Reading, writing and interpreting regression equations

QAEco coding club: Oct 5 2017

Heavy use of equations impedes communication among biologists

Tim W. Fawcett[1] and Andrew D. Higginson

School of Biological Sciences, University of Bristol, Bristol BS8 1UG, United Kingdom

Edited[†] by Robert M. May, University of Oxford, Oxford, United Kingdom, and approved June 6, 2012 (received for review April 4, 2012)

Most research in biology is empirical, yet empirical studies rely fundamentally on theoretical work for generating testable predictions and interpreting observations. Despite this interdependence, many empirical studies build largely on other empirical studies with little direct reference to relevant theory, suggesting a failure of communication that may hinder scientific progress. To investigate the extent of this problem, we analyzed how the use of mathematical equations affects the scientific impact of studies in ecology and evolution. The density of equations in an article has a significant negative impact on citation rates, with papers receiving 28% fewer citations overall for each additional equation per page in the main

for enhancing the presentation of mathematical models to facilitate progress in disciplines that rely on the tight integration of theoretical and empirical work.

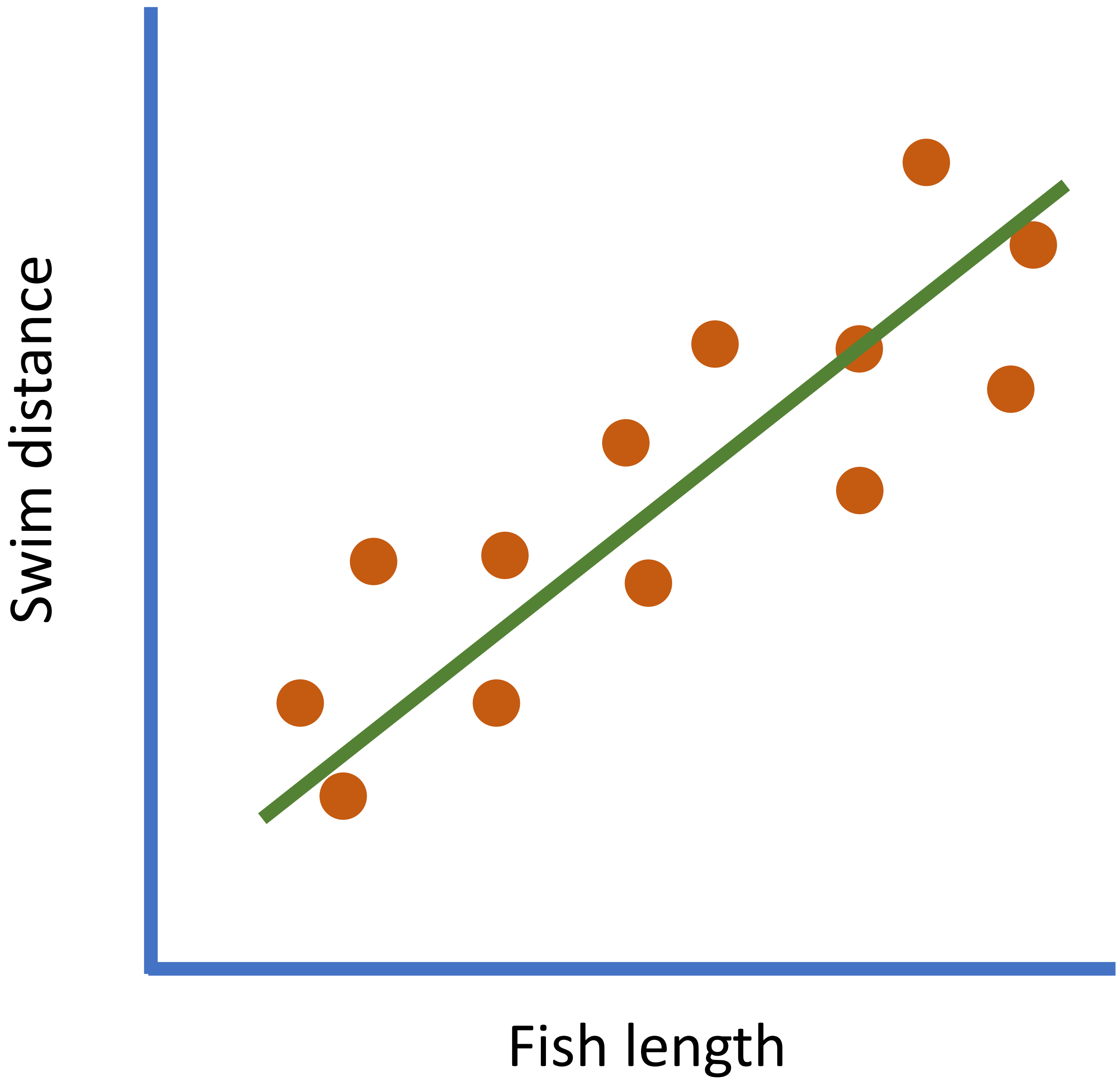
Results

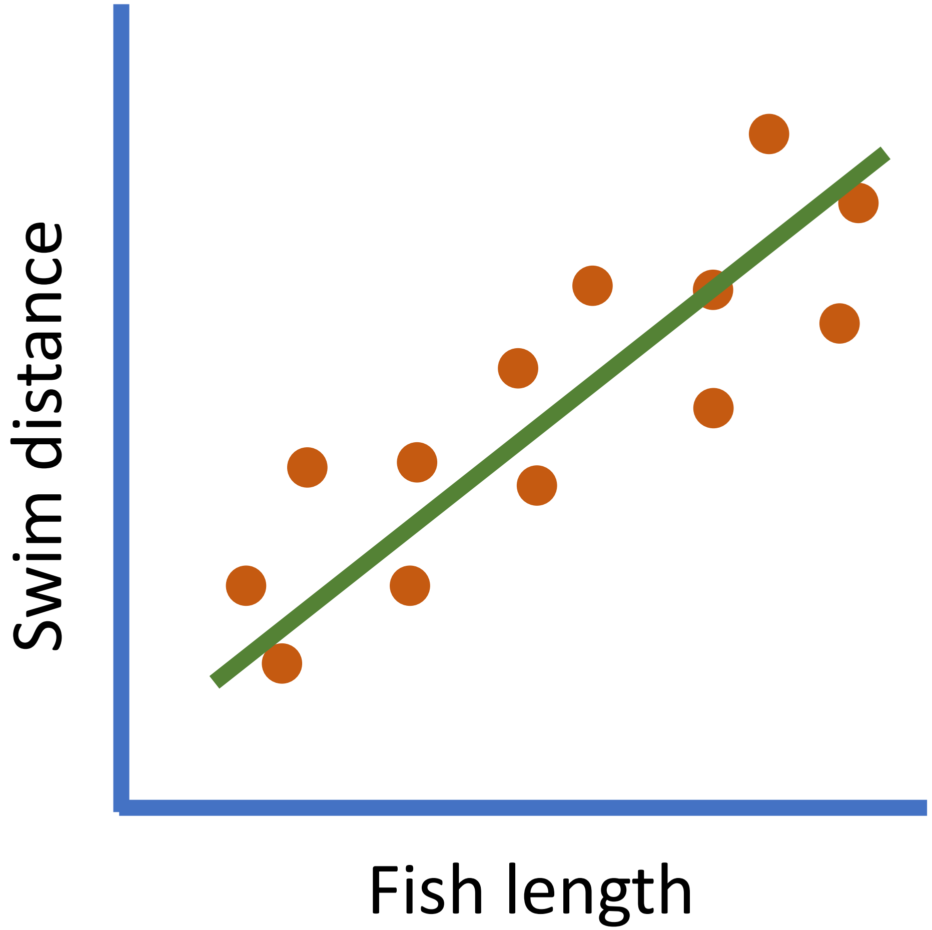
To quantify the technical level of any theory presented in the articles, we counted equations, inequalities, and other mathematical expressions (hereafter referred to simply as "equations") in the main text and any printed appendixes. We divided this count by the number of pages to give a measure of equation density, which ranged from 0 to 7.29 equations per page (mean ± SEM: 0.43 ±

Level 1: linear regression

How big is it?

