RenR 480/711

Regression and correlation

Regression and Correlation

When to use it?

→ when independent variable is not categorical

Correlation analysis ← technically the same → Regression analysis

but

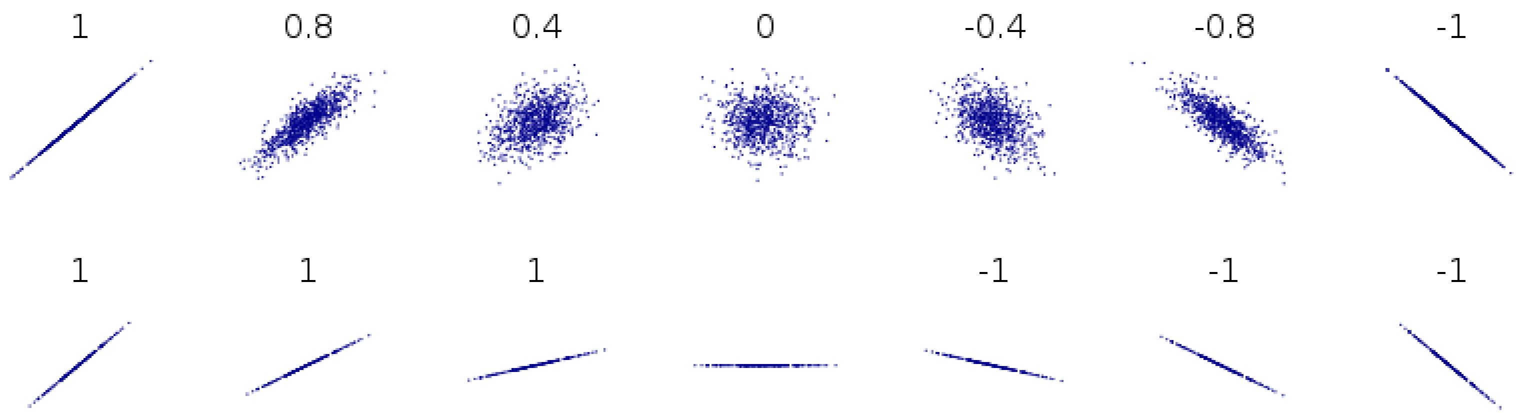
Different question/objective

Is there a relationship between variables?

Can we predict one variable from another?

Pearson correlation coefficient

Ranges from 1 to -1



Measure of strength and direction of a relationship
Only linear relationships are reflected

https://en.wikipedia.org/wiki/Correlation_and_dependence

Pearson correlation coefficient

Assumptions:

→ Normal distribution of experimental errors

→ Homogeneity of variances

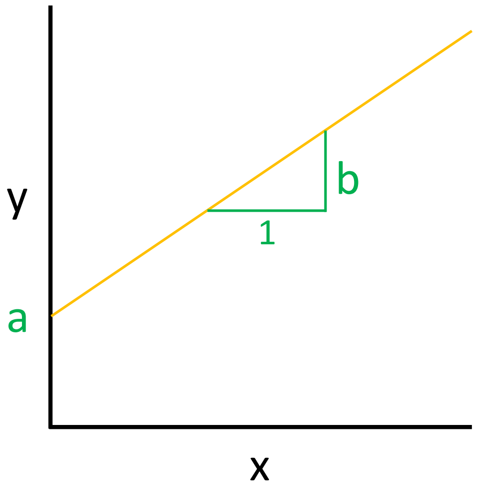
→ Independence of samples

Alternatives when assumptions are not met:

Rank based methods:

