Statistical models : linking data to theory

Presentation – a brief tour on statistical models

Sara Mortara

27 Jan 2025

Who we are (in order of appearance)

Biologists trying to help you to learn statistical modelling faster and easier than we did

Sara, Sara Mortara (she/her)

Diogro, Diogo Melo (he/him)

Who we are

Professors who also have contributed to this course

Andrea, Andrea Sanchez Tapia (she/her)

Paulinha, Paula Lemos da Costa (she/her)

PI, Paulo Inácio Prado (he/him)

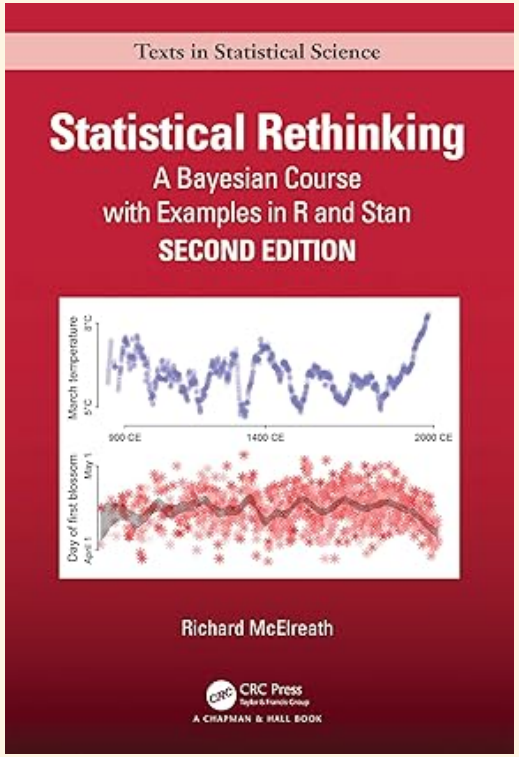
What we want

- To help you to be good users of statistical models in ecological research
- Kindness and persistence
  - To make you learn this faster and easier than we did
  - In doing that, to improve our own knowledge on statistical models

> "Ninguém educa ninguém, ninguém educa a si mesmo, os homens se educam entre si, mediatizados pelo mundo"
>
> Paulo Freire

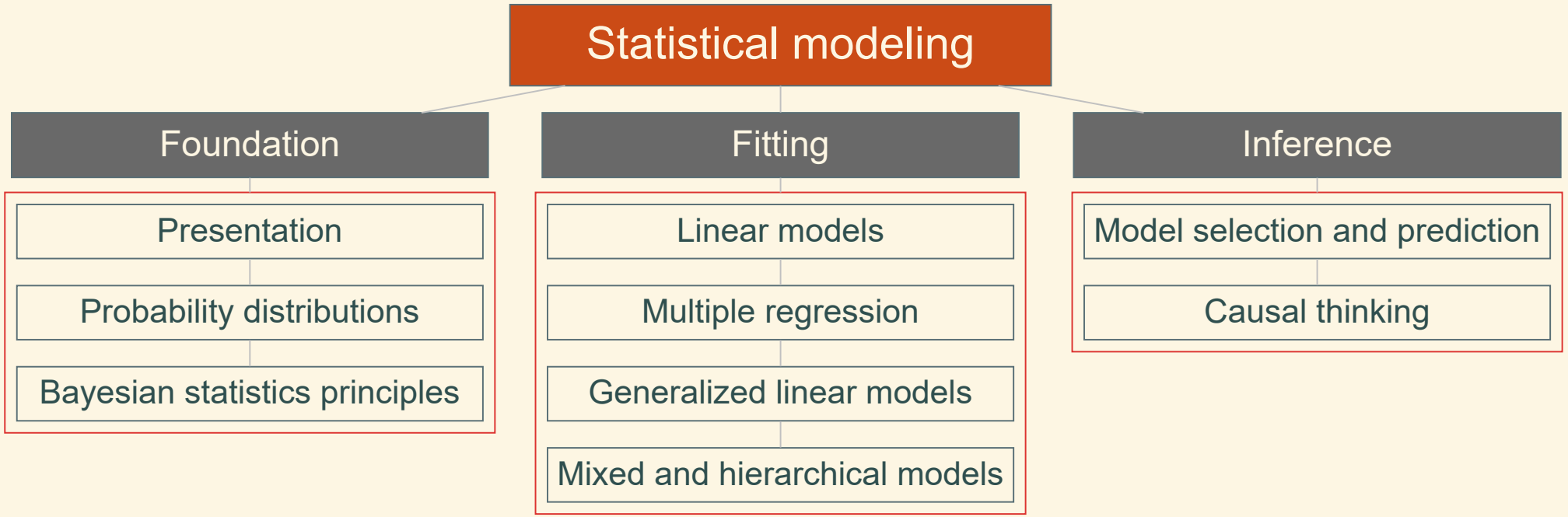
Emphasis on rethinking

broader lens to view aspects of statistical modeling



- emphasize the interplay between data, theory, and uncertainty
- incorporate prior knowledge and domain expertise
- tools for reasoning about model uncertainty and predictive accuracy
- we avoid significance testing -> we aim statistical clarity (sensu Dushoff et al 2019)

