Introduction to Linear Modelling

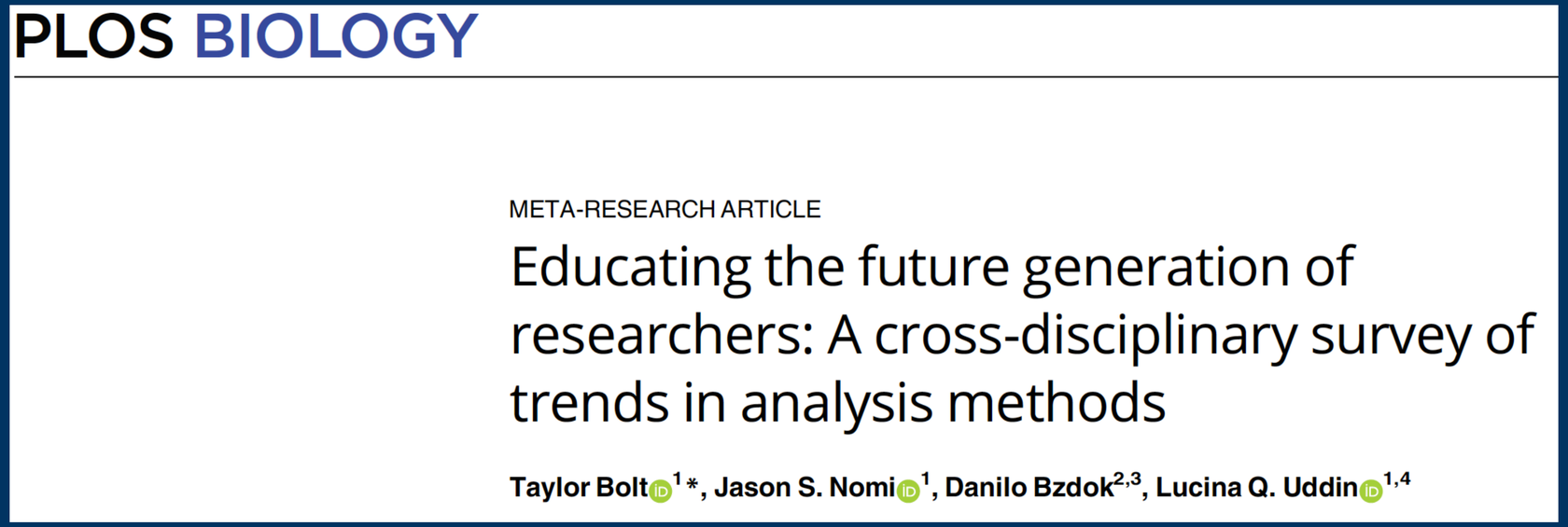
NGSchool 2022

Nuno Sepúlveda, 16.09.2022
nuno.Sepulveda@mini.pw.edu.pl
http://www.immune-stats.net

Your thoughts about linear regression

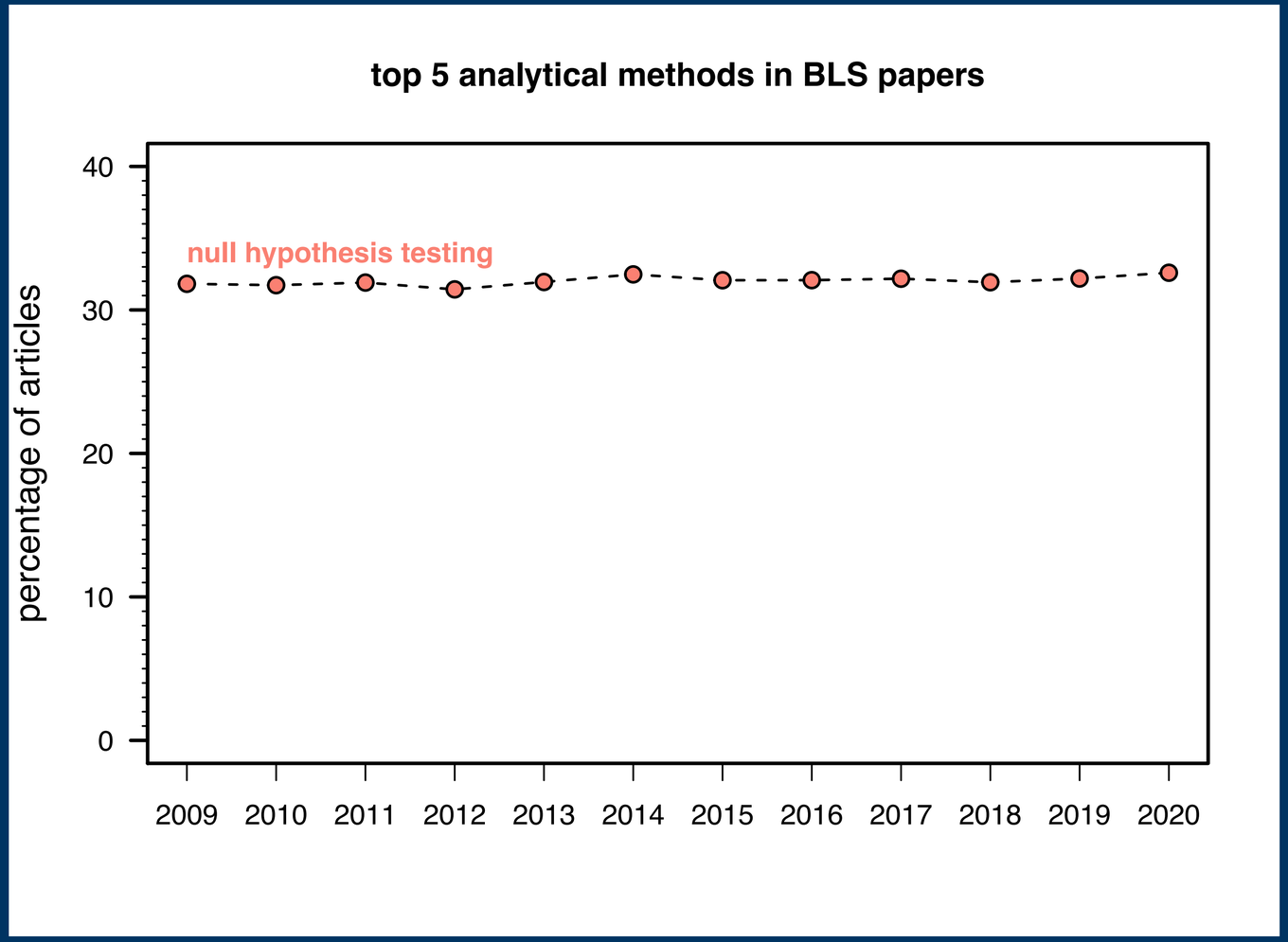
Why learning linear regression?

Why learning linear regression?

