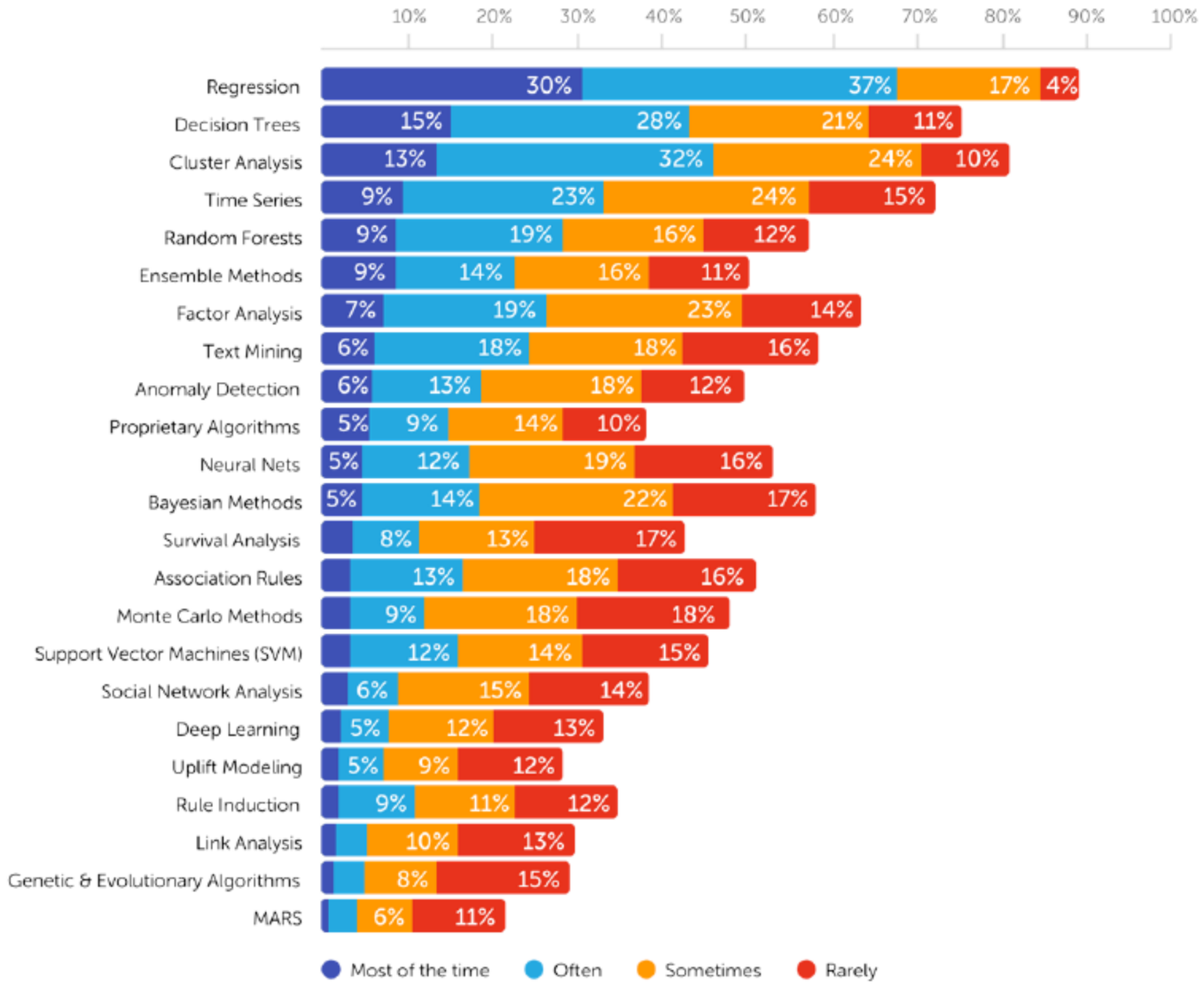
You just did a **regression**

Regression is a statistical method that estimates relationship between two or more variables by finding a line of best fit.



Why regression?