Course overview



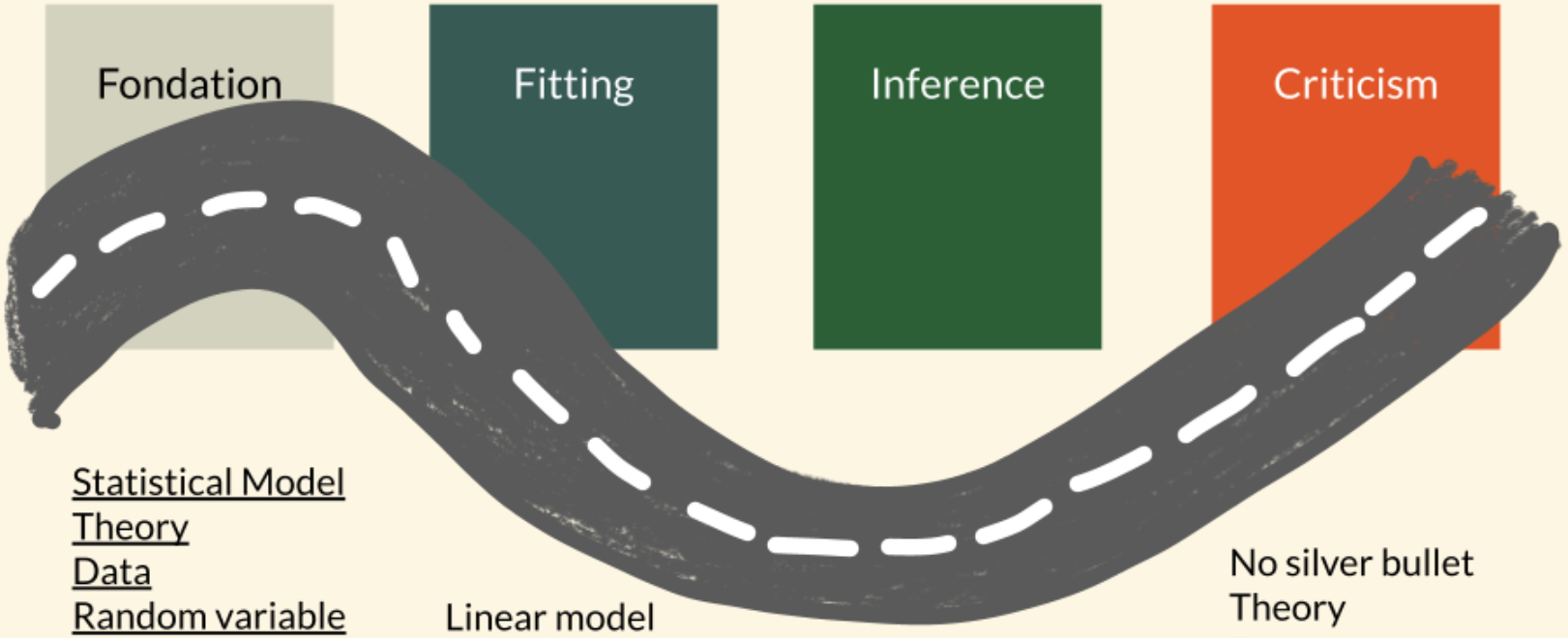
Teaching strategies

- Lectures and computer labs
- Labs: tutorials in R language
- RStudio server available for the course
- Q&A (in your preferred language)
- Office hours if possible (and if interest you)
- All resources at https://statistical-modeling.github.io/2025
- Learning resources will be available as we progress

Course schedule

Week	Time	Monday	Tuesday	Wednesday	Thursday	Friday
1	09:00-10:30	Course presentation	Extending Linear Models	Q&A	Generalized Linear Models - 2	Bayesian Statistics Principles
1	11:00-12:30	Simple Linear Models	Linead Models Lab	Generalized Linear Models - 1	GLM Lab	Q&A
2	14:00-15:30	Model selection, prediction and overfitting	Causal Thinking and model building	Causal Thinking lab	Mixed and hierachical models	Mixed and hierachical models lab
2	16:00-17:30	Bayesian statistics Lab		Q&A		Final Q&A