How far did it swim?


Swim distance
Fish length



In words

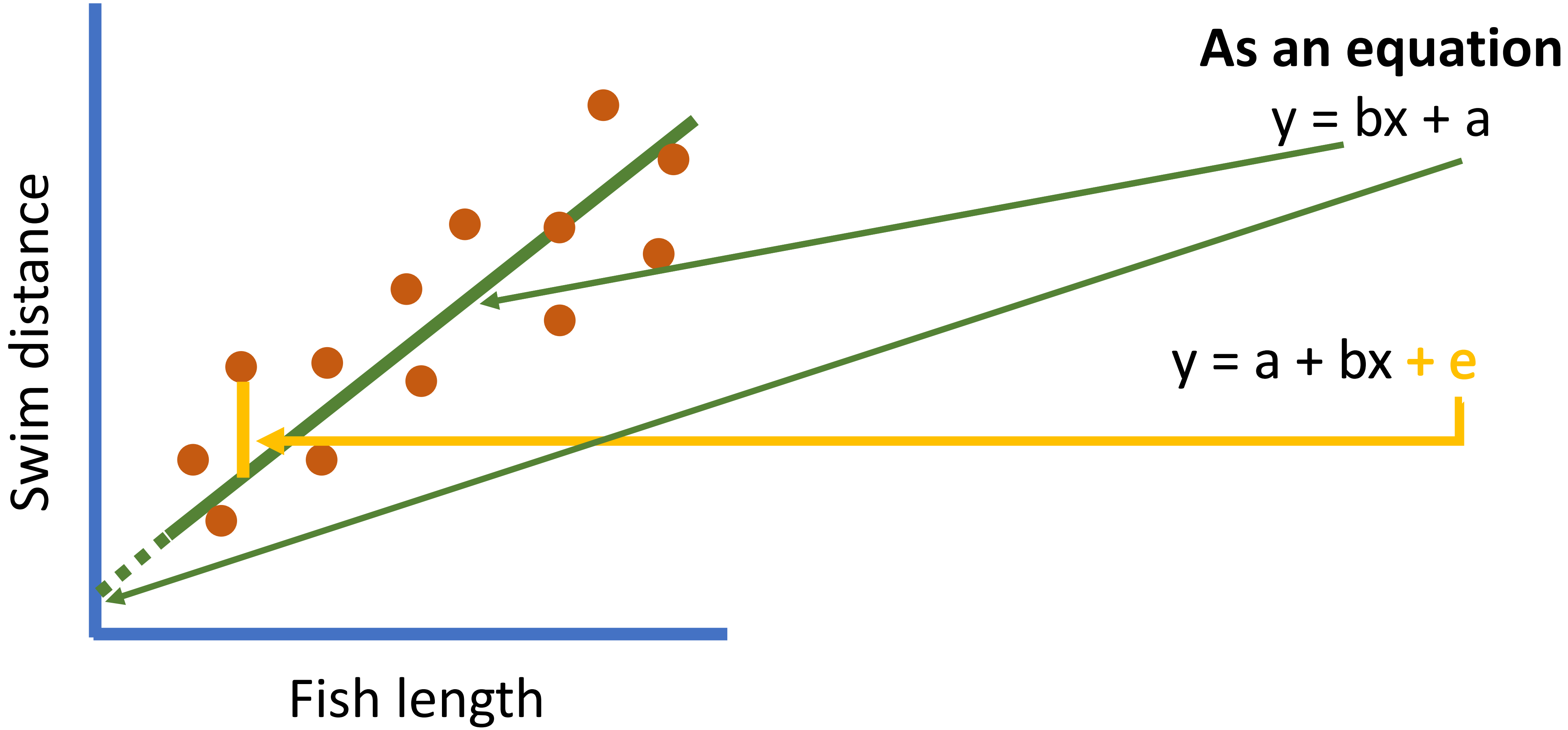
We fitted a linear regression model to data on swim distance (response variable) and fish length (predictor variable).

In R

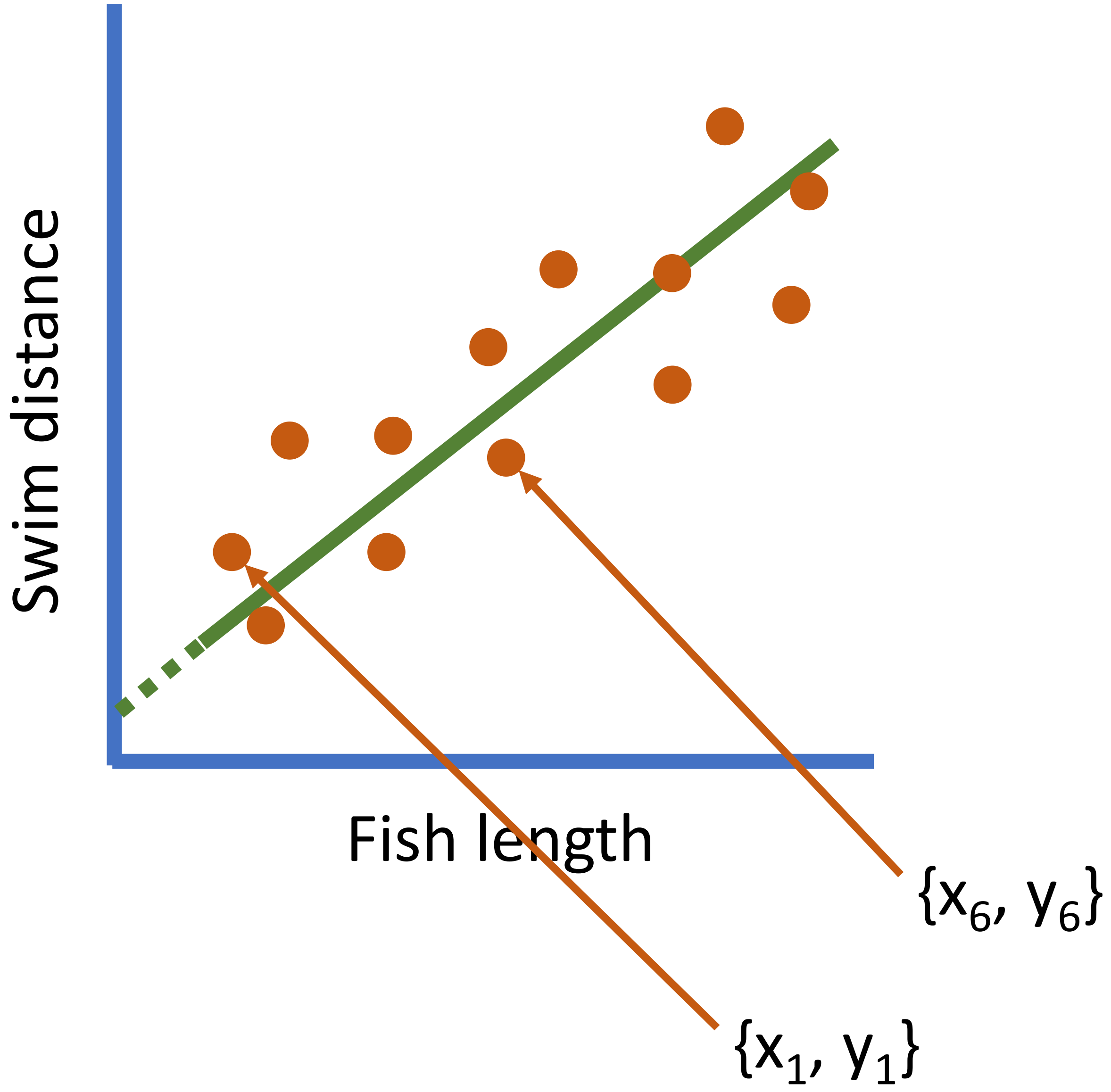
lm(swimDistance ~ fishLength)