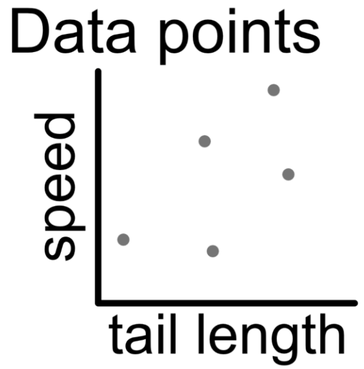
- The most used technique
- Powerful
- Flexible

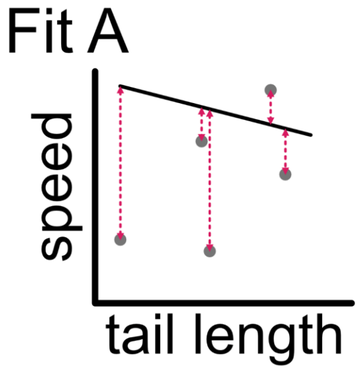


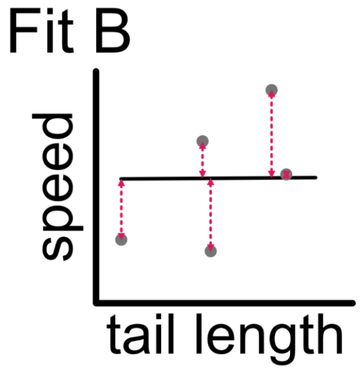
https://www.rexeranalytics.com/data-science-survey

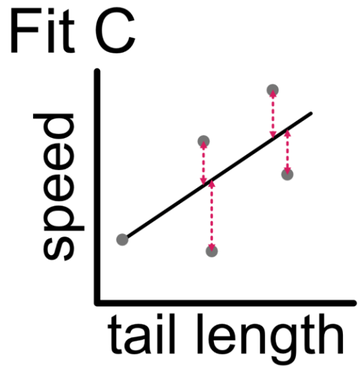
How the best fit line is found?

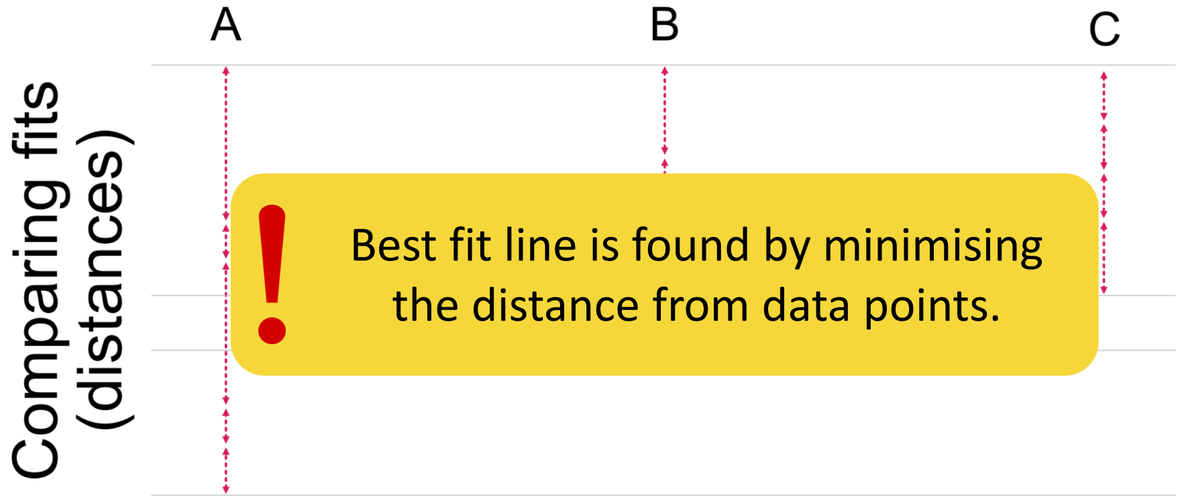
Residual: the difference between the observed value and the estimated value.