1. Kendall’s Correlation Coefficient
2. Spearman’s Correlation Coefficient

Linear regression



$y = a + bx$

a: intercept with y-axis
b: slope of one unit of x

$b = r$ if the data is normalized

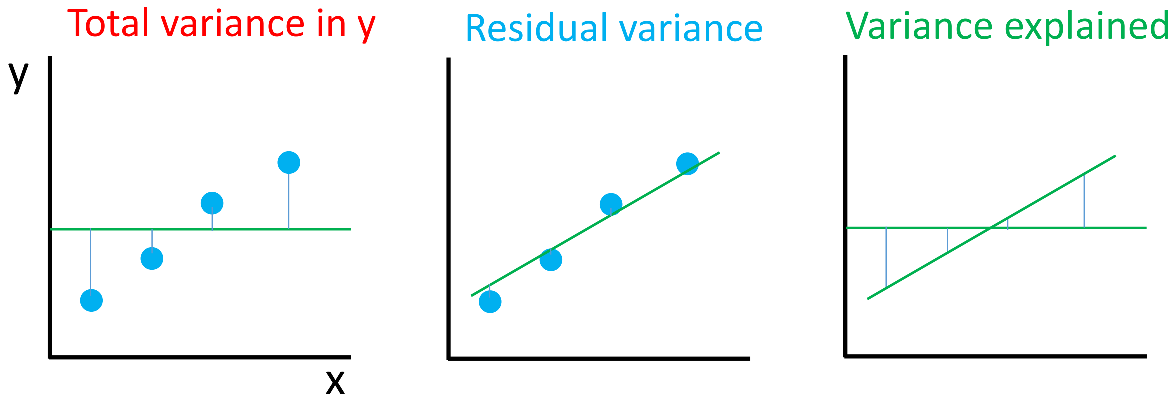
Linear regression

Calculation of r^2:

$$r^2 = \frac{Signal}{Noise}$$

$$r^2 = \frac{Variance\ explained}{Total\ Variance} \qquad r = \sqrt{\frac{Variance\ explained}{Total\ Variance}}$$

Variance components:



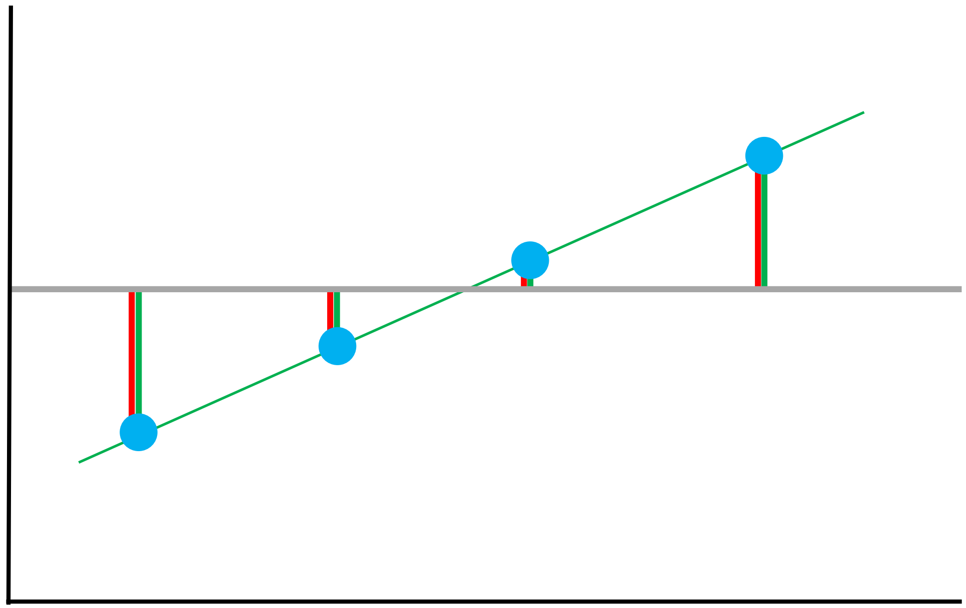
r^2 ranges from 0 to 1 and therefore does not indicate direction