Course roadmap



Linking data and theory

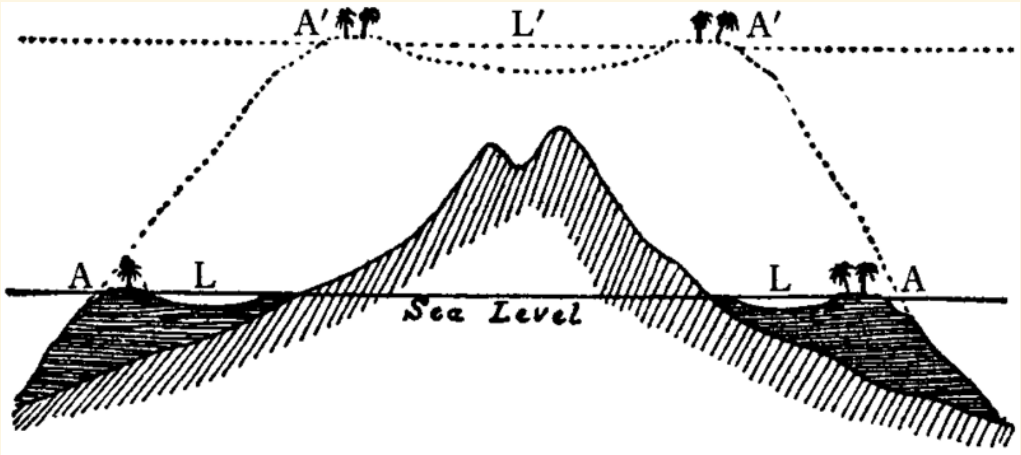
Ibã Huni Kuin, Nai Basa Masheri, 2014

How do we address ecological problems characterized by complexity and uncertainty?

- Use a **model** to simplify
- **Theories**: summarize the understanding of how
- **Hypotheses**: proposed explanations of how
- **Data**: measurable variables carrying uncertainty
- **Statistical models** confront theories with data

Use a model to simplify

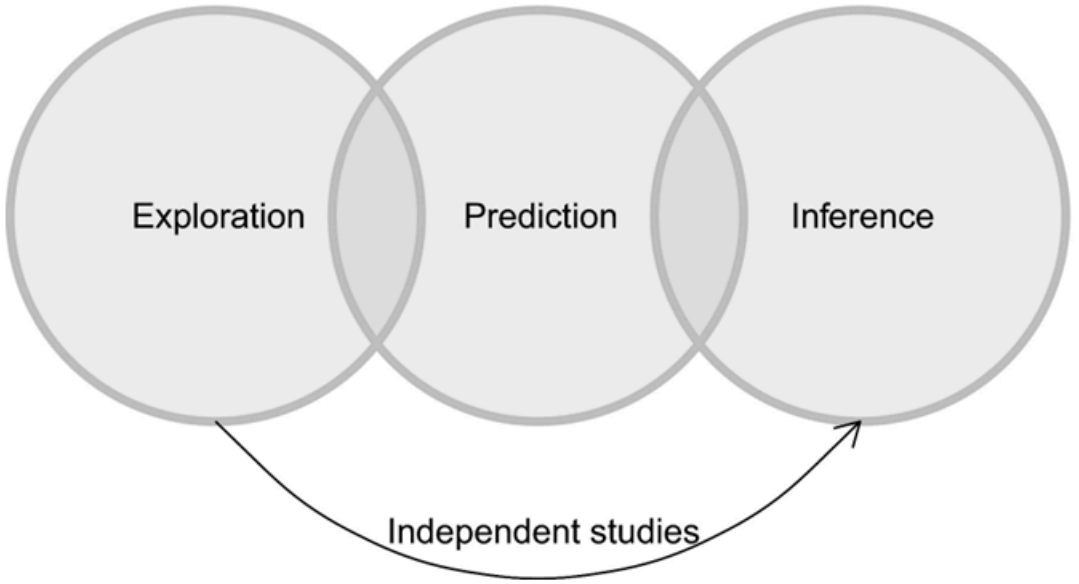
Darwin's theory of coral atolls



Hypothesis: Atolls represent a sequential stage in the lifecycle of coral reefs influenced by the interplay of island subsidence and coral growth.

Modeling goals in Ecology

A practical guide to selecting models for exploration, inference, and prediction in ecology



Ecology, Volume: 102, Issue: 6, First published: 12 March 2021, DOI: (10.1002/ecy.3336)

What is the relation between photosynthesis and forest growth?

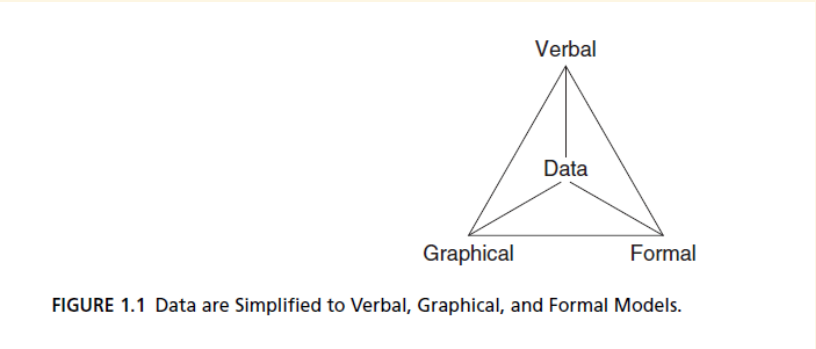


FIGURE 1.1 Data are Simplified to Verbal, Graphical, and Formal Models.