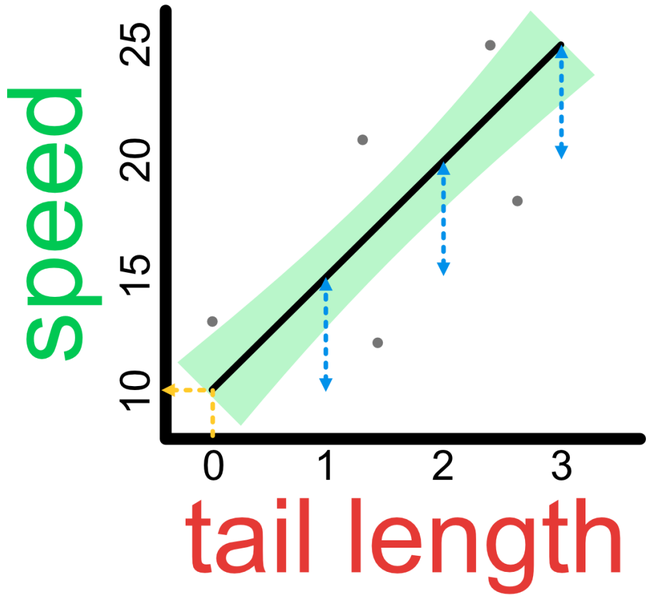






Best fit line is found by minimising the distance from data points.

A line can be described mathematically as an **equation**



Coefficients

Y
Predicted output
Outcome variable
Response variable
Dependent variable

Intercept
Constant

Slope

X
Explanatory variable
Predictor variable
Independent variable

Residual standard deviation
Residual standard error
Error
Disturbance

$$speed = a + b \times tail\_length + e$$

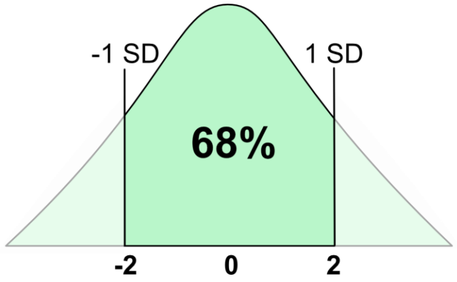
1. **Intercept:** starting point
2. **Slope:** inclination
3. **Residual SD:** accuracy of relationship

Equations are good for answering different study questions

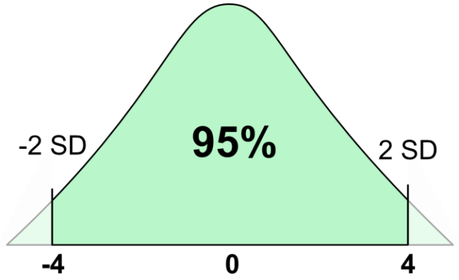
tail length

$$20 = 10 + 5 \times \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 5 \end{matrix} + 2$$