Linear regression

Total variance

Residual variance

Explained variance



100% of total variance explained

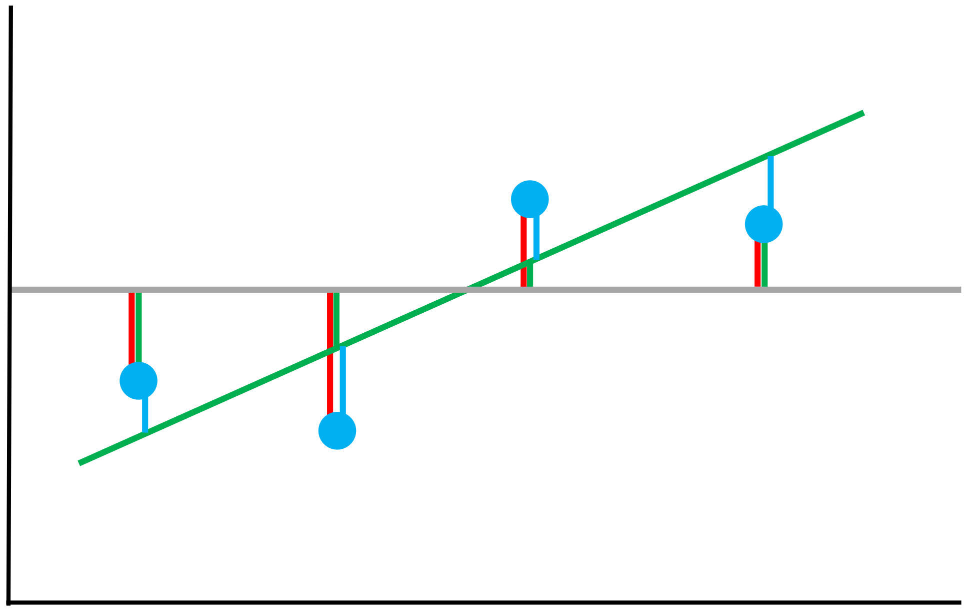
$r^2 = 1$

Linear regression

Total variance

Residual variance

Explained variance



~60% of total variance explained

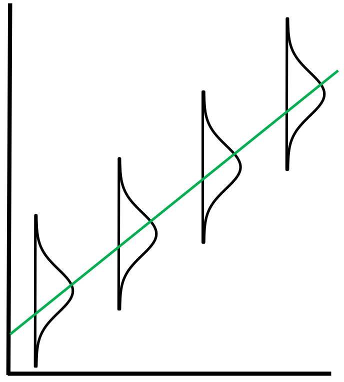
$r^2 = 0.6$

Linear regression

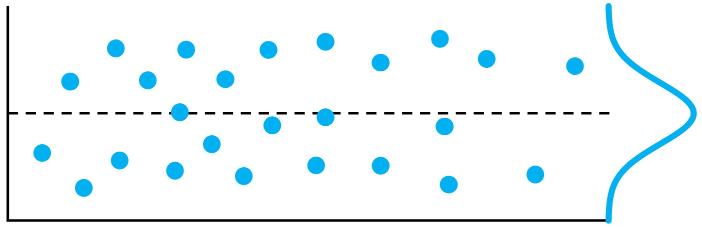
<u>Assumptions for linear regression</u>