or

lm(y ~ x)



response = intercept + slope x predictor + error

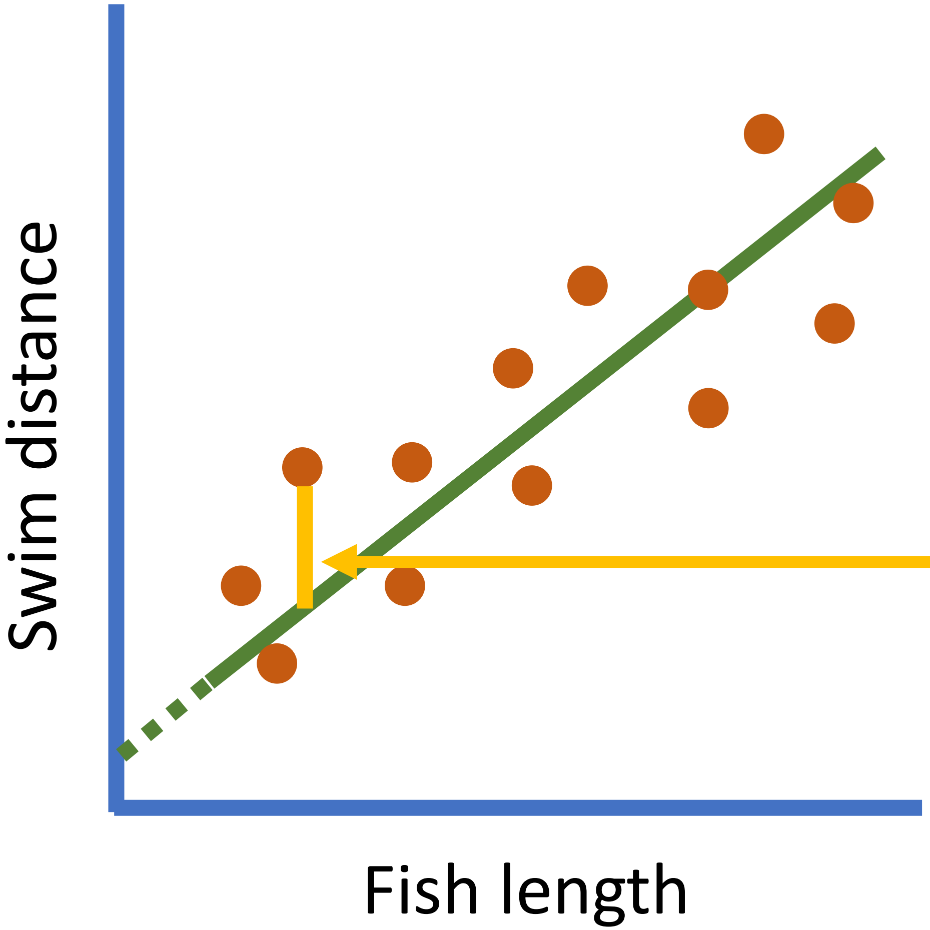


As an equation

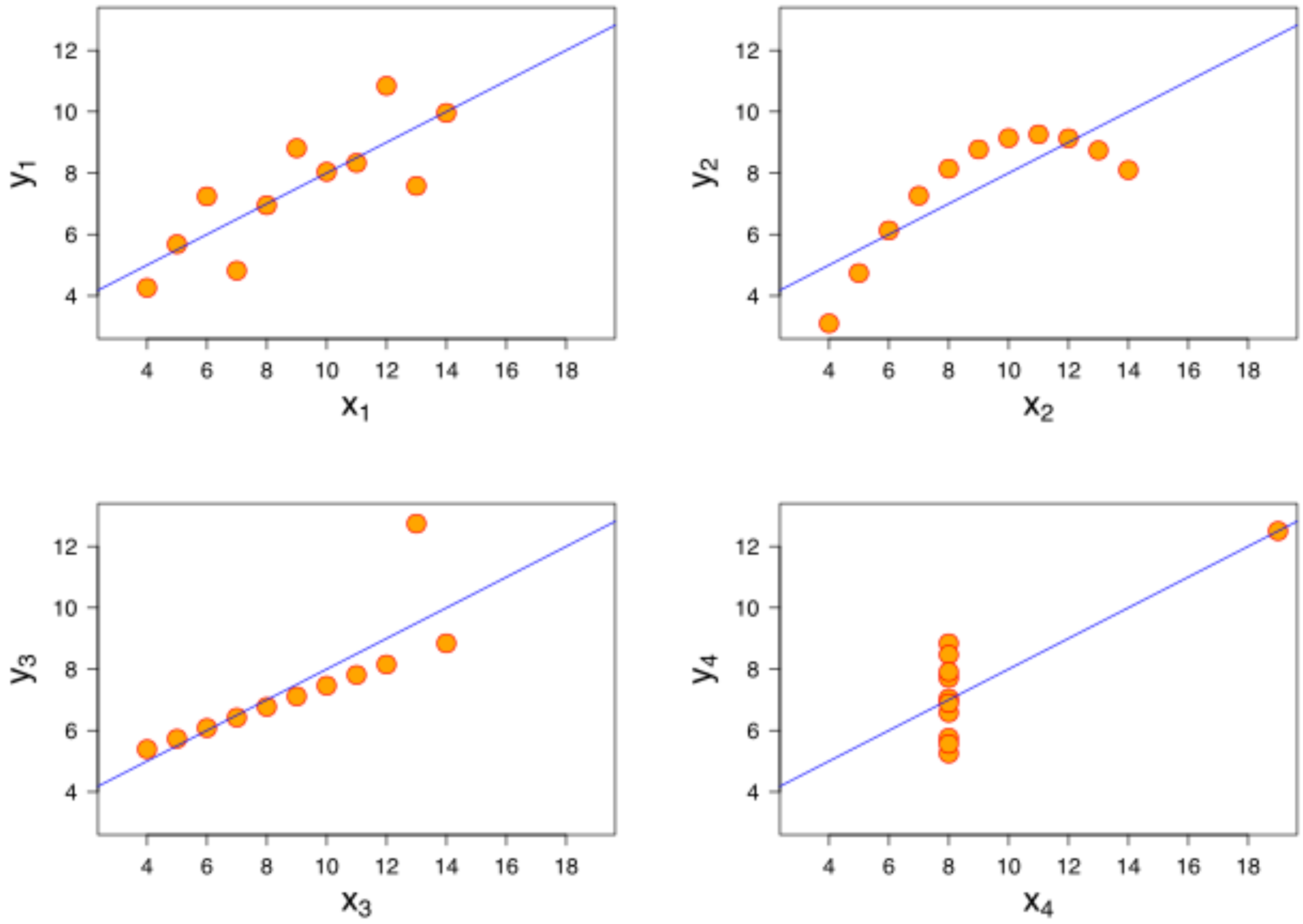
$$y = a + bx + e$$

$$y_i = a + bx_i + e_i$$

where y_i is the distance fish i has swum, a is the swim distance when fish length is zero, b is the association between the length of fish i (x_i) and swim distance, and e_i is residual variation in swim distance of fish i.



$$y_i = a + bx_i + e_i$$



$$y_i = a + bx_i + e_i$$

Jian’s rules

Greek letters for parameters; Roman letters for data

$y_i = a + bx_i + e_i$