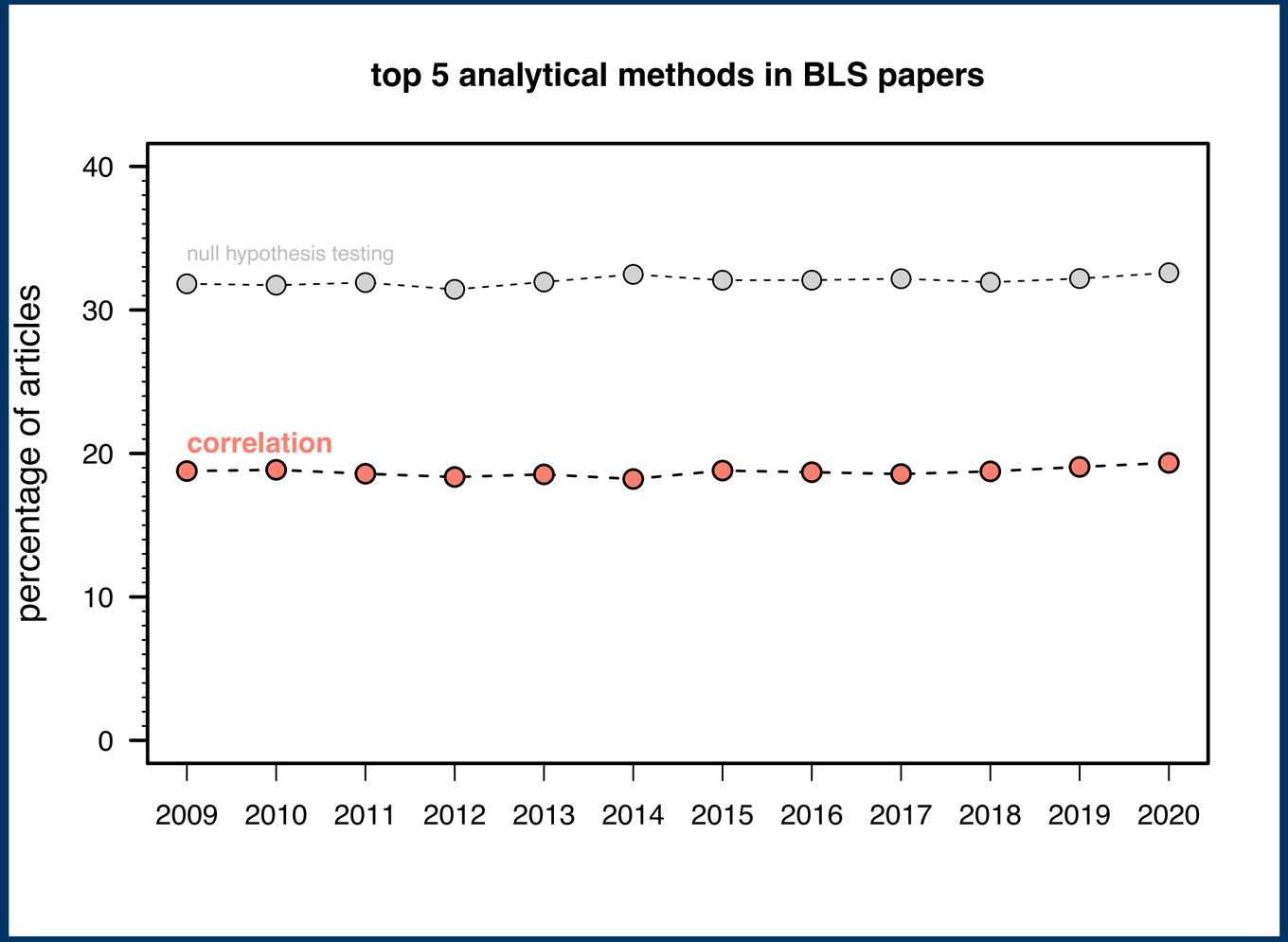


Bolt T, Nomi JS, Bzdok D, Uddin LQ (2021) Educating the future generation of researchers: A cross-disciplinary survey of trends in analysis methods. PLoS Biol 19(7): e3001313.

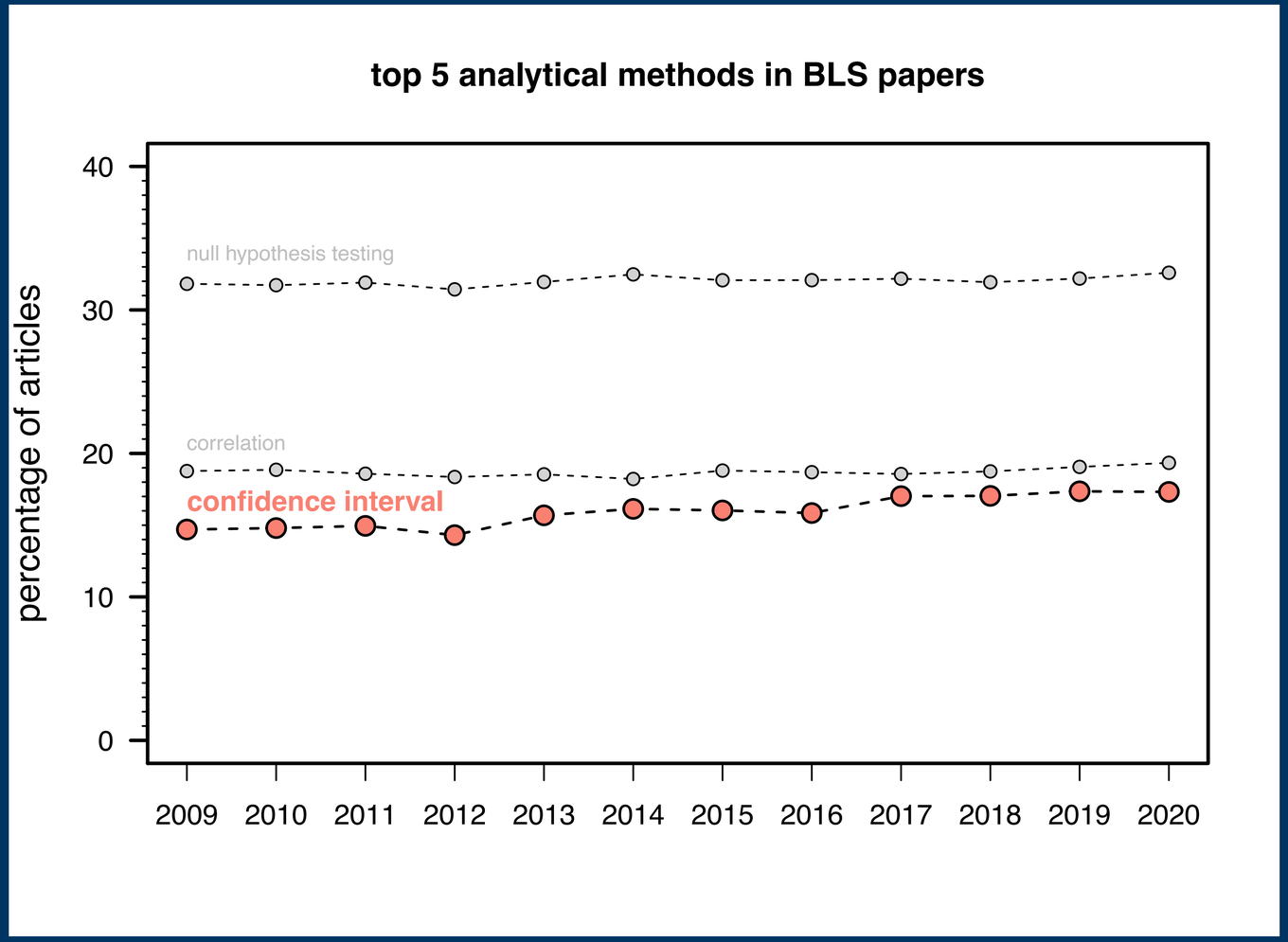
Why learning linear regression?



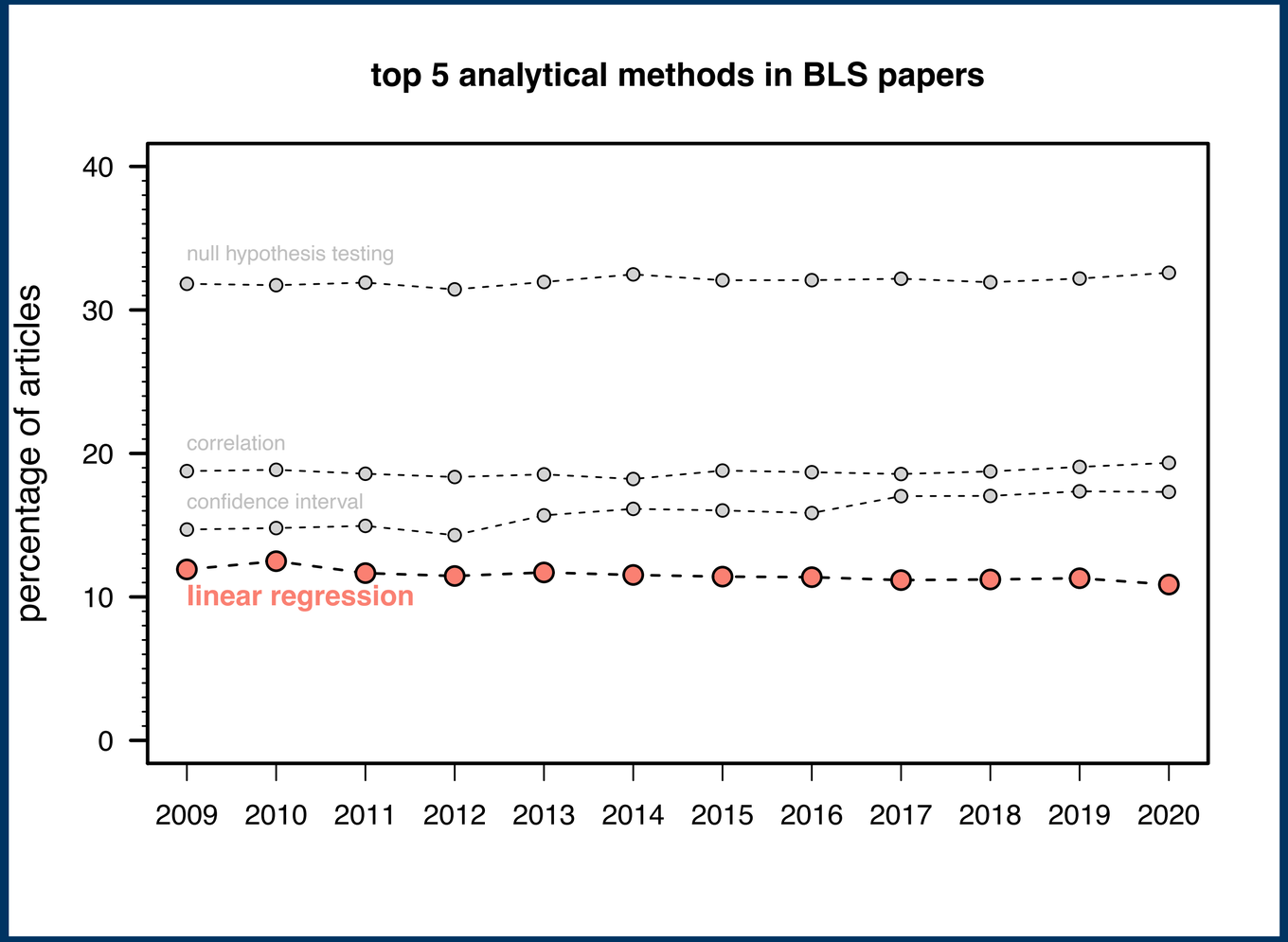
Why learning linear regression?