→ normality in y for a given x

→ equal variances in y for a given x



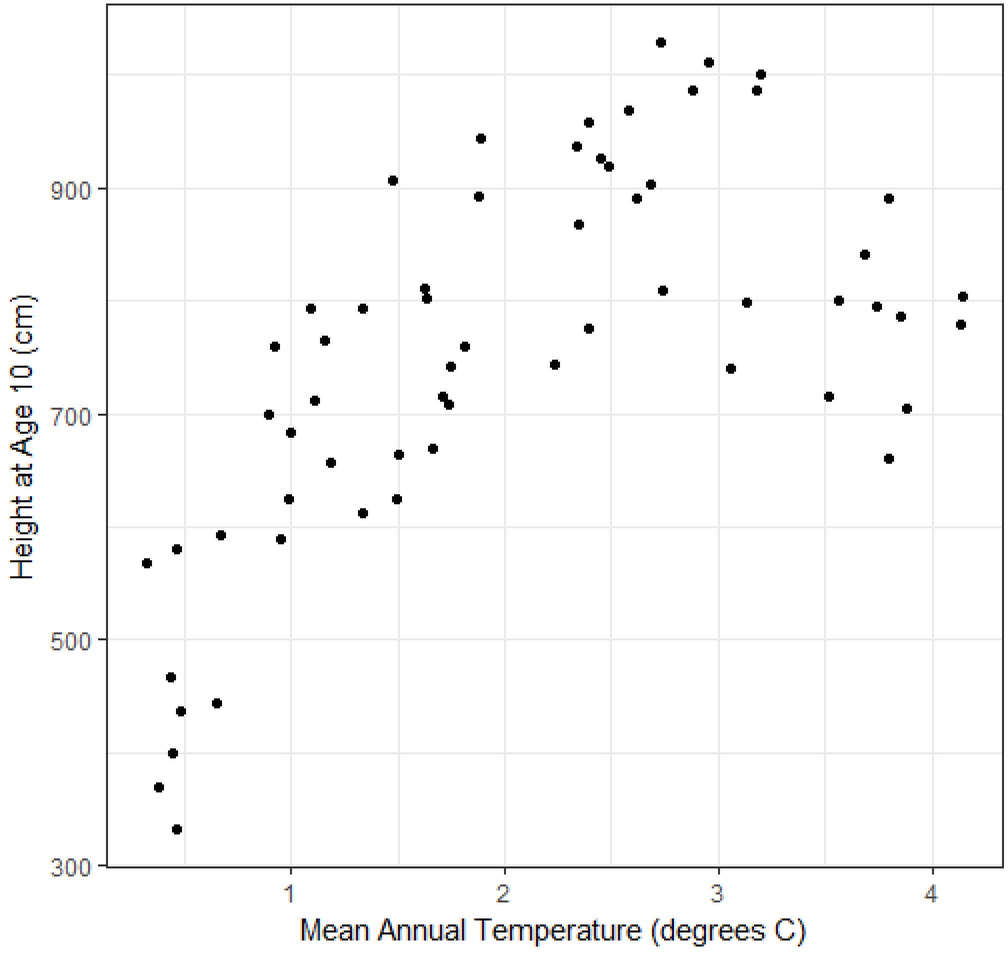
Test with residual plots



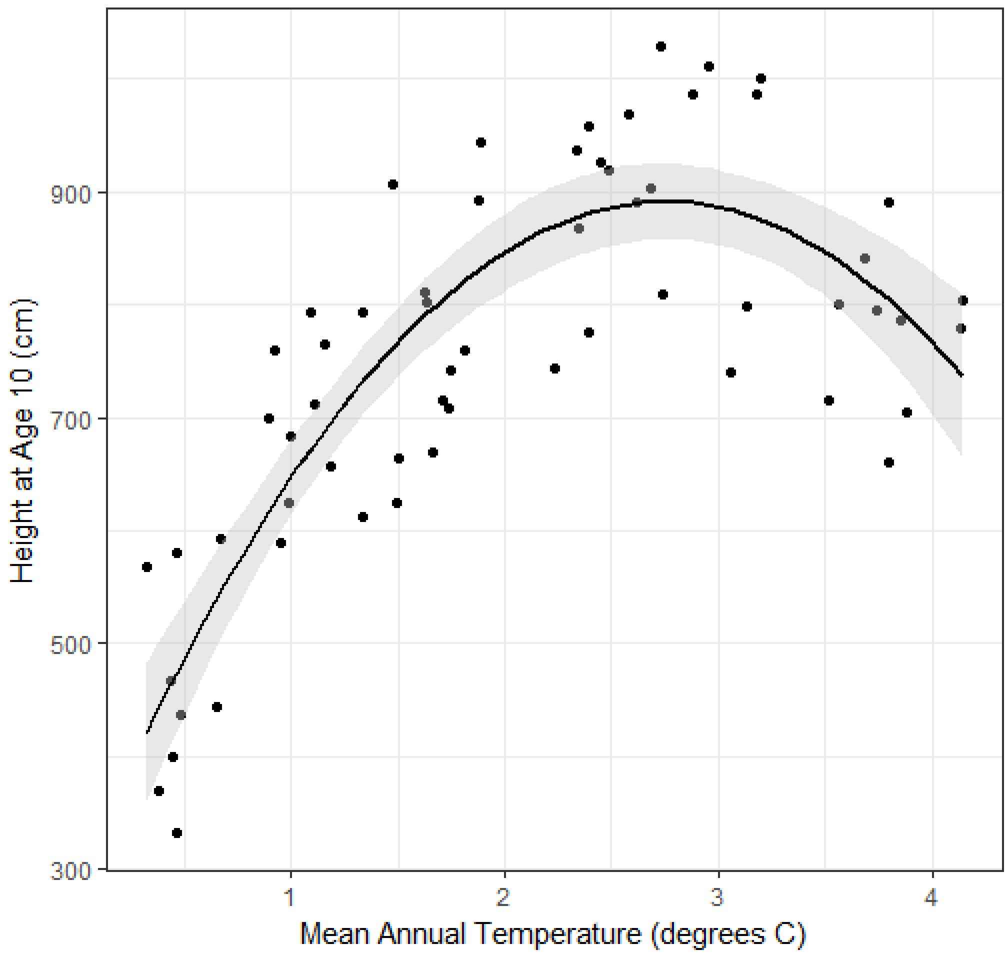
RenR 480/711

Non-linear Regression

Non-linear regression



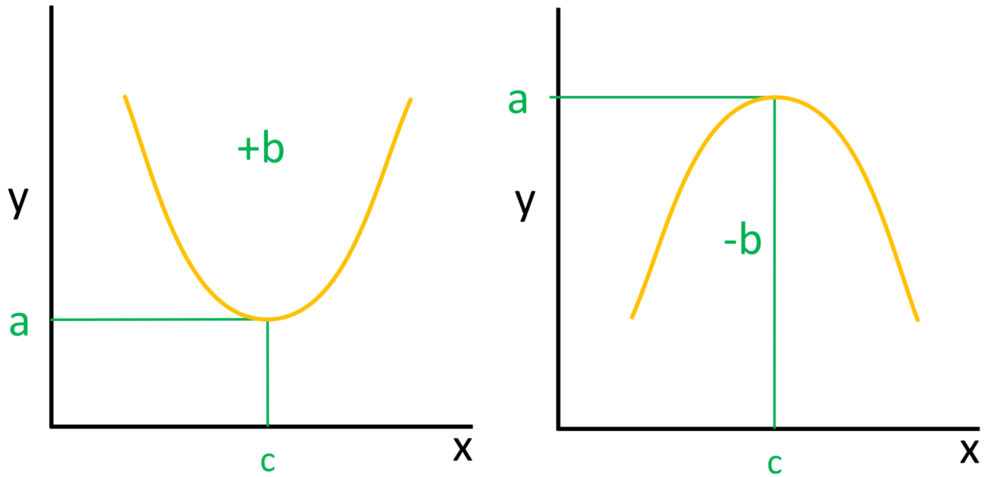