Define all terms

Single letters where possible

Jian's rules

Greek letters for parameters; Roman letters for data

$y_i = \alpha + \beta x_i + \varepsilon_i$

Define all terms

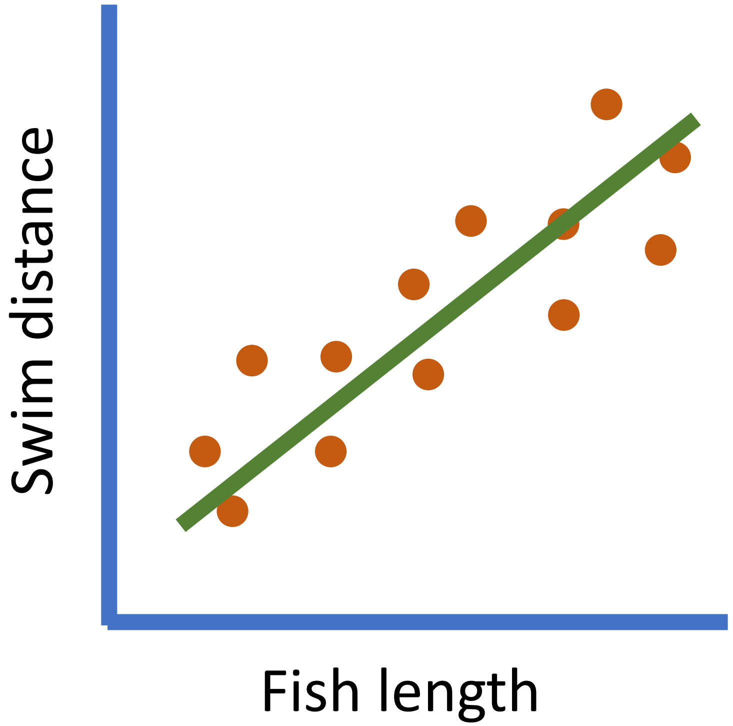
Single letters where possible

Jian’s rules

Greek letters tters for data

Si e

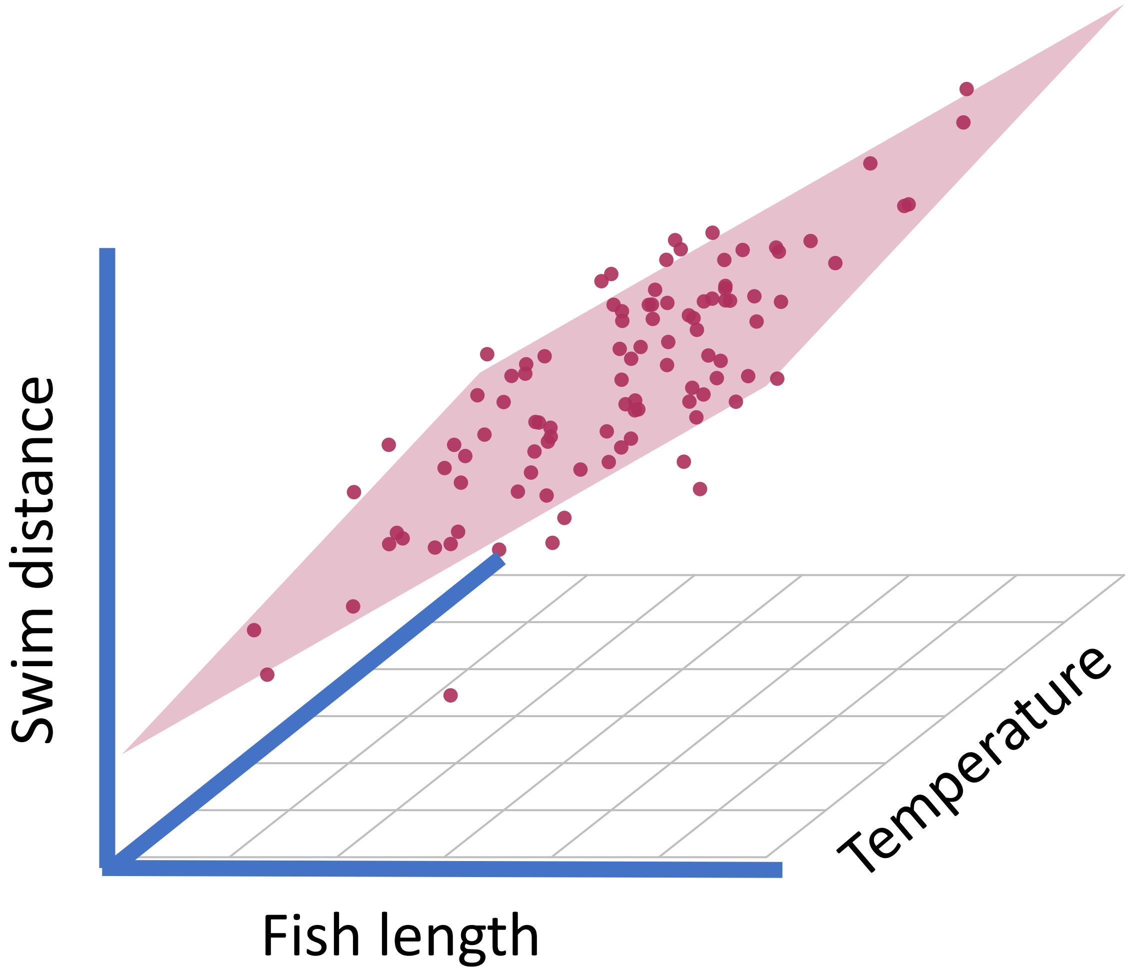
Level 2: vectors and matrices



$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{pmatrix} = \alpha + \beta \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \alpha + \beta\mathbf{x} + \boldsymbol{\varepsilon}$$