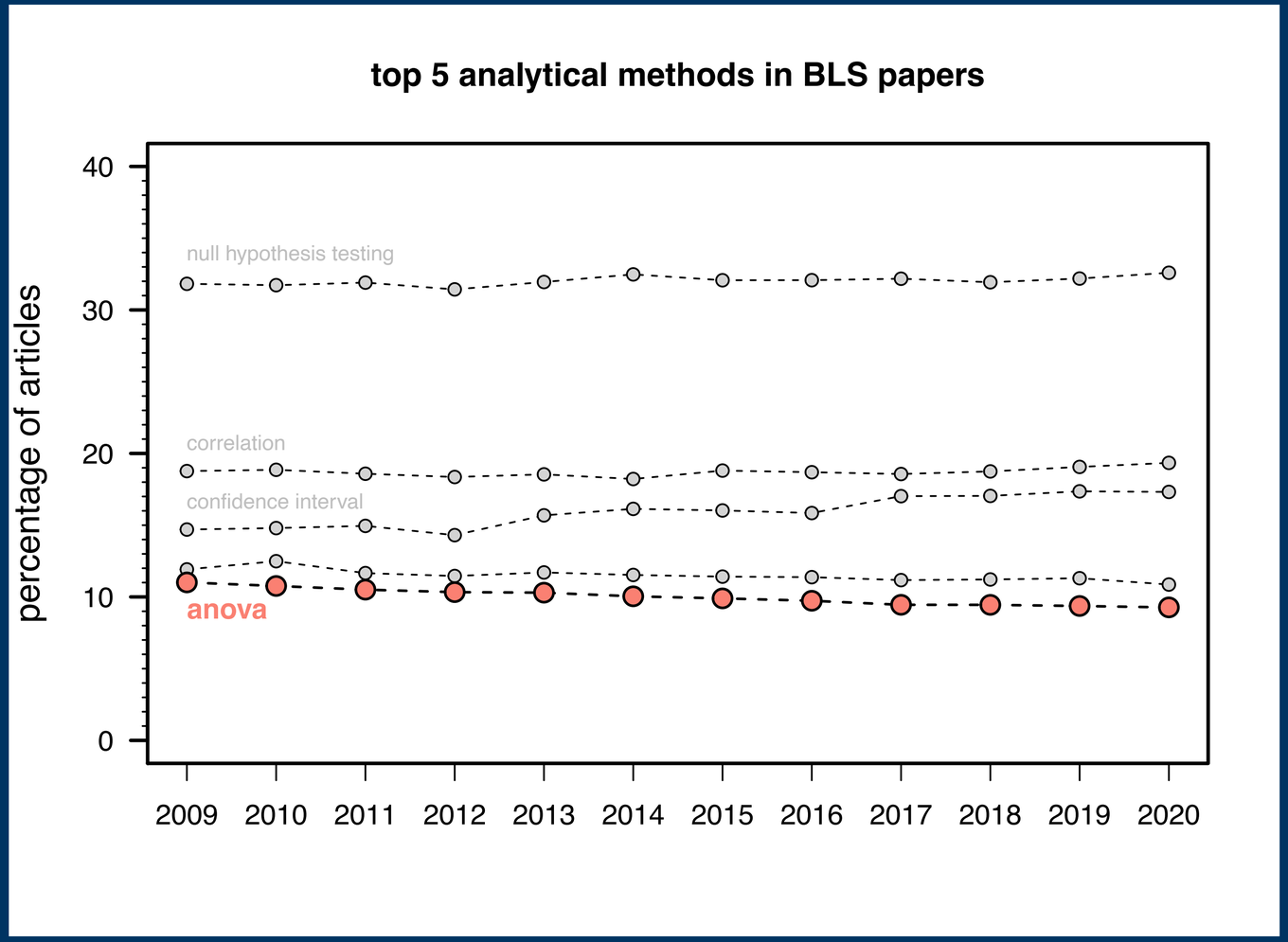
Why learning linear regression?



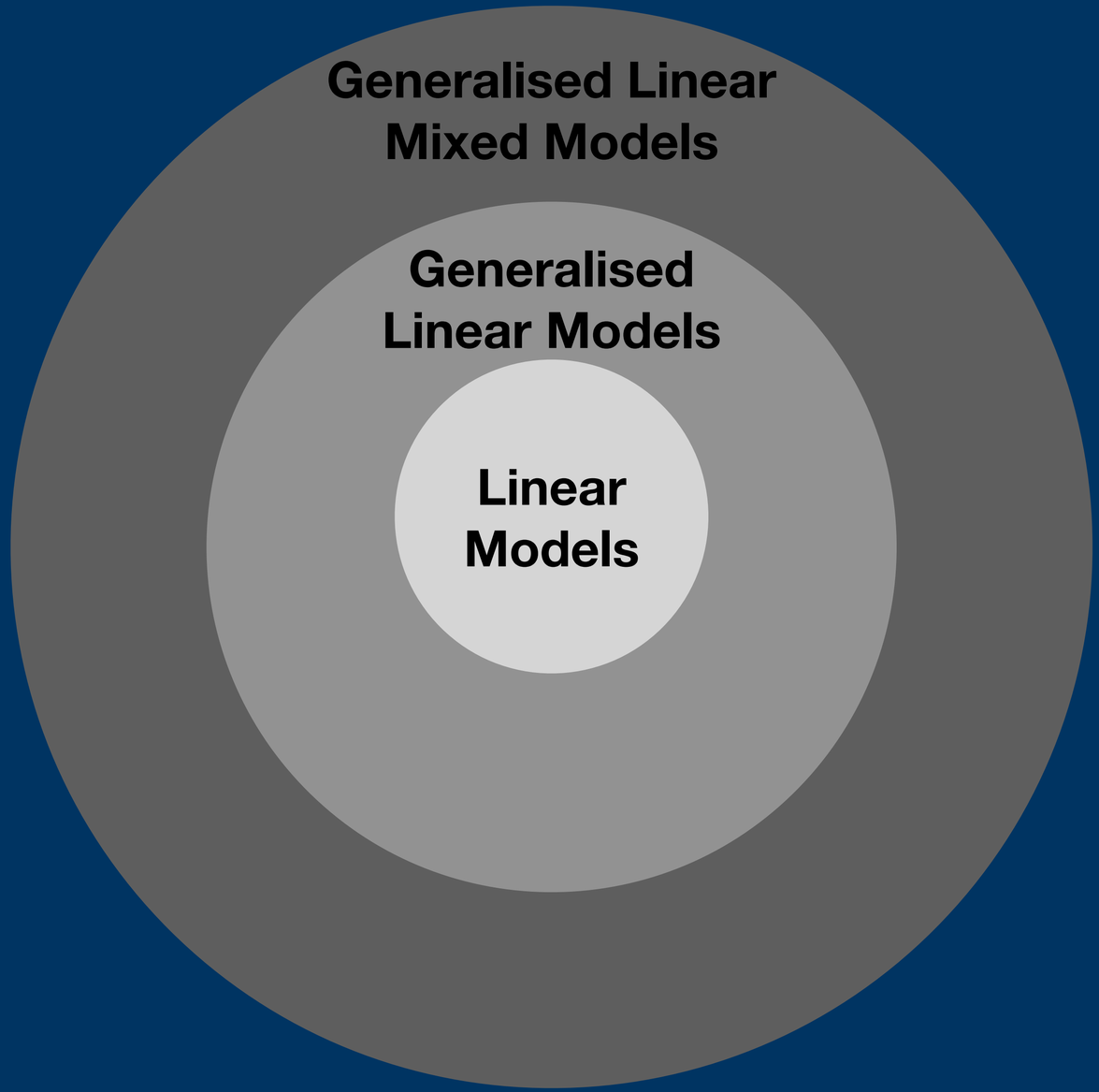
Why learning linear regression?