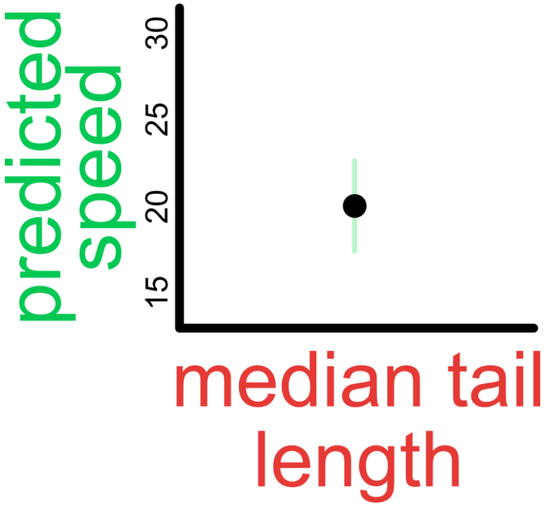
Descriptive



Lower confidence limit: 20-2 = **18**
Upper confidence limit: 20+2 = **22**



Lower confidence limit: 20-4 = **16**
Upper confidence limit: 20+4 = **24**

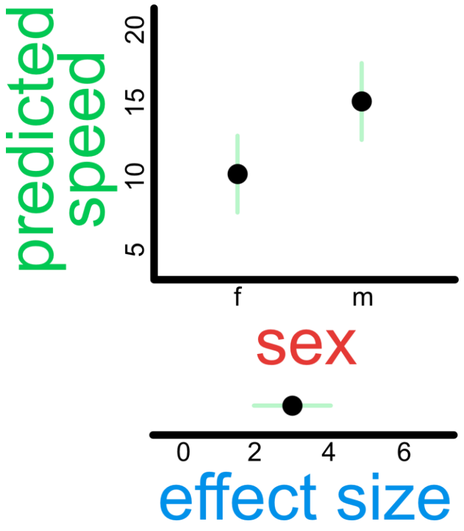


Equations are good for answering different study questions

Comparative

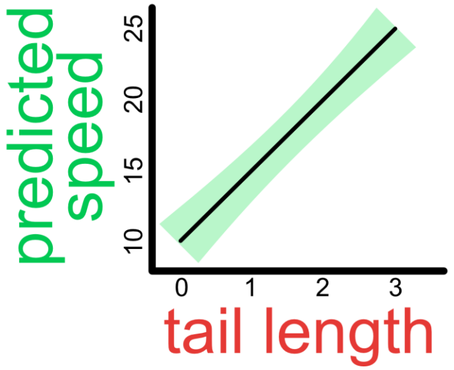
$$10 = 10 + 5 \times \begin{cases} f \\ m \end{cases} + 2 \qquad \text{sex}$$