Non-linear regression



Non-linear regression

Parabolic relationship

$$y = a + b(x - c)^2 \qquad y = a - b(x - c)^2$$



Steep increase/decrease and reversal of the trend

Example: Growth and survival over an environmental gradient

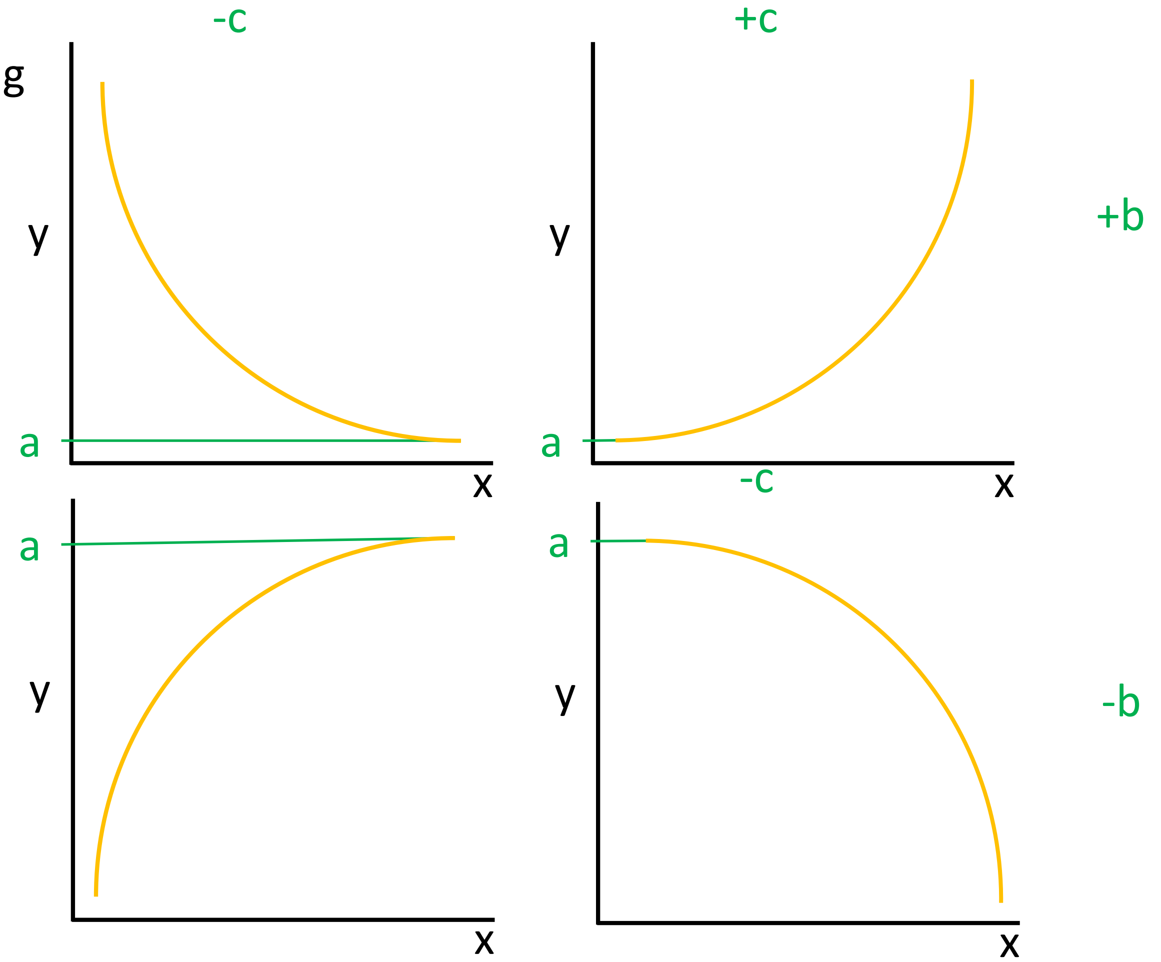