Why learning linear regression?



Linear regression as a foundation for more complex models



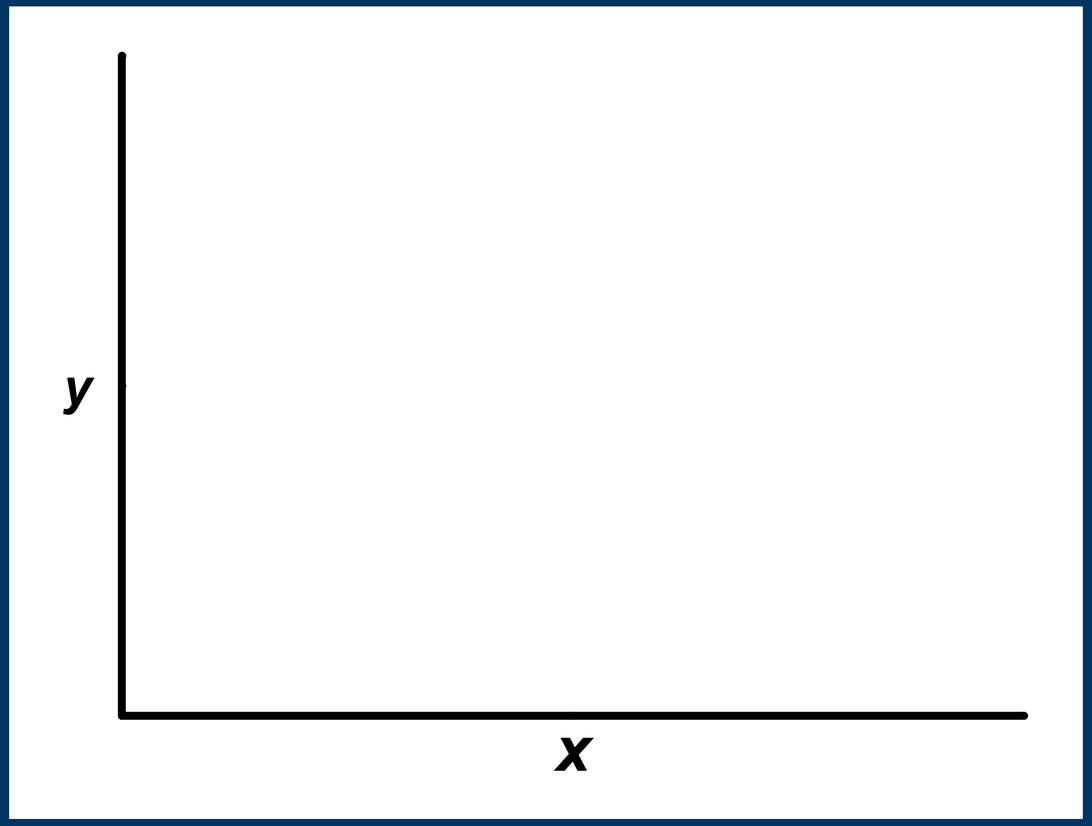
Objectives

Touch-base on simple linear regression

- How to estimate parameters?
- How to interpret the results?
- How to be cautious in reporting?