In R

lm(swimDistance ~ fishLength + temperature)

or

lm(y ~ x_1 + x_2)

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

$$\mathbf{y} = \alpha + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{X} = \begin{pmatrix} \text{length}_{1,1} & \text{temp}_{1,2} \\ \text{length}_{2,1} & \text{temp}_{2,2} \\ \text{length}_{3,1} & \text{temp}_{3,2} \\ \ldots & \ldots \\ \text{length}_{n,1} & \text{temp}_{n,2} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \text{effect of length} \\ \text{effect of temp} \end{pmatrix}$$

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ x_{3,1} & x_{3,2} \\ \dots & \dots \\ x_{n,1} & x_{n,2} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

indexing denotes row and column:
$\textit{value}_{\textit{row,column}}$

indexing denotes row:
$\textit{value}_{\textit{row}}$

$$\mathbf{y} = \alpha + \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{X}\,\boldsymbol{\beta} = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ x_{3,1} & x_{3,2} \\ \dots & \dots \\ x_{n,1} & x_{n,2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} x_{1,1}\beta_1 + x_{1,2}\beta_2 \\ x_{2,1}\beta_1 + x_{2,2}\beta_2 \\ x_{3,1}\beta_1 + x_{3,2}\beta_2 \\ \dots \\ x_{n,1}\beta_1 + x_{n,2}\beta_2 \end{pmatrix}$$

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{pmatrix} = \alpha + \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ x_{3,1} & x_{3,2} \\ \dots & \dots \\ x_{n,1} & x_{n,2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{pmatrix} = \alpha + \begin{pmatrix} x_{1,1}\beta_1 + x_{1,2}\beta_2 \\ x_{2,1}\beta_1 + x_{2,2}\beta_2 \\ x_{3,1}\beta_1 + x_{3,2}\beta_2 \\ \dots \\ x_{n,1}\beta_1 + x_{n,2}\beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

In a linear model the distribution of $\mathcal{Y}$ is multivariate normal,

$$\mathcal{Y} \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{o}, \sigma^2 \boldsymbol{W}^{-1}), \tag{1}$$

where n is the dimension of the response vector, $\boldsymbol{W}$ is a diagonal matrix of known prior weights, $\boldsymbol{\beta}$ is a p-dimensional coefficient vector, $\boldsymbol{X}$ is an $n \times p$ model matrix, and $\boldsymbol{o}$ is a vector of known prior offset terms. The parameters of the model are the coefficients $\boldsymbol{\beta}$ and the scale parameter σ.

In a linear mixed model it is the *conditional* distribution of $\mathcal{Y}$ given $\mathcal{B} = \boldsymbol{b}$ that has such a form,

$$(\mathcal{Y}|\mathcal{B} = \boldsymbol{b}) \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{o}, \sigma^2 \boldsymbol{W}^{-1}), \tag{2}$$

where $\boldsymbol{Z}$ is the $n \times q$ model matrix for the q-dimensional vector-valued random-effects variable, $\mathcal{B}$, whose value we are fixing at $\boldsymbol{b}$. The unconditional distribution of $\mathcal{B}$ is also multivariate normal with mean zero and a parameterized $q \times q$ variance-covariance matrix, $\boldsymbol{\Sigma}$,