Non-linear regression

Exponential

$y = a + bc^x$

Increasing/decreasing rate of change

Example:
Population growth
Compound interest



Non-linear regression

Logarithmic

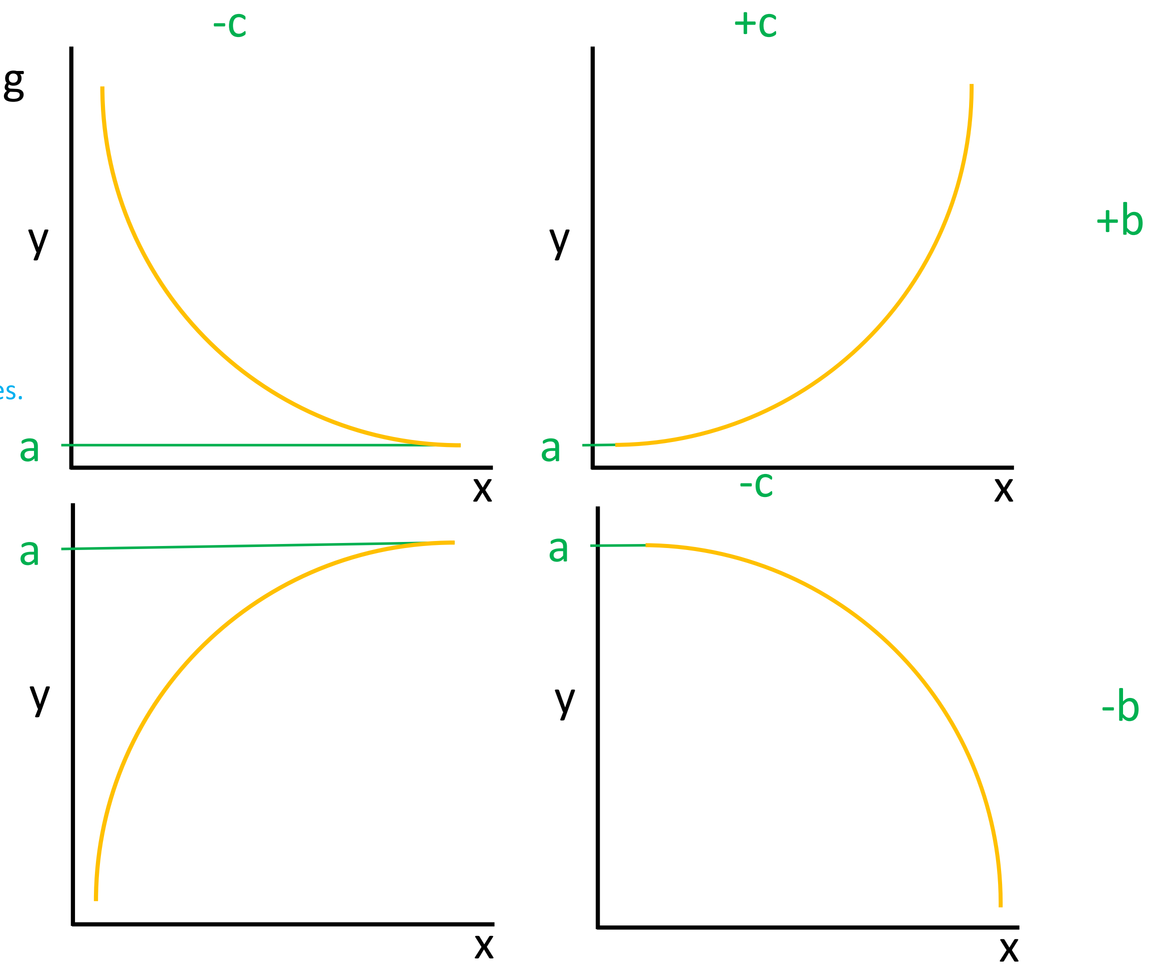
$$y = a + bx^c$$

Increasing/decreasing rate of change

Example:

Human perception:

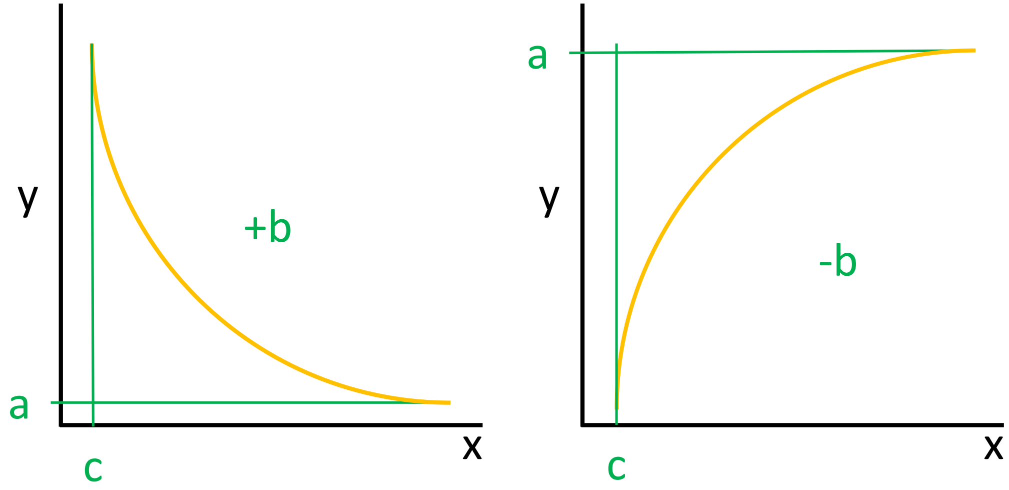
Time to make a decision is the logarithm of the number of choices.



Non-linear regression

Hyperbolic

$$y = a + \frac{b}{x + c}$$



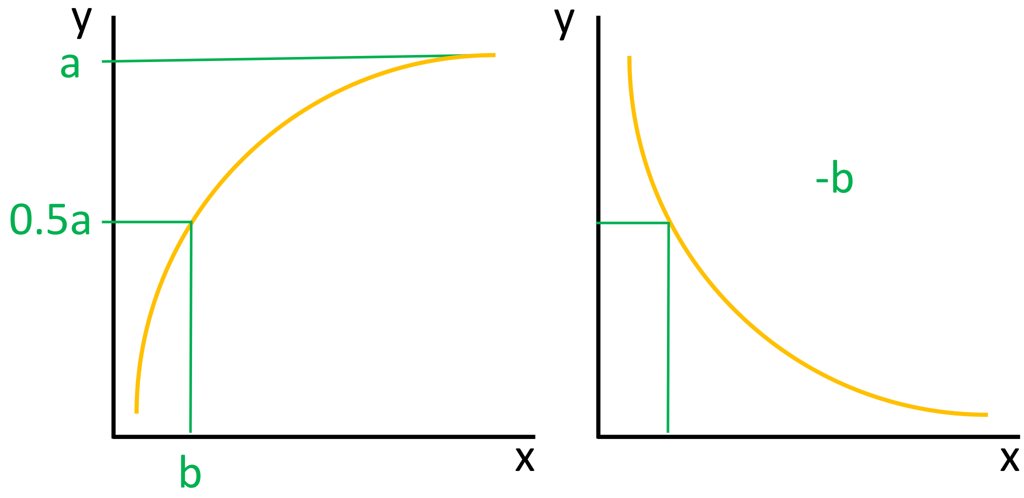
Similar to exponential/logarithmic but with two asymptotes