Use R to conduct data analysis

Study of the relationship between x and y



Simple linear regression

$$y = a + bx$$

$y =$ response variable

dependent variable

outcome variable

$x =$ explicative variable

independent variable

covariate/predictor

feature

Simple linear regression

$$y = a + bx$$

a = intercept

b = slope

Simple linear regression

$$y = a + bx$$

a = intercept

b = slope

Footnote

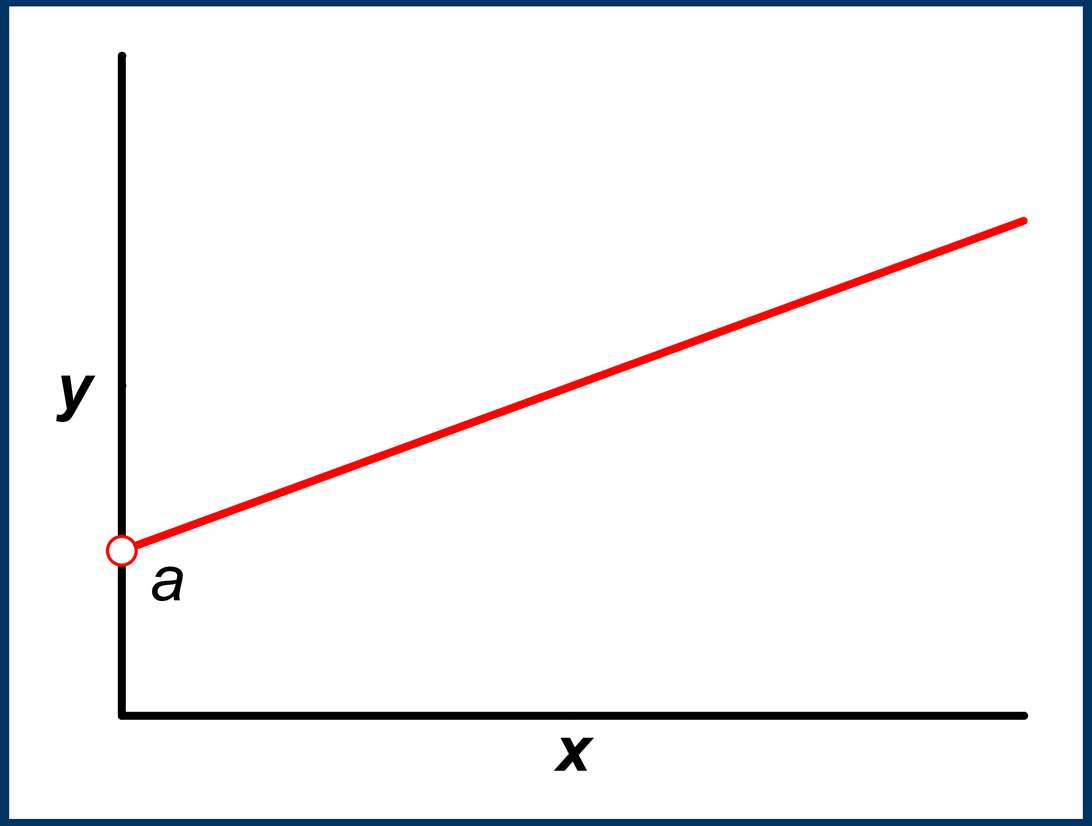
$$y = \alpha + \beta x$$

$$y = \beta_0 + \beta_1 x$$

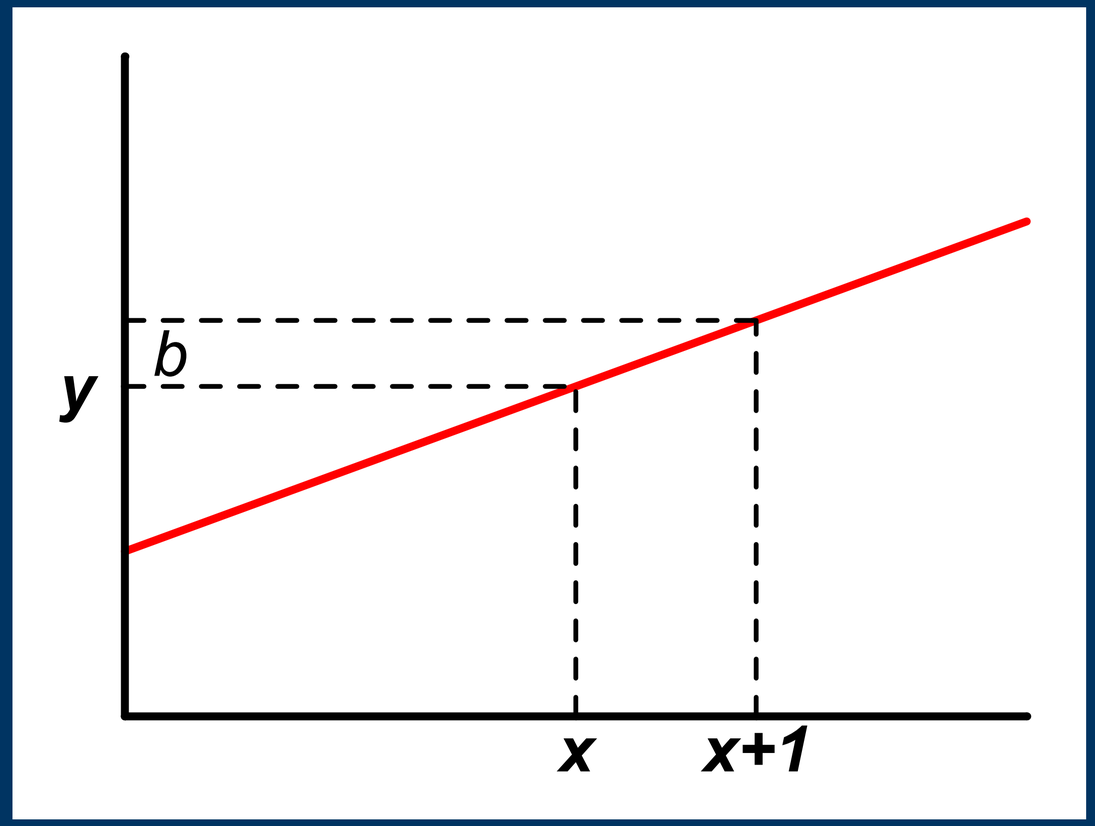
Statisticians/Mathematicians

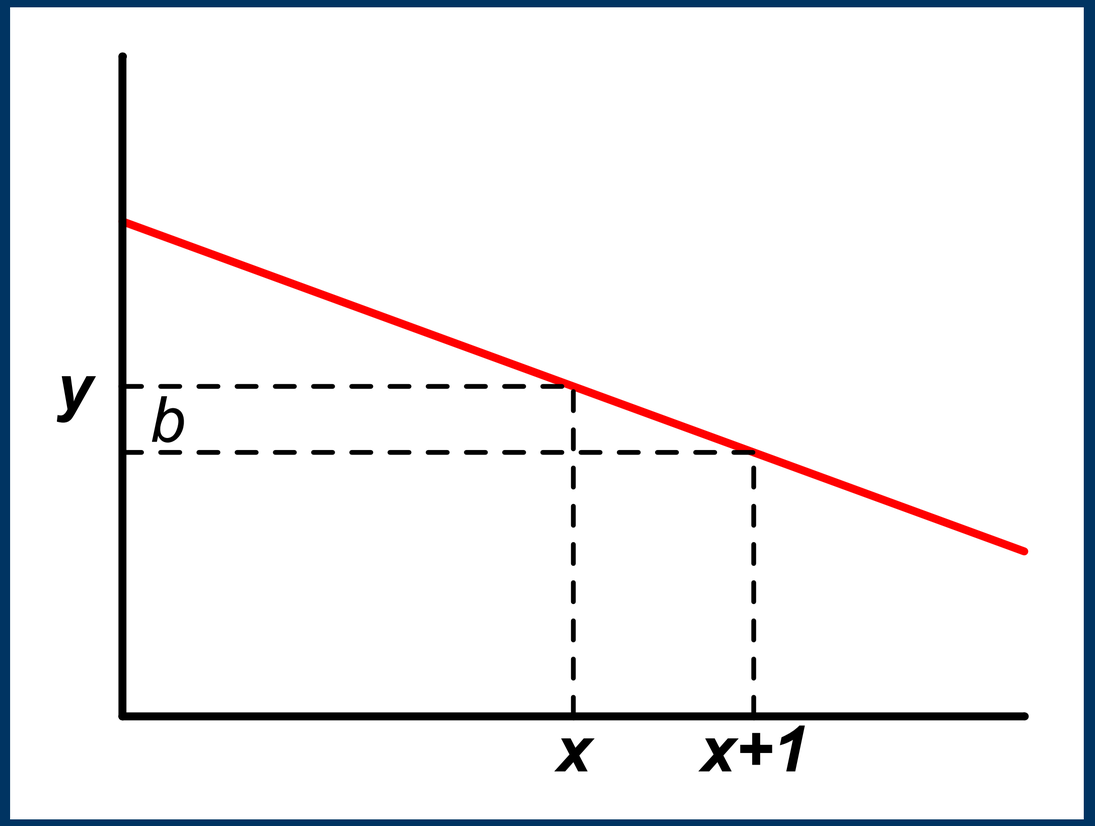
What is the advantage of writing the equation like this?

What does the intercept represent?