$$\mathcal{B} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{3}$$

Jian’s rules

Greek letters for parameters; Roman letters for data

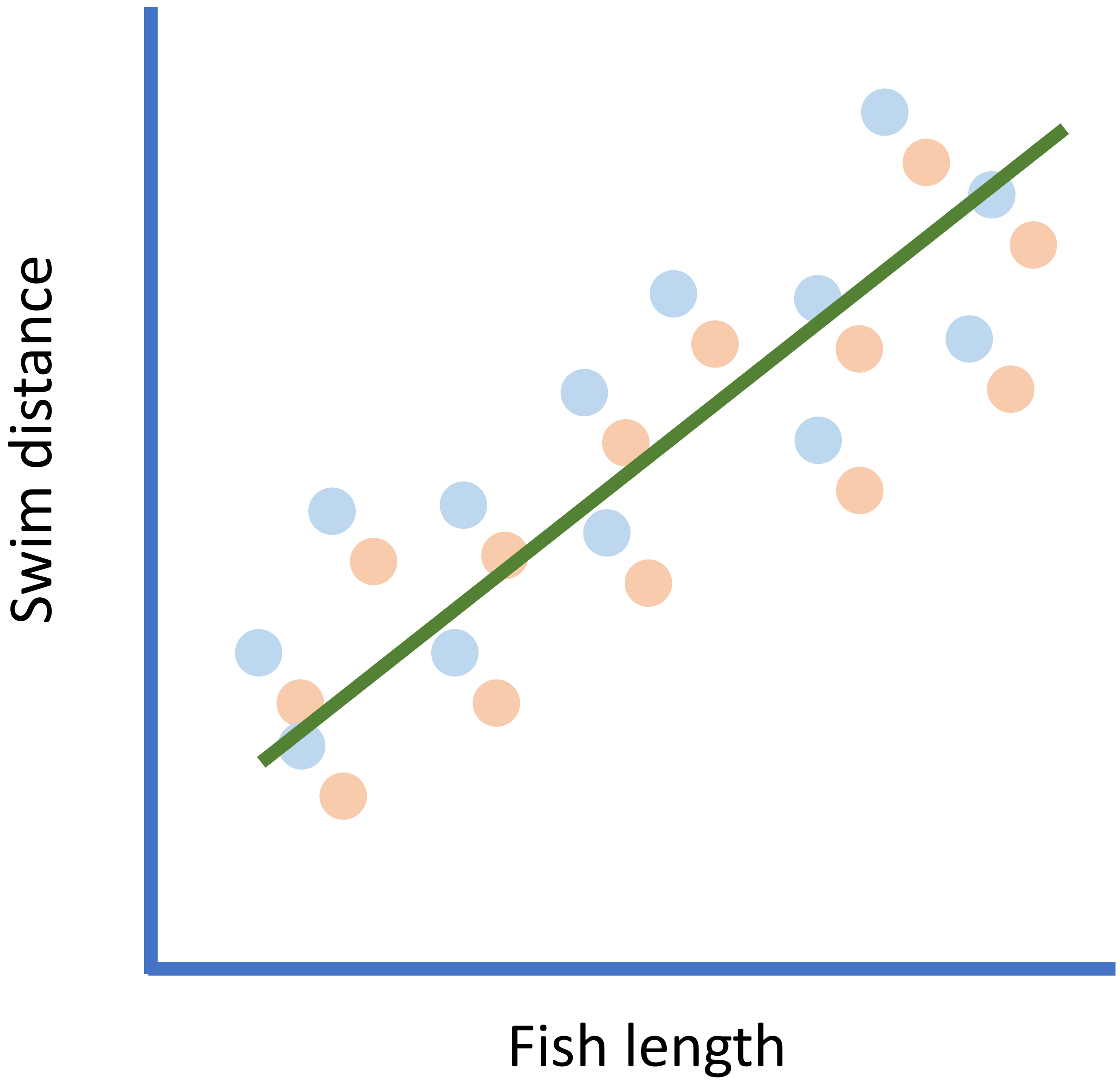
Define all terms

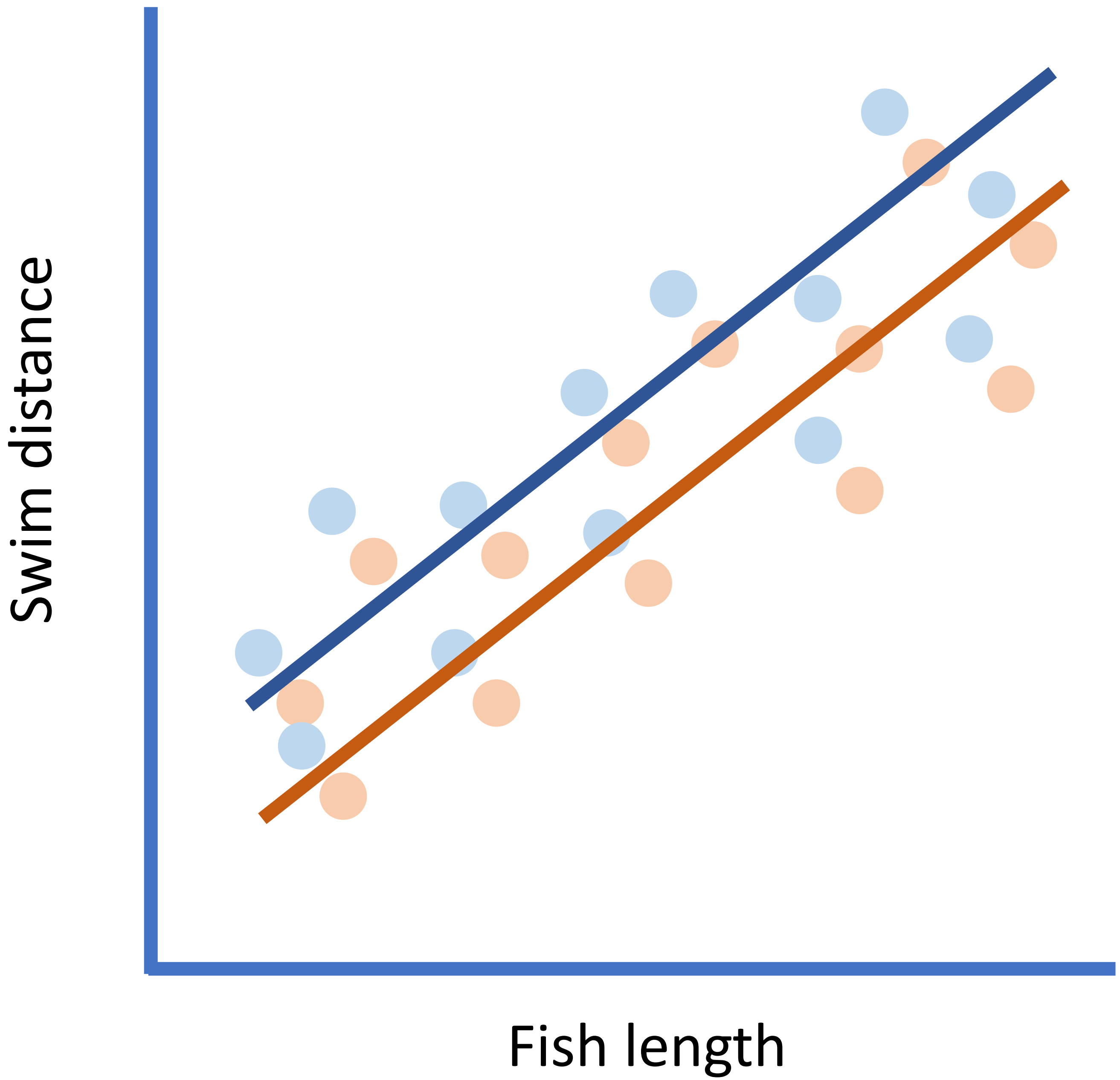
Single letters where possible

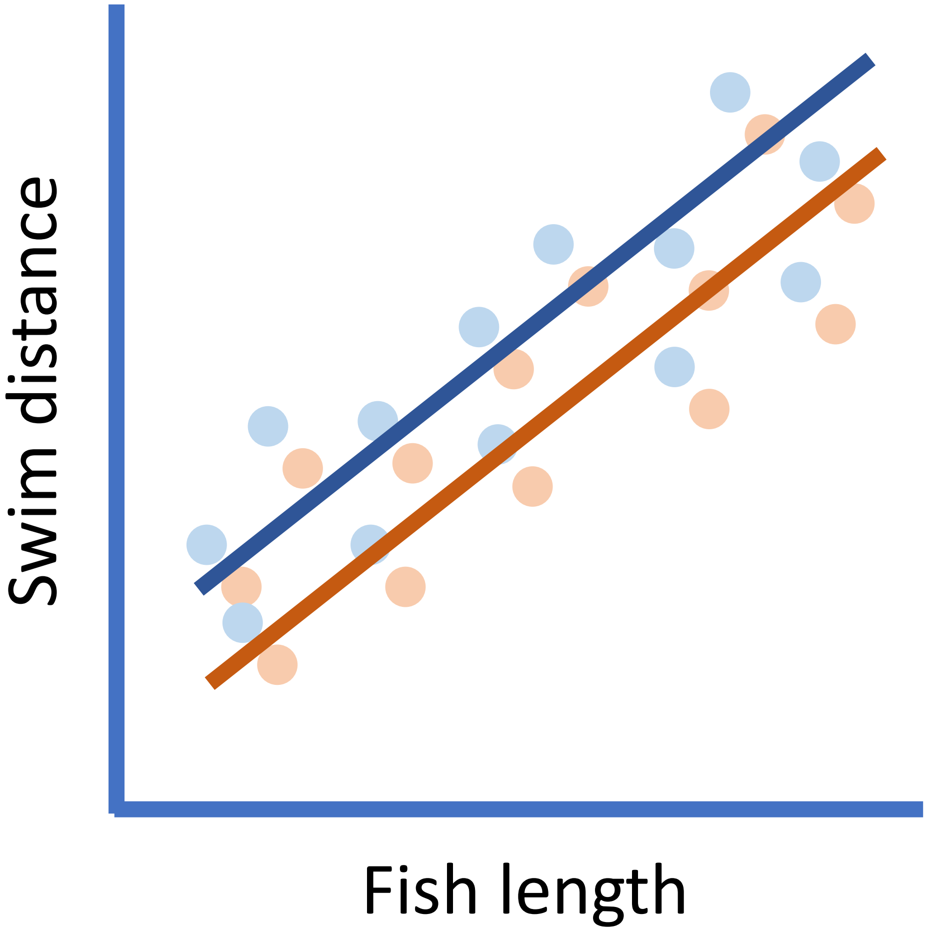
Boldface lower-case for vectors

Boldface upper-case for matrices

Level 3: mixed effects models


Swim distance
Fish length


Swim distance
Fish length



In words

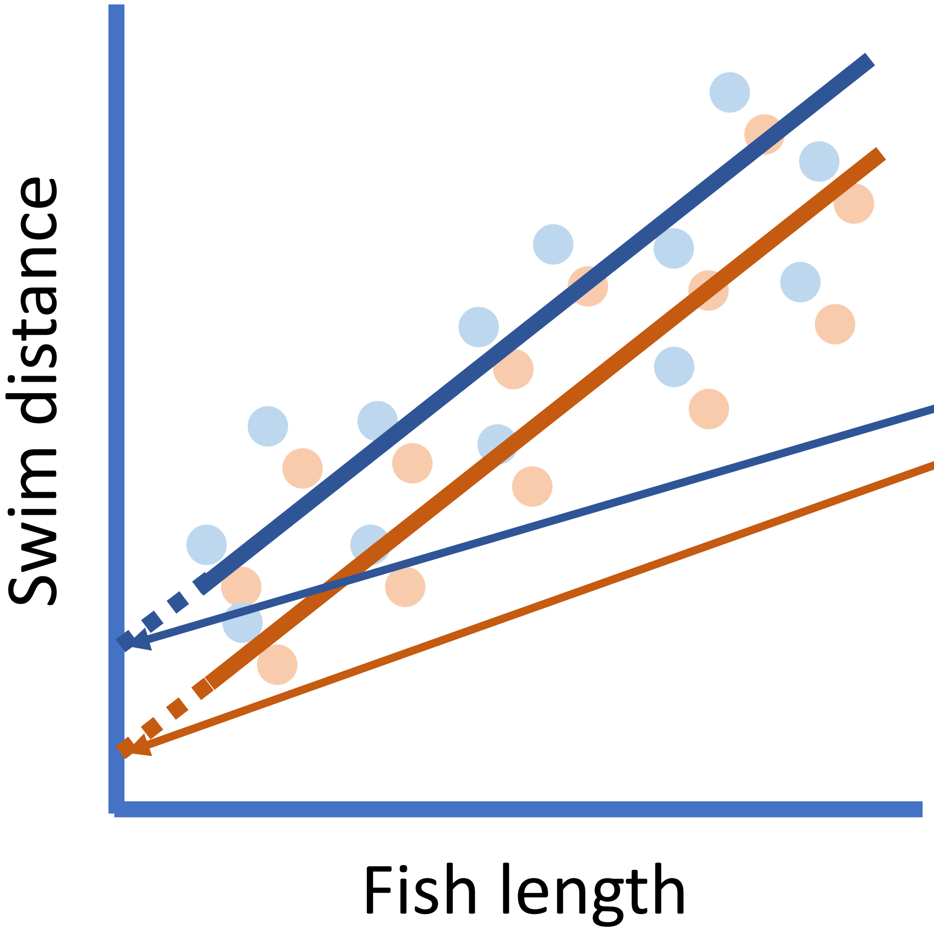