Example: Gravitational pull on a cable fixed between two points

Non-linear regression

Michaelis Menten

$$y = \frac{ax}{b + x}$$



Similar to exponential/logarithmic but with two parameters

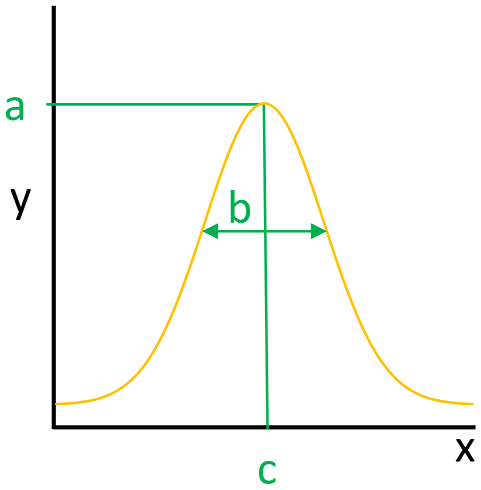
Example: growth as a function of resource availability

Kinetic models

Non-linear regression

Gaussian

$$y = ab^{(x-c)^2}$$



Normal curve

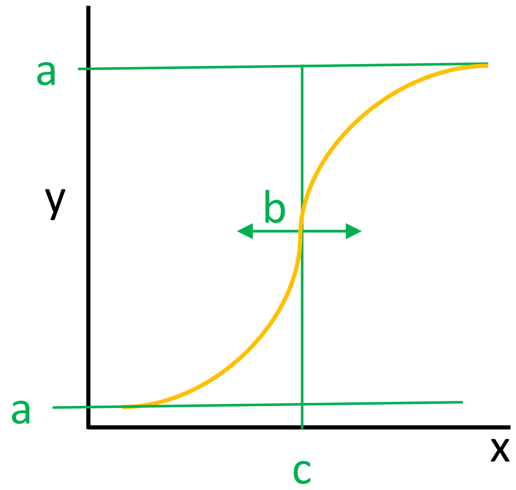
Example: Distribution of sample means

Growth and survival over an environmental gradient

Non-linear regression

Sigmoidal

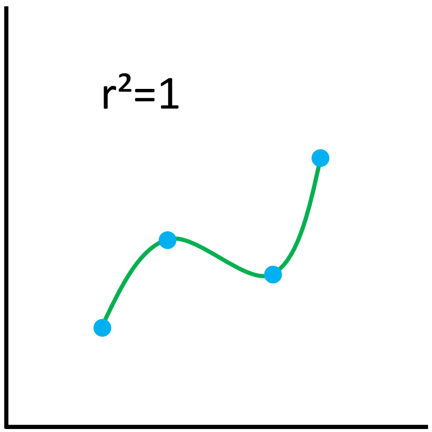
$$y = \frac{a}{1 + b^{(x-c)}} + d$$



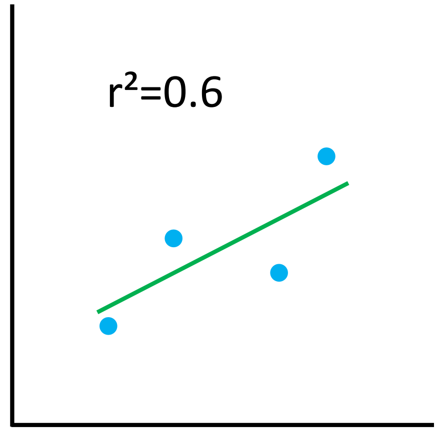
Stable rate of change followed by step increase followed by again stable change

Example: Learning, restricted growth

Overfitting and AIC



$$y = a + b_1 x + b_2 + x^2 + b_3 x^3$$



$$y = ax + b$$

Occam's razor: If there is a variety of explanations, we should select the simplest one.

Akaike's Information Criterion (AIC)

→ Measure of model complexity weighted by explained variance

→ lower AIC value = better model