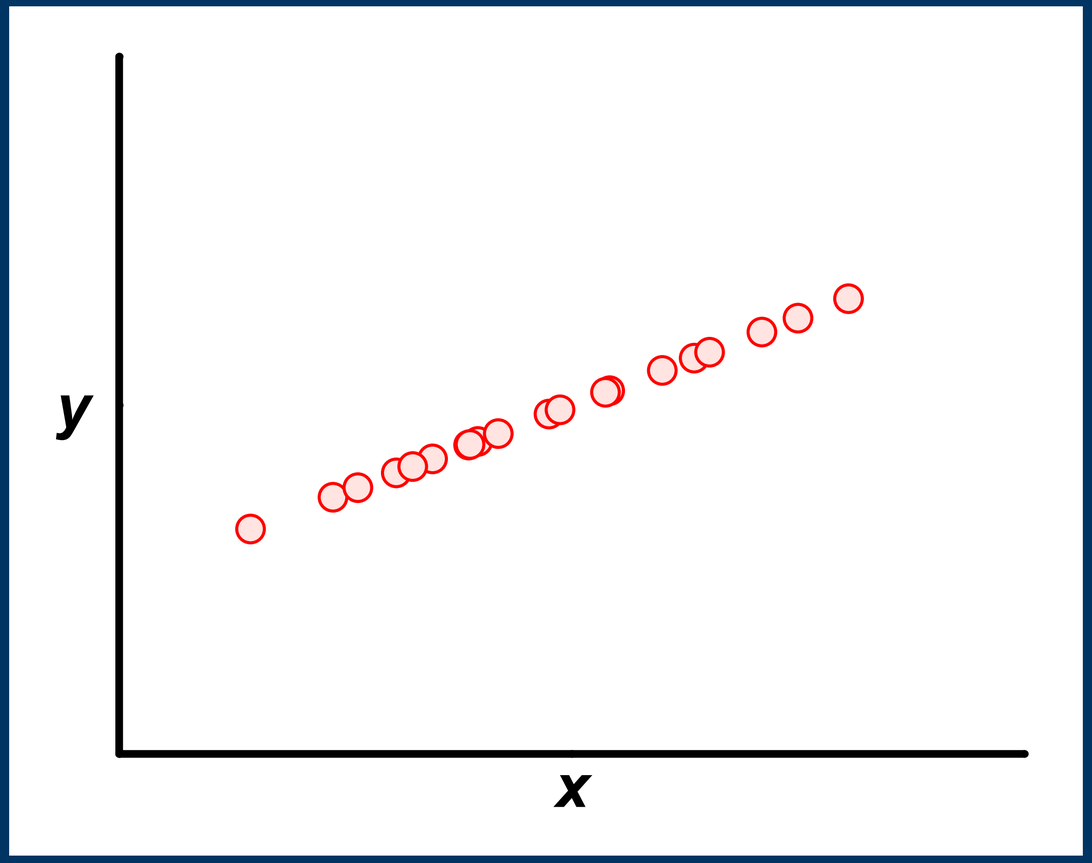


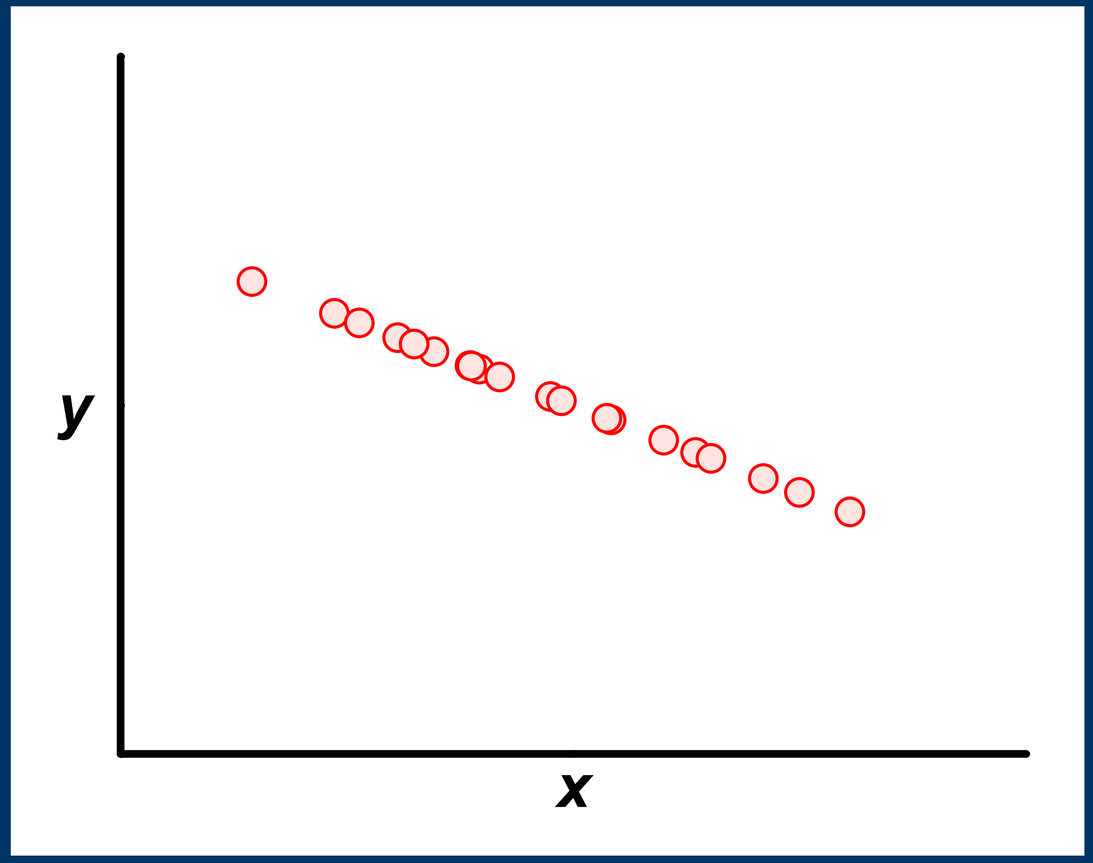
What does the slope represent?