We fitted a linear regression to data on swim distance (response variable) and fish length (predictor variable) and included a random intercept for sampling location.

In R

lmer(swimDistance ~ fishLength + (1 | site))

or

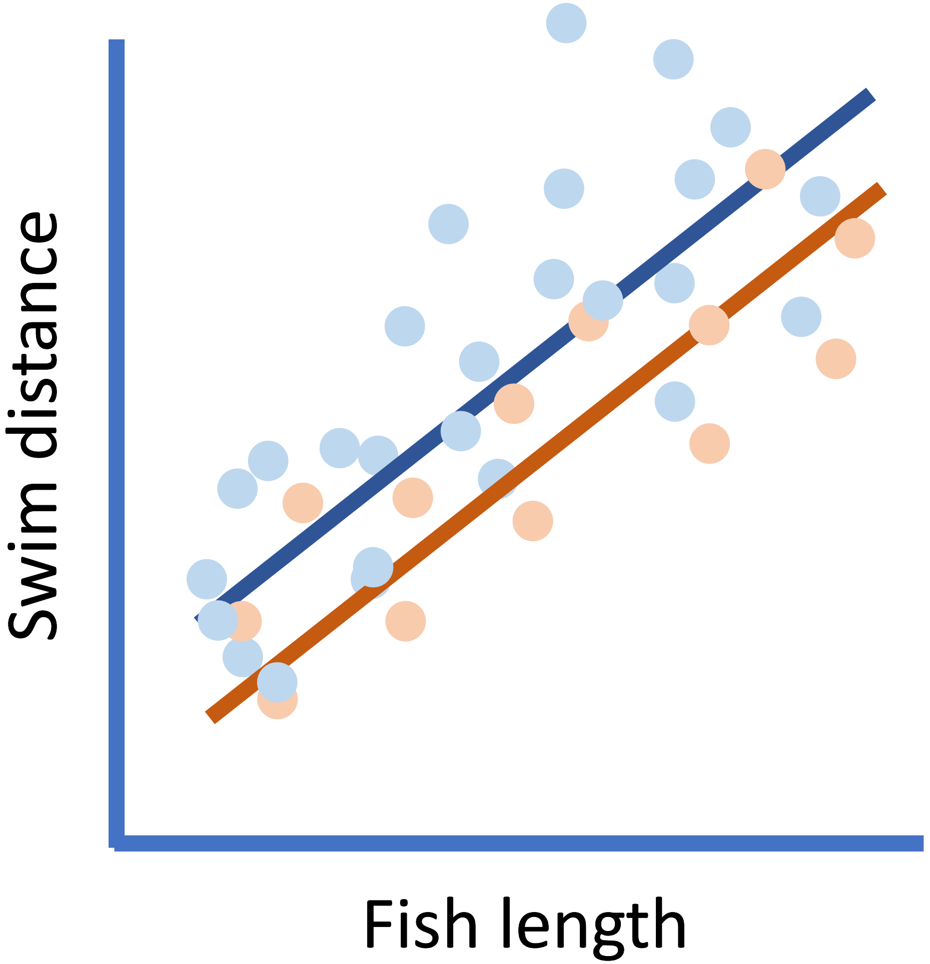
lmer(y ~ x + (1 | group))



As an equation

$$y_i = \alpha + \beta x_i + \gamma_{s(i)} + \varepsilon_i$$

where y_i is the distance fish i has swum, α is the swim distance of a fish of length zero, β is the association between the length of fish i (x_i) and swim distance, $\gamma_{s(i)}$ is a random intercept in site s(i) and ε_i is residual variation in swim distance of fish i.