Carbon cycle of forests

- Relationship between photosynthesis, productivity, and biomass
- From early naturalists: assumptions that tall, majestic forests must be more productive that shorter forests
- Recent ecosystem theory: causal linking chain linking Gross Primary Productivity (GPP) to biomass -> Verbal model

What is a statistical model?

A description of how your data should behave according to some theory

Thus, statistical models are hypotheses about how a given data set was generated

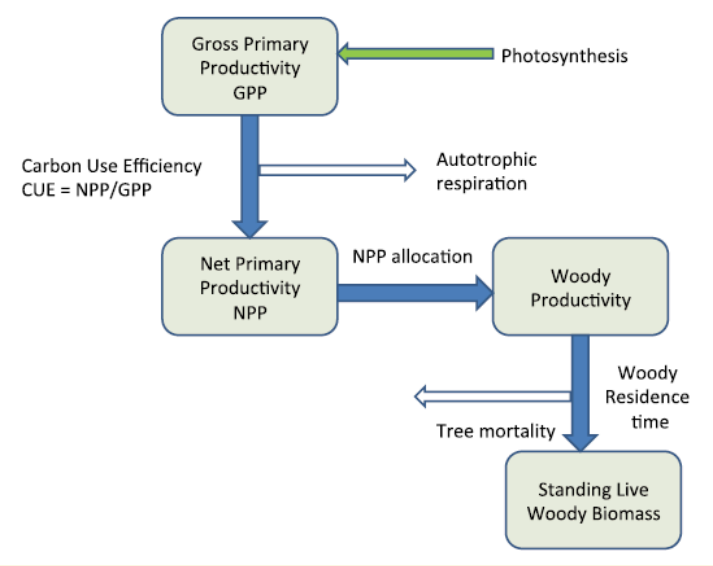
Why is this different from a mathematical model?

Mathematical models describe the behavior of theoretical quantities

Statistical models describe the behavior of measurements used to express theoretical quantities

What we expect from theory?

Carbon cycle of forests



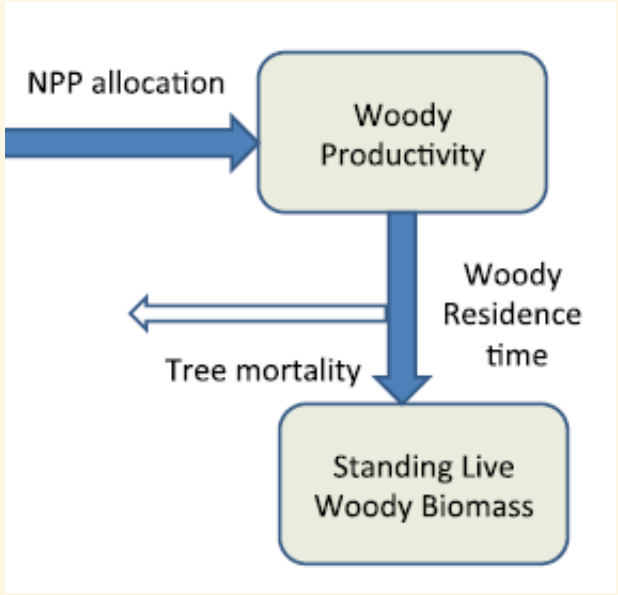
Malhi and collaborators 2012; 2015

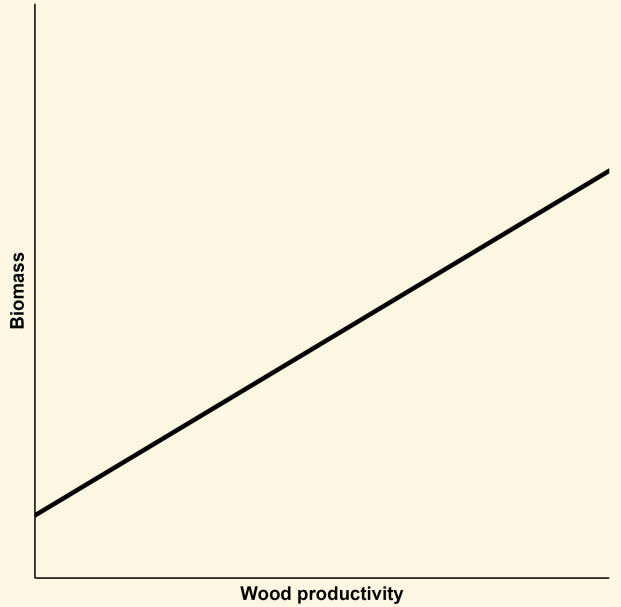
- GPP: total photosynthetic carbon fixation by the forest canopy
- Autotrophic Respiration: the fraction of GPP used for metabolic processes
- NPP: remaining carbon allocated to biomass production
- Mortality and residence time: high productivity forests show higher turnover and lower biomass residence times
- Biomass Accumulation: results from the interplay between productivity, allocation, and mortality

What we expect from theory?