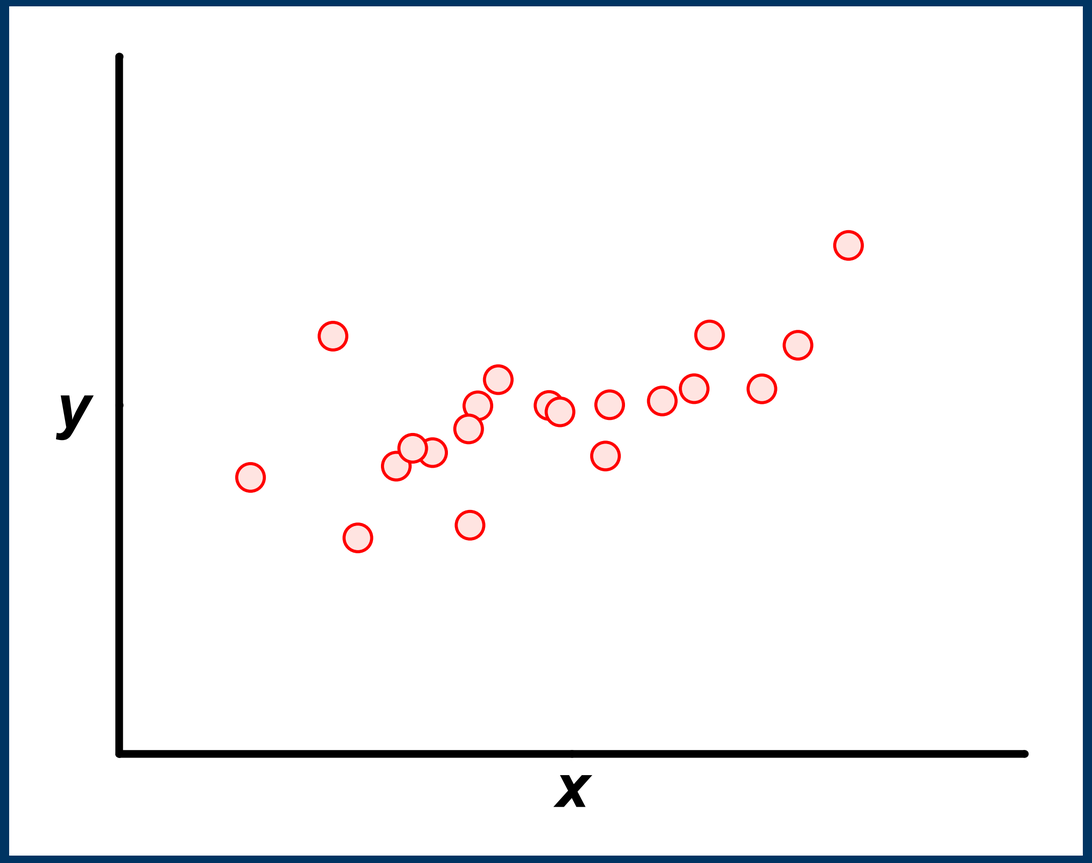


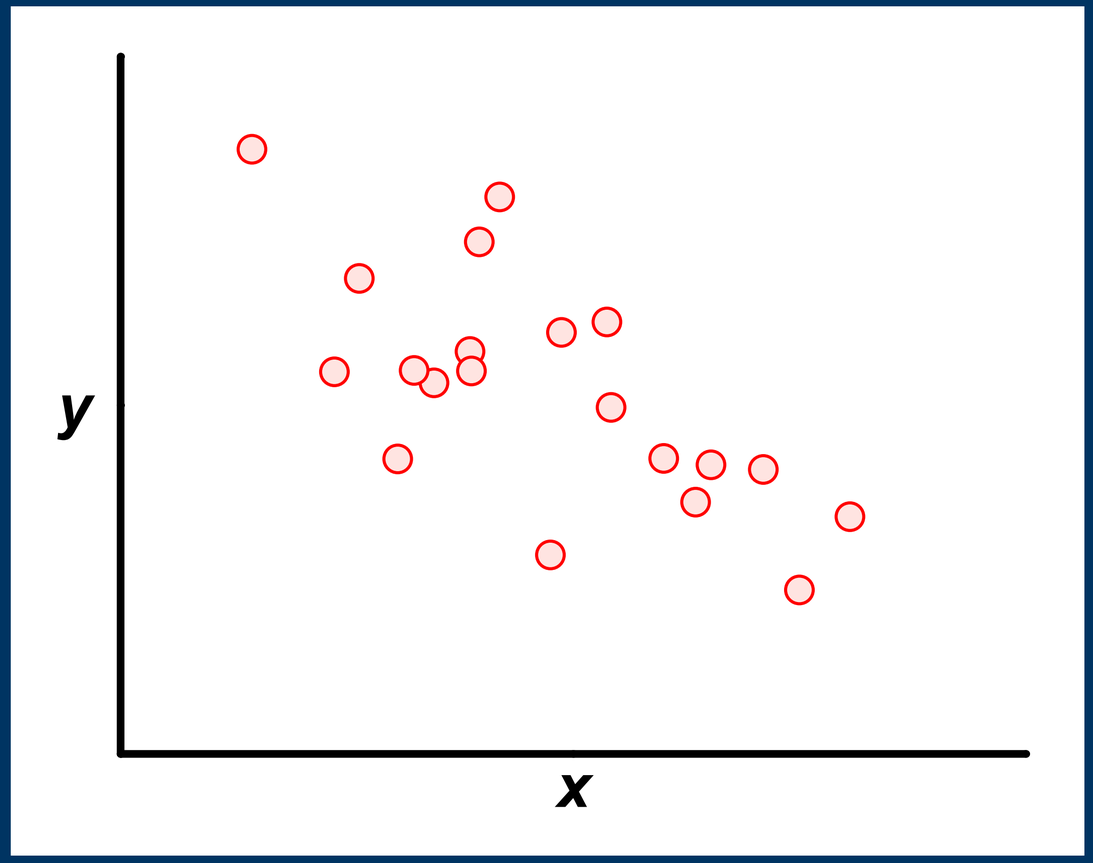
In a perfect world





In a not-so-perfect world





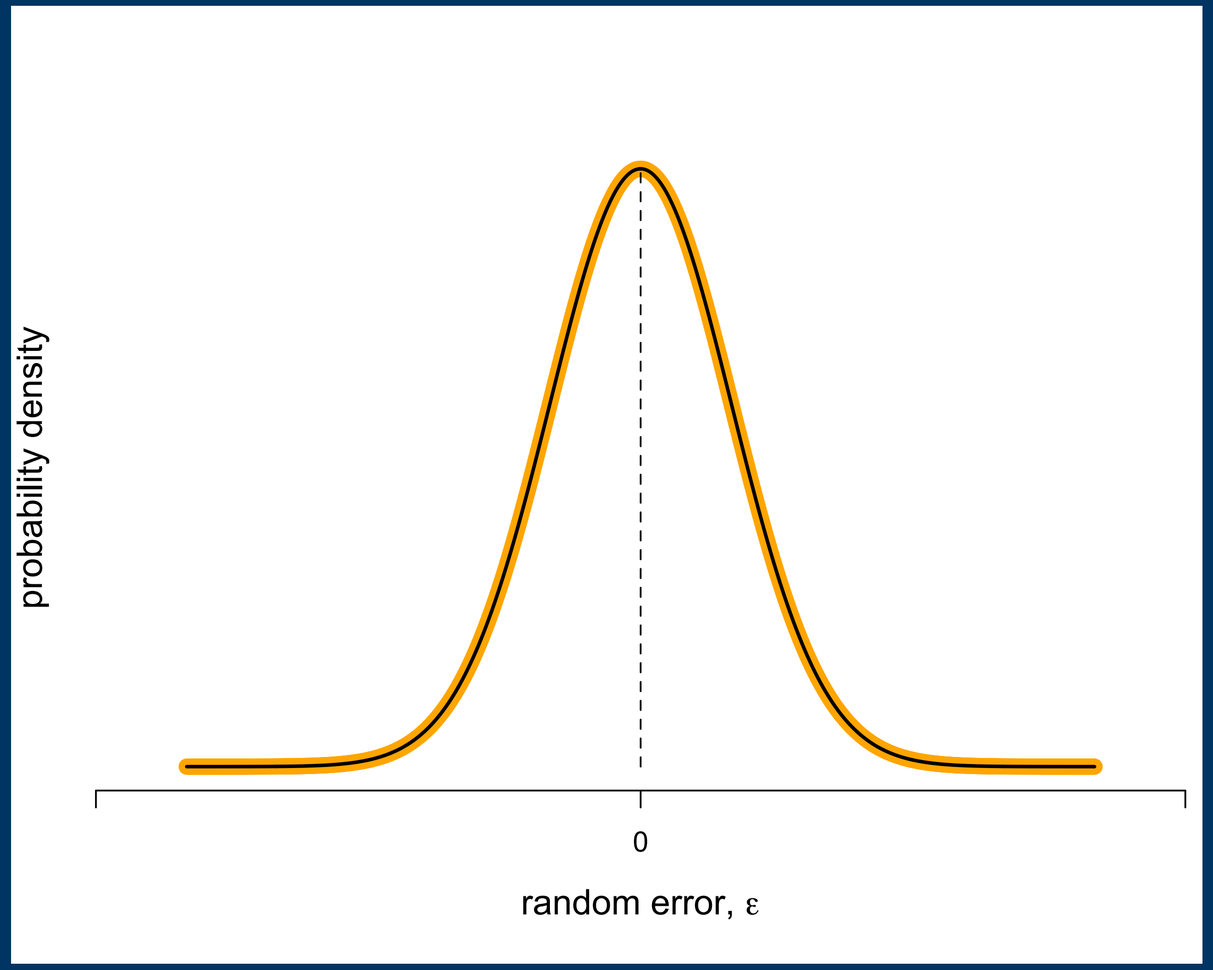
Introducing uncertainty/randomness

$$y_i = a + bx_i + \epsilon_i, \quad i = 1, \ldots, n$$

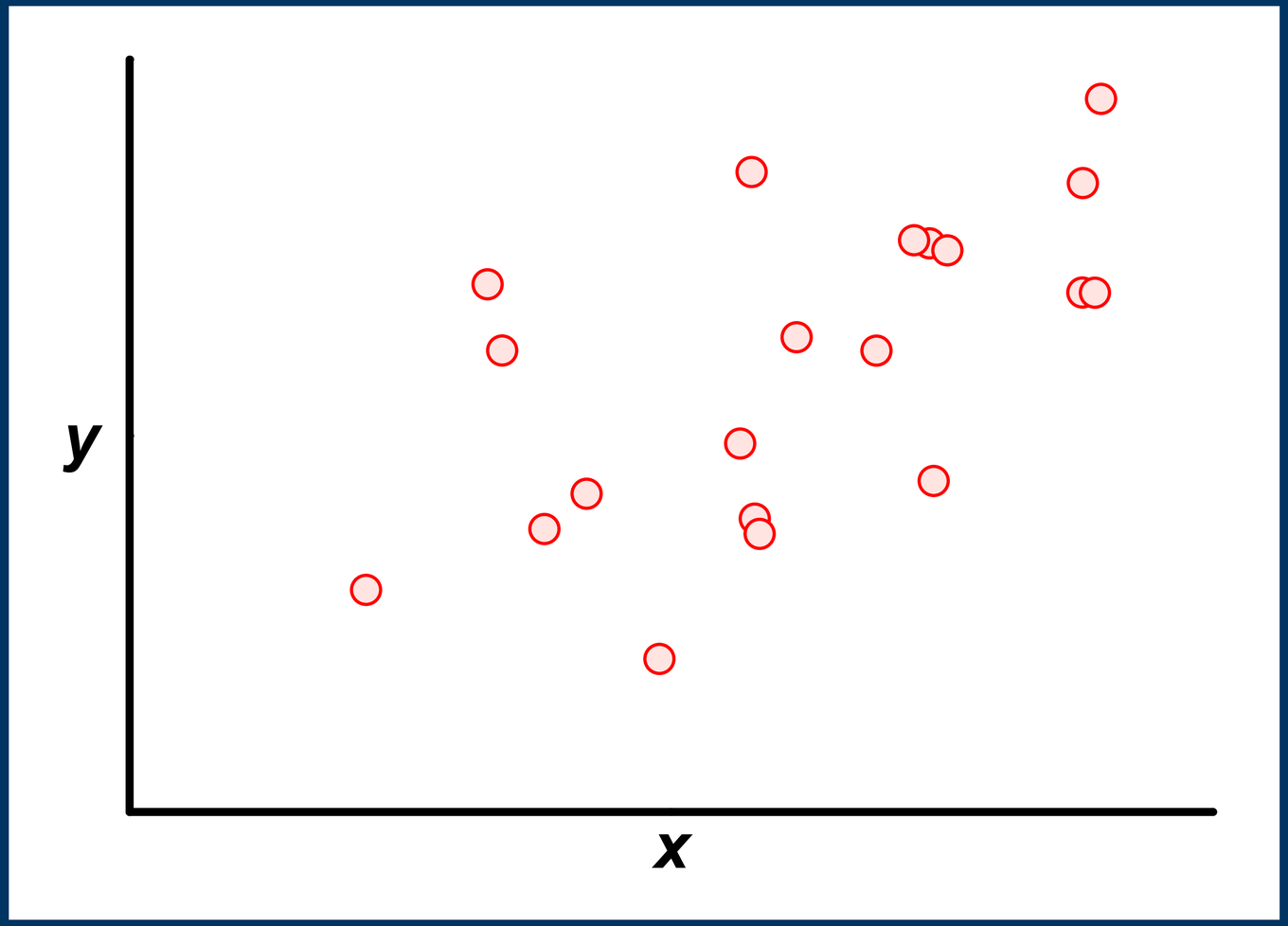
$$\epsilon_i \rightsquigarrow N(\mu = 0, \sigma)$$