$$15 = 10 + 5 \times \begin{cases} f \\ m \end{cases} + 2 \qquad \text{sex}$$

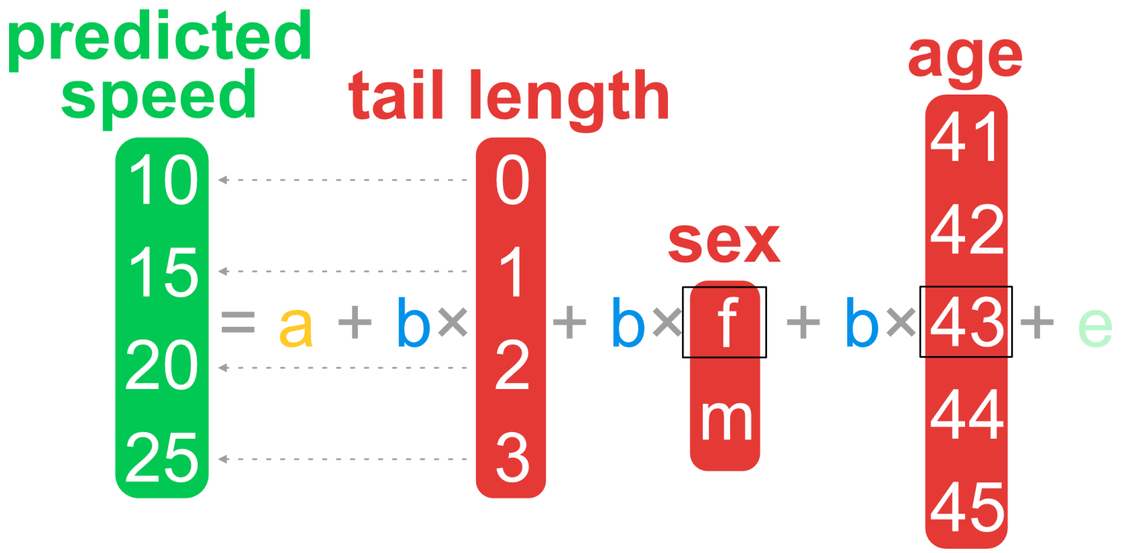


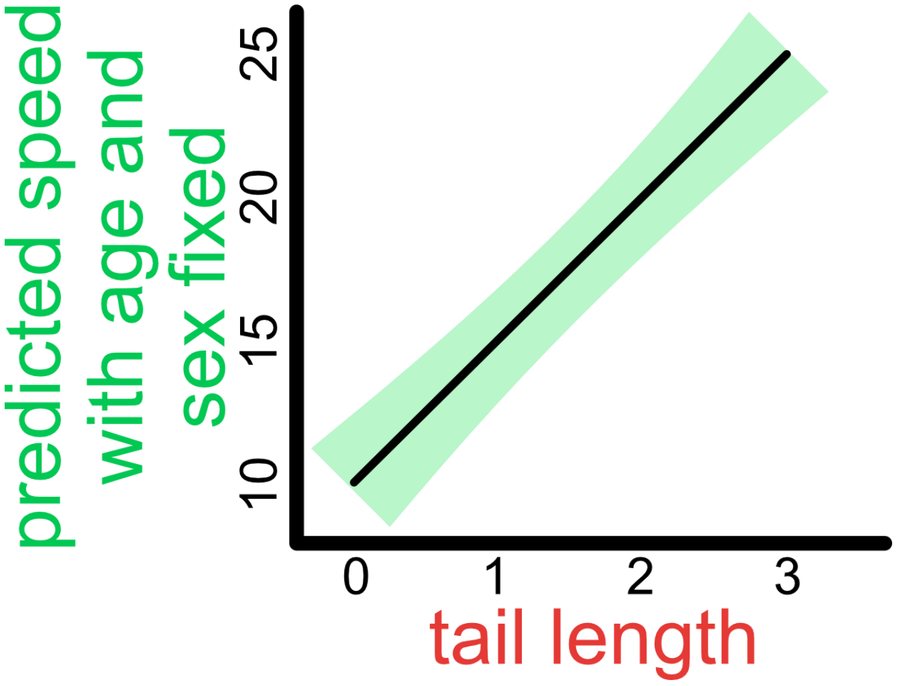
Relationship

$$\begin{matrix} \text{predicted speed} \\ \begin{cases} 10 \\ 15 \\ 20 \\ 25 \end{cases} \end{matrix} = 10 + 5 \times \begin{matrix} \text{tail length} \\ \begin{cases} 0 \\ 1 \\ 2 \\ 3 \end{cases} \end{matrix} + 2$$



Multiple variables are not a problem



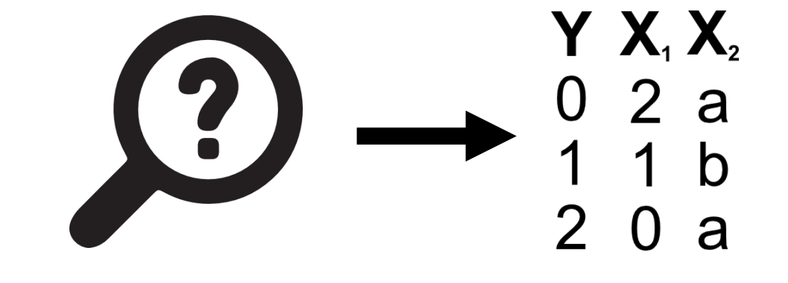


! Solves many kind of study questions!

Regression equations allow analysing the world

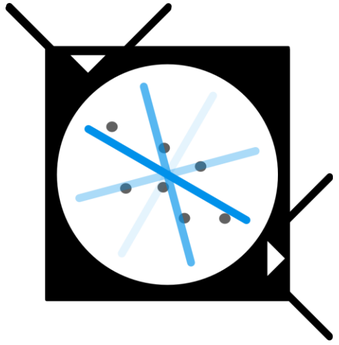
https://pixfeeds.com/images/2/180821/1200-95683243-plant-growing-in-sunlight.jpg

Line fitting machine’s input and output

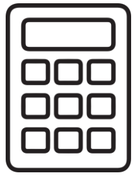


Y	X_1	X_2
0	2	a
1	1	b
2	0	a

Y ~ X



Y ~ **a** + **b**X + **e**



! Regression finds the line of best fit and calculates coefficients for our input formula and residual standard deviation.

Regression in R language

	speed	tail_length
1	20	15
2	25	12
3	32	18
4	37	14
5	34	24
6	39	22

Use variable names and tilde for determining formula

speed ~ tail_length

Use the formula inside a regrssion function

function_name(speed ~ tail_length, data = yourdata)