Biomass in old-growth Amazonian forests

Malhi et al 2006 GCB

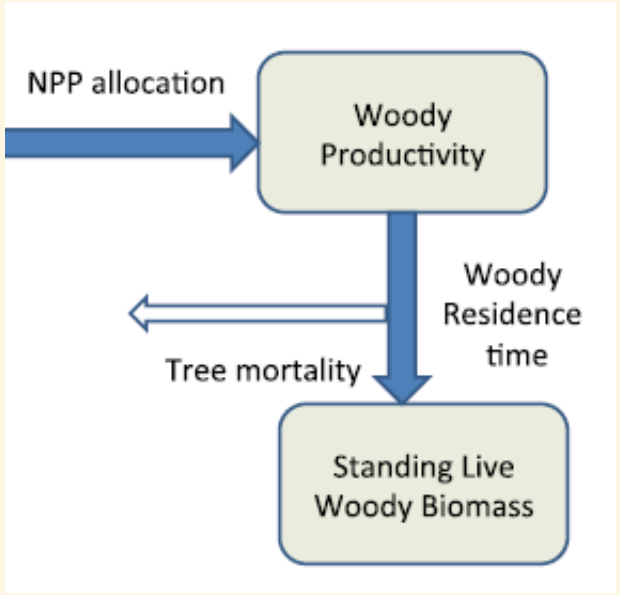




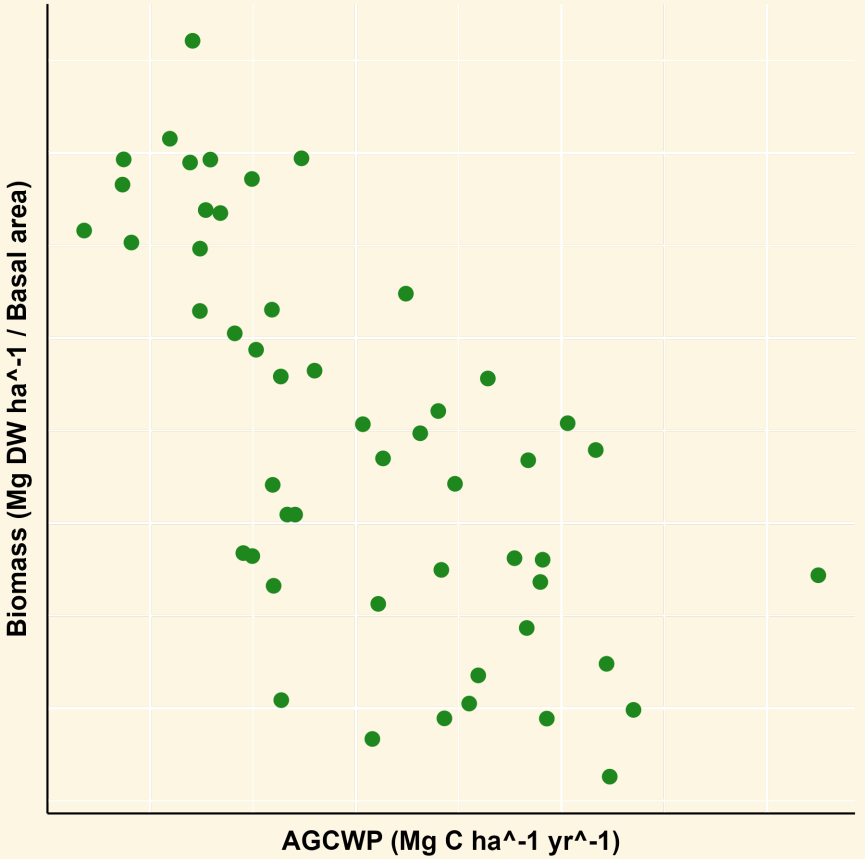
What we expect from theory?