Random error

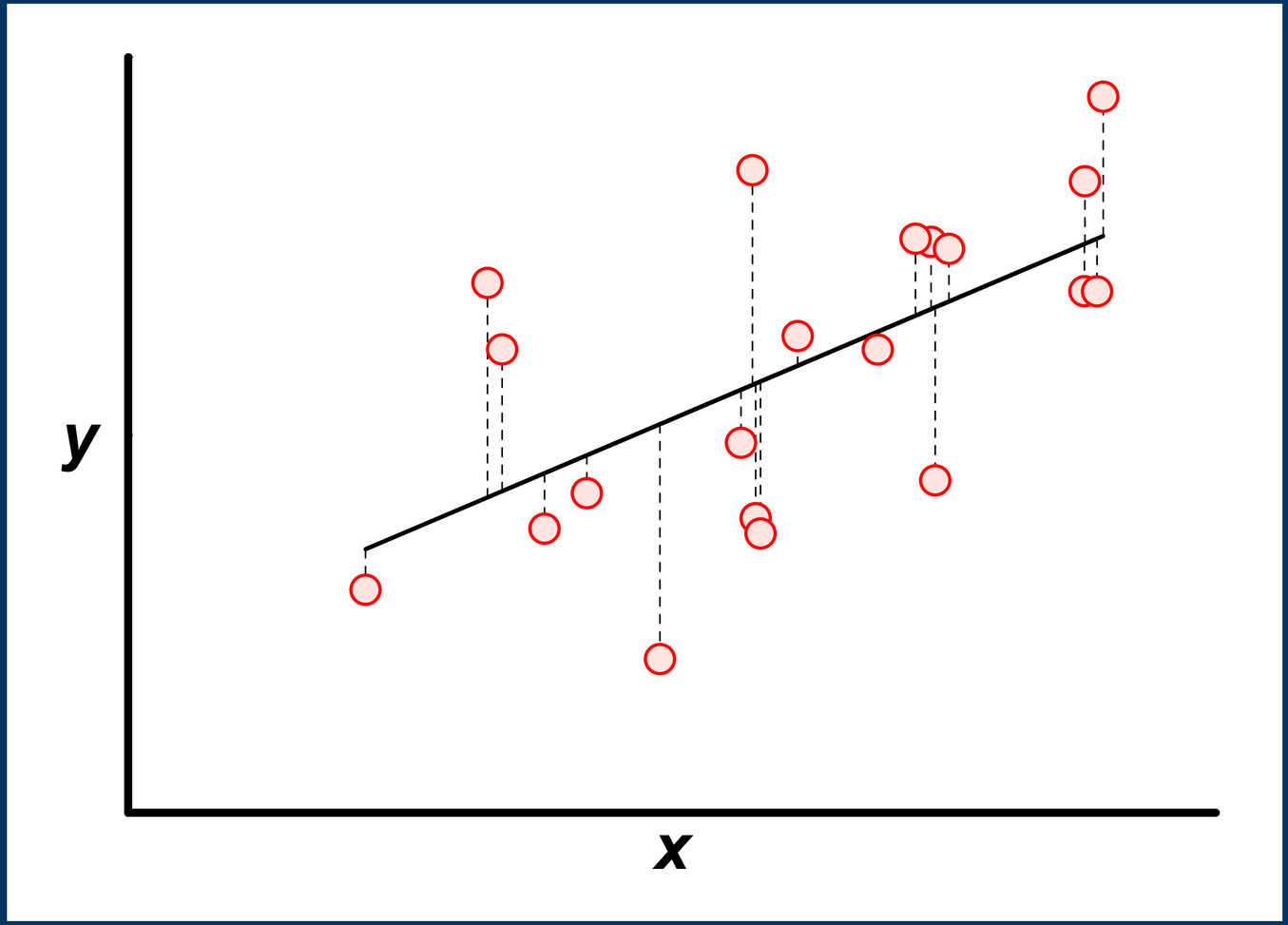
What is the source of this random error?



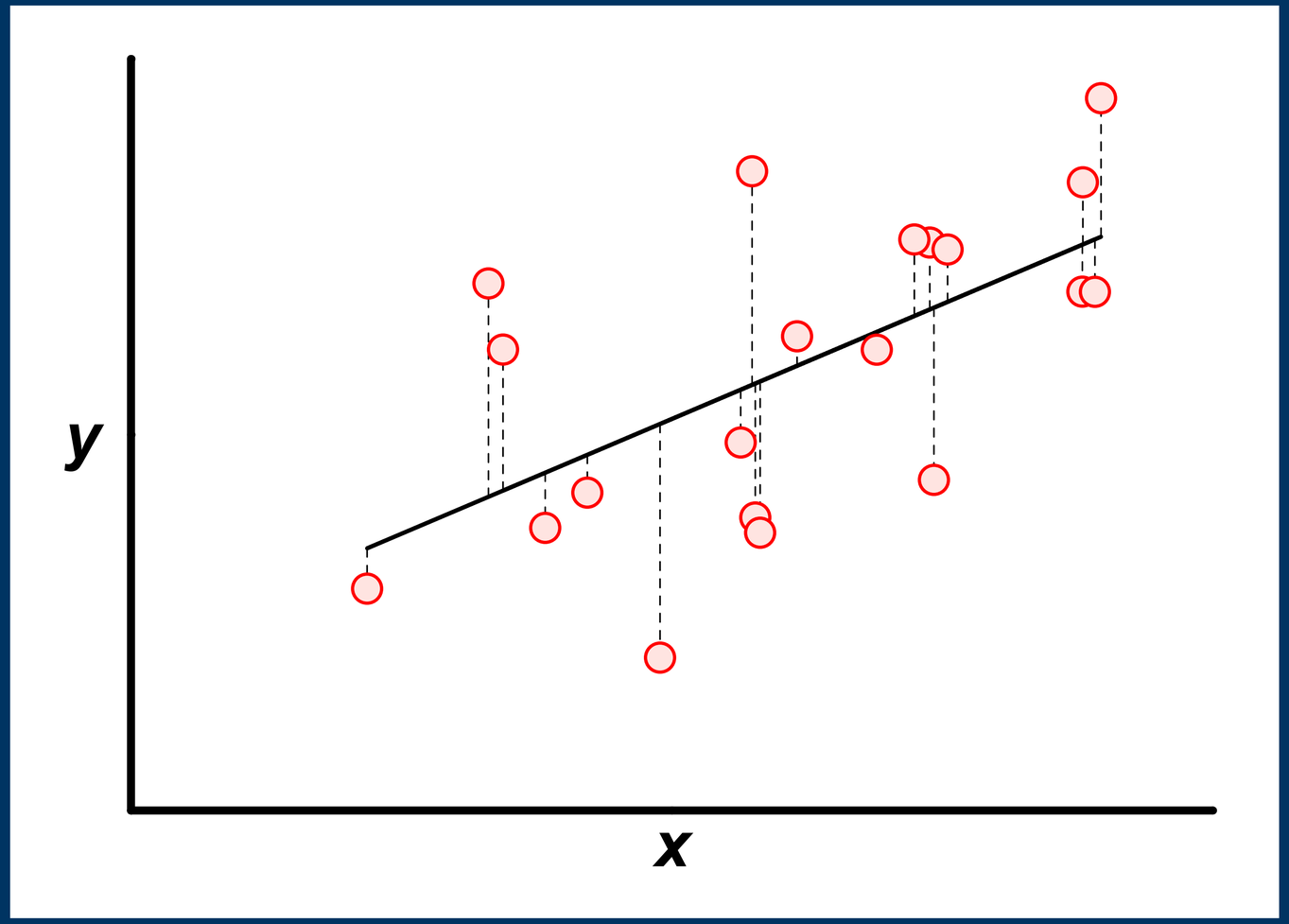
How to estimate *a* and *b*?



Ordinary Least Squares Method



Ordinary Least Squares Method



$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Inference of interest (in large samples)

$H_0 : a = 0$ versus $H_1 : a \neq 0$

$$t = \frac{\hat{a}}{se\left(\hat{a}\right)} \mid H_0 \rightsquigarrow N\left(\mu = 0, \sigma = 1\right)$$

$H_0 : b = 0$ versus $H_1 : b \neq 0$

$$t = \frac{\hat{b}}{se\left(\hat{b}\right)} \mid H_0 \rightsquigarrow N\left(\mu = 0, \sigma = 1\right)$$

p-value < 0.05, reject H_0

p-value ≥ 0.05, not reject H_0

0.05 is the **significance level of the test**

Warnings

Technical warning

Be aware of the model assumptions and their validity in the data

Interpretative warnings

Be aware of the dangers of extrapolating

Be aware of the dangers of inferring causality

It is R time!