In words

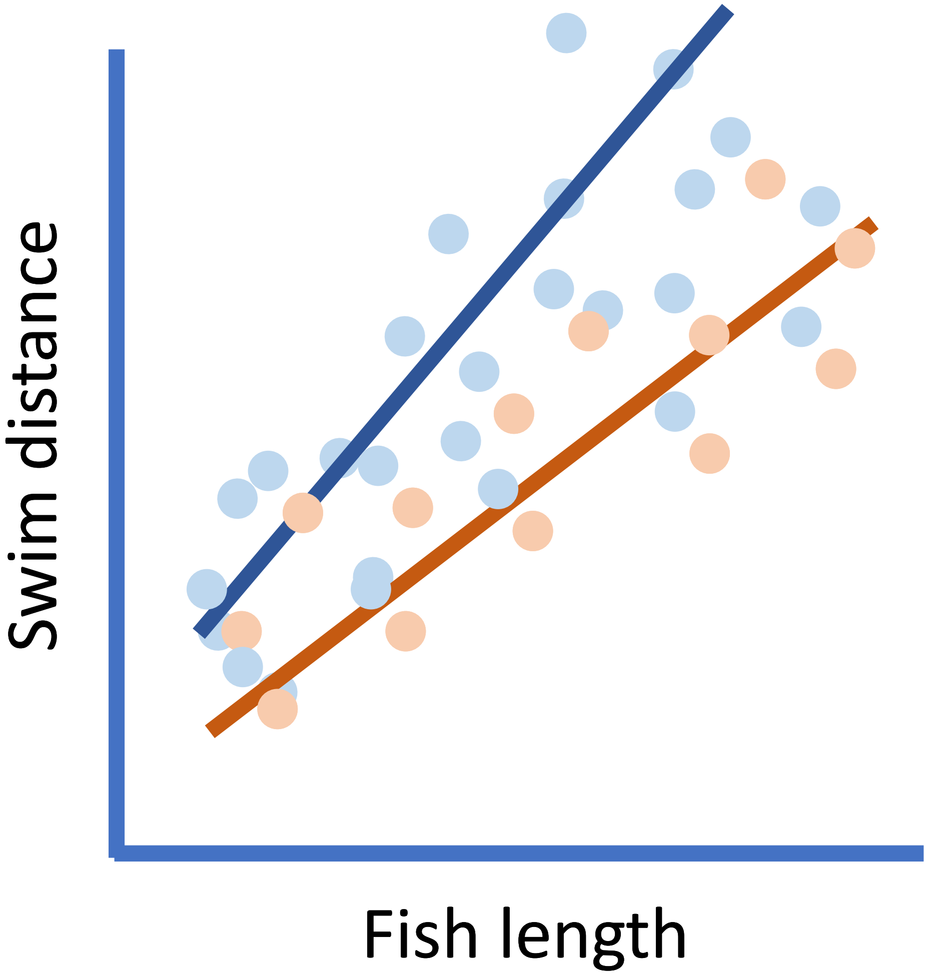
We fitted a linear regression to data on swim distance (response variable) and fish length (predictor variable) and included a random intercept and slope for sampling location.

In R

lmer(swimDistance ~ fishLength + (fishLength | site))

or

lmer(y ~ x + (x | group))



In words

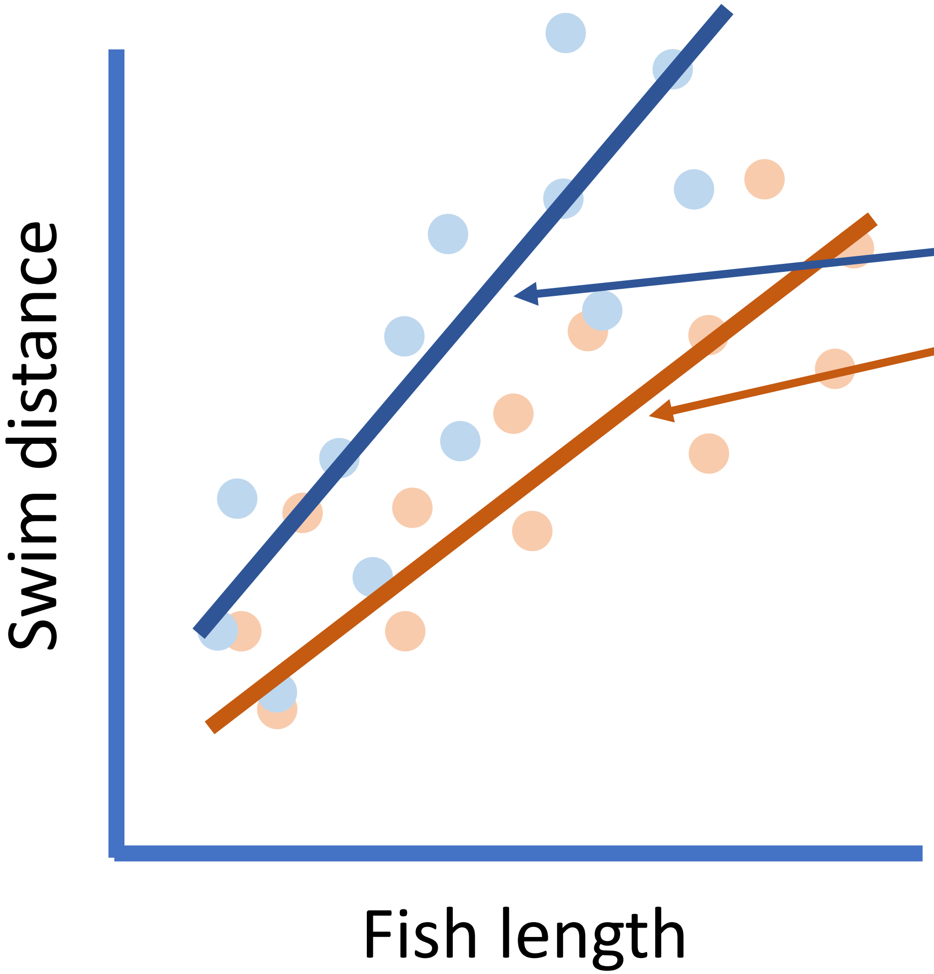
We fitted a linear regression to data on swim distance (response variable) and fish length (predictor variable) and included a random intercept and slope for sampling location.

In R

lmer(swimDistance ~ fishLength + (fishLength| site))

or

lmer(y ~ x + (x | group))



As an equation

$$y_i = \alpha + \beta x_i + \boldsymbol{\delta}_{s(i)} x_i + \gamma_{s(i)} + \varepsilon_i$$

where y_i is the distance fish i has swum, α is the swim distance of a fish of length zero, β is the association between the length of fish i (x_i) and swim distance, $\boldsymbol{\delta}_{s(i)}$ is a random association between fish length and swim distance in site s(i), $\gamma_{s(i)}$ is a random intercept in site s(i) and ε_i is residual variation in swim distance of fish i.

Why use equations for mixed models?

Equations are explicit

(”random effect” has several meanings[1])

Equations highlight similarities among models

(e.g., mixed effects, repeated measures, random intercept)

[1] http://andrewgelman.com/2005/01/25/why_i_dont_use/

Jian’s rules

Greek letters for parameters; Roman letters for data

Define all terms

Single letters where possible

Boldface lower-case for vectors

Boldface upper-case for matrices

Be consistent!

Take-home messages

Are equations really better?

Equations have their place. . .

. . . but so do written descriptions and R formulas

Equations might be hard to interpret but they’re equally hard to misinterpret

Level 4: do we still have time?

Some things to cover on the whiteboard

Nonlinear models: $y_i = \alpha + f(x_i) + \varepsilon_i$

BUGS/JAGS notation: $y \sim \text{Normal}(\mu, \sigma)$

Generalised models and link functions