Biomass in old-growth Amazonian forests

Malhi et al 2006 GCB



What the data shows

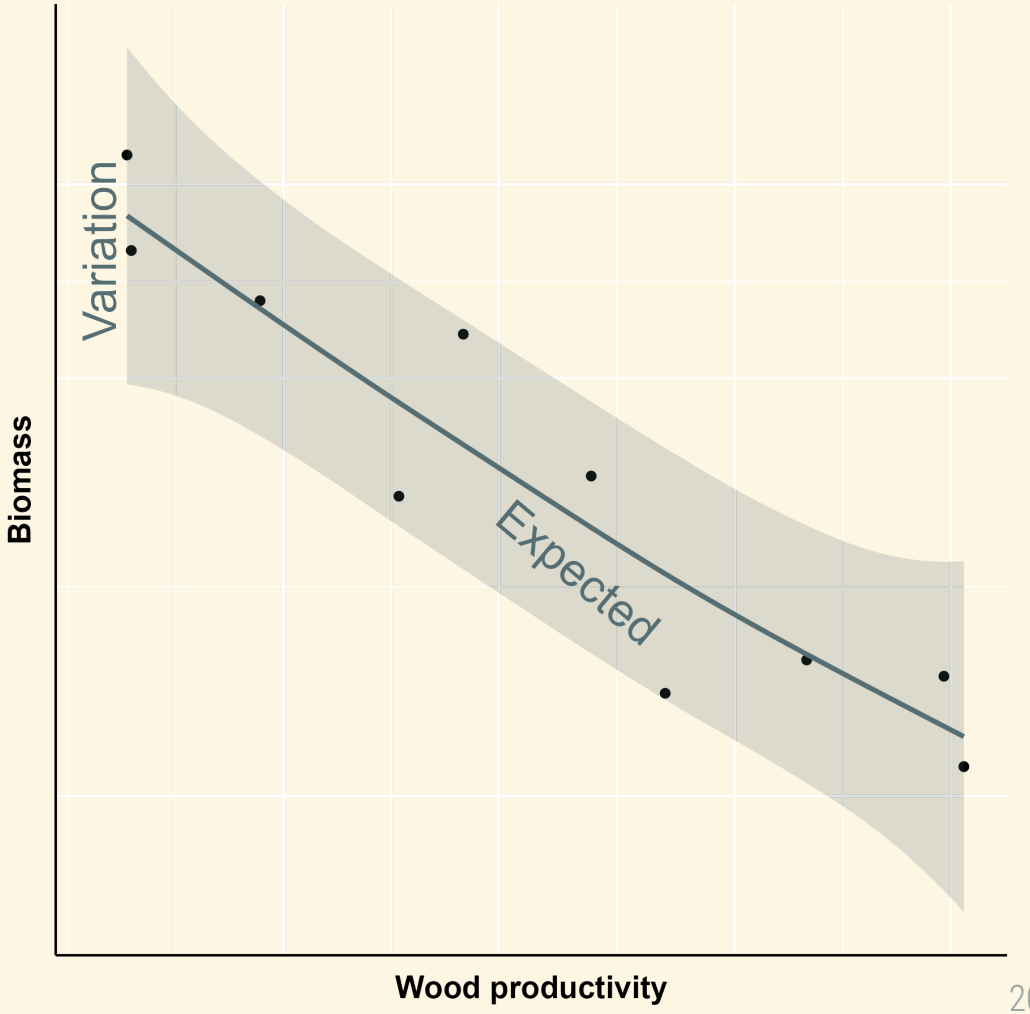


Empirical data, a.k.a. random variables

- We can never control all sources of variation that affect our measurements or counts
- Thus, any measurement or count done in a scientific research has some degree of uncertainty
- In other words, any quantitative data of scientific interest is a random variable

From the mathematical to the statistical model

- We can not assure the exact value of a random variable but we can think on how likely would be each possible value
- A simple hypothesis: large departures from the theoretical expectation are less likely than small ones
- In this approach data is partitioned into expected values and unexplained variation (a.k.a. *"residual"*, *"error"* or *"noise"*)
- $Data = mean(Y) + Residual$



The normal or Gaussian probability distibution

- Ascribes a probability to each value a measurement can assume
- It is a model for the sum of many sources of variation
- Small deviations from the expected value are more likely than large deviations
- The probabilities of the deviations are symmetrical around the expected value

The normal distribution is a statistical model

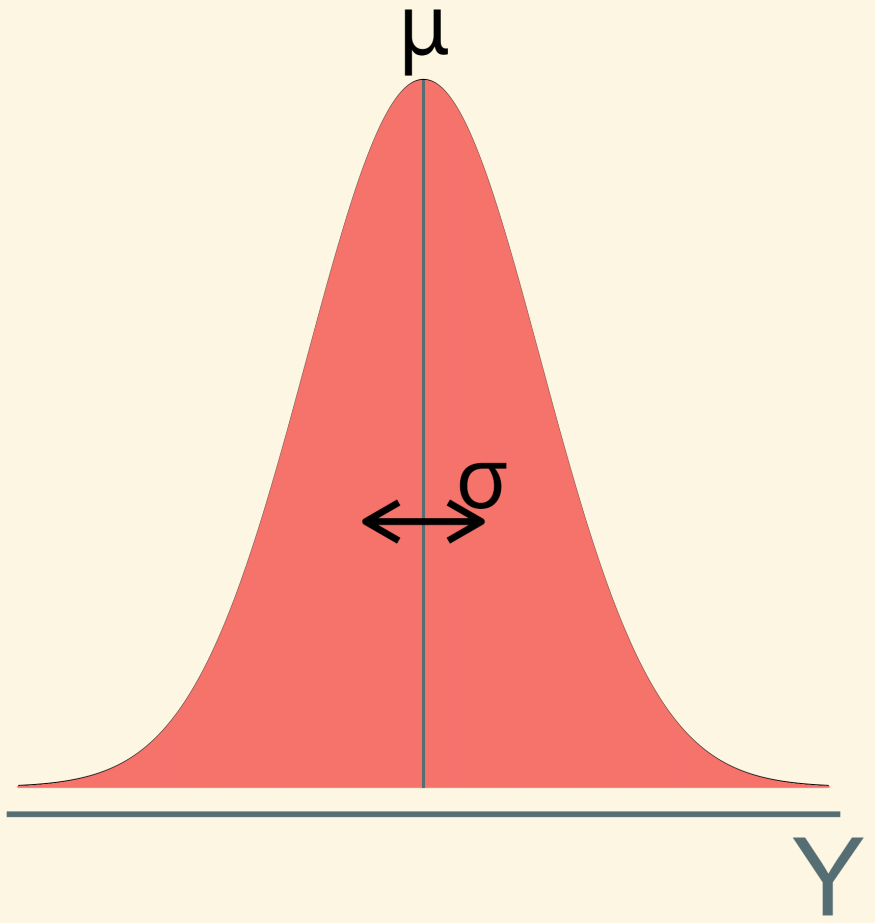
$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \, e^{-\frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2}$$

Where:

- Y: the value of a measurement
- μ: the expected value
- σ: the standard deviation

Notation:

- $Y \sim \mathcal{N}(\mu, \sigma)$ means "Y *is a random variable that follows a normal distribution*"



A syntax for statistical models

Our model:

$$Y_i \sim \mathcal{N}(\mu, \sigma)$$

$$\mu = \alpha + \beta * AGCWP$$

Usually:

$$\sigma = C$$

For us: σ is a random variable with its own distribution

In both cases, σ is always positive

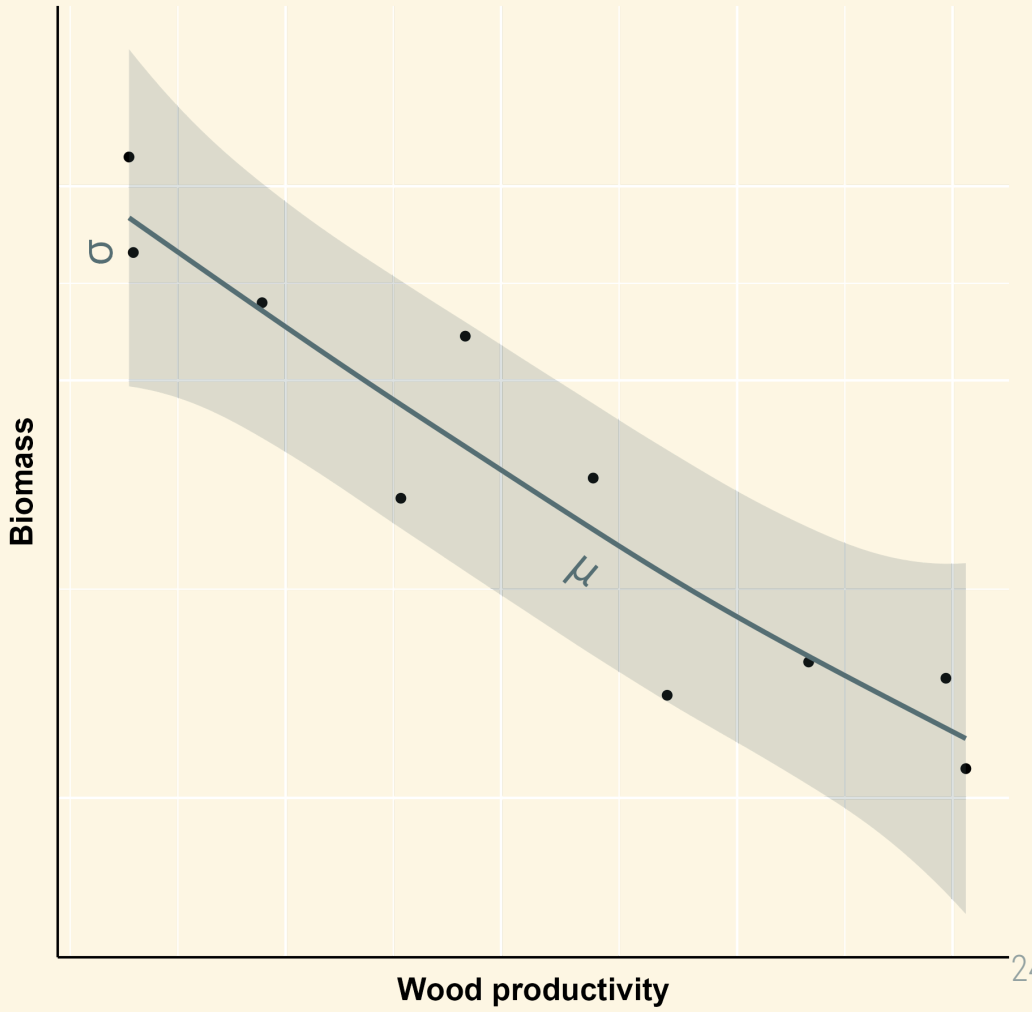
Which means: This model assumes that biomass per basal area Y_i is normally distributed around a mean value μ that is linearly related to above ground course wood productivity (AGCWP), and with a standard deviation σ following a particular distribution. The slope β quantifies how productivity affects biomass, while the intercept α represents the baseline biomass when AGCWP is zero.

What's next?

Model fitting: use the data to find the best guesses for the free parameters of the model (c, z)

Inference: Use the guessed parameters and the fitted model to learn about the study system and to make predictions

Model selection: fit alternative models and compare them



Linear regression: a model to rule them all

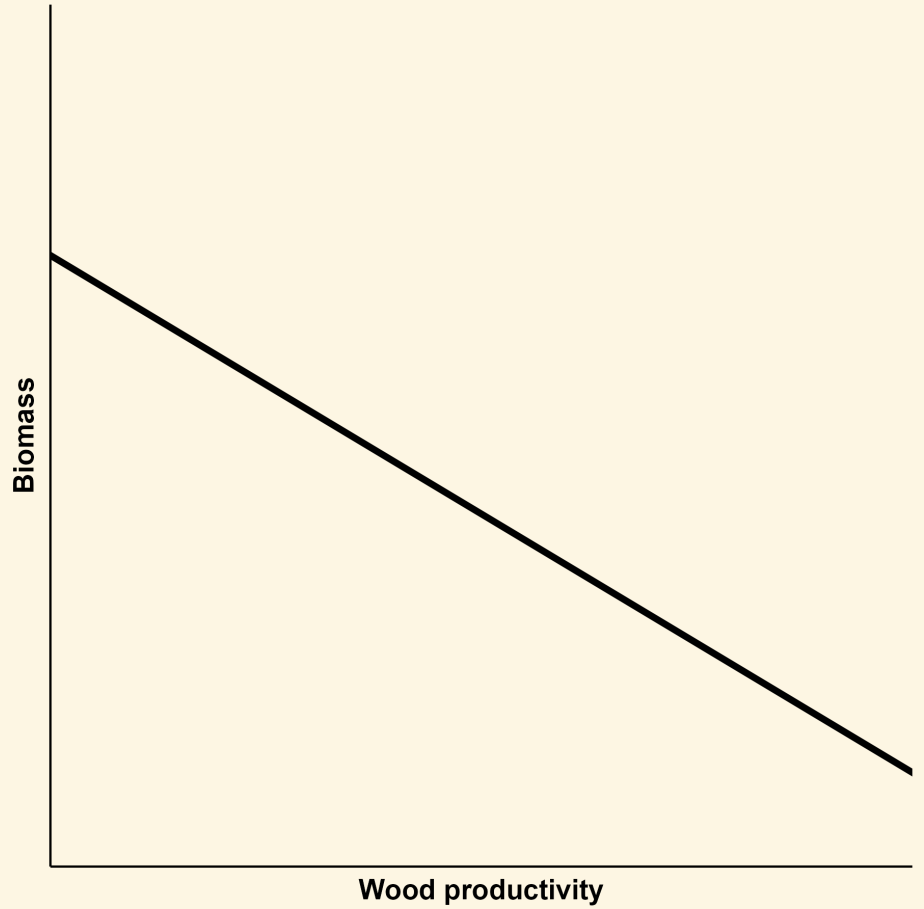
- All statistical model has a mathematical expression that describes the expected behavior of a response or dependent variable
- The simplest mathematical behavior is a linear relationship: the response is proportional to other measurements (known as a predictor or independent variables)
- Linear regression is the statistical model that describe proportionalities
- Linear regression is a foundational statistical model, from which a lot of other models were developed
- Thus, it is general a good idea to deduce linear relationships from your theoretical models and then fit a linear regression. Next we will show an example with the logistic equation

What we expect from theory

$$Y_i \sim \mathcal{N}(\mu, \sigma)$$

$$\mu = \alpha + \beta \times AGCWP$$

The biomass is a linear function of the woody productivity, and the slope would depend on the effect of mortality



Which data do we need?

- The biomass in a forest plot, e.g. biomass per basal area -> response variable
- Wood productivity, e.g. aboveground coarse wood productivity -> predictor variable

```
head(df)
```

```
##      AGCWP  Biomass
## 1 1.678107 11.58105
## 2 1.864691 11.82921
## 3 1.870414 11.96556
## 4 2.095071 12.07726
## 5 2.205525 12.60637
## 6 2.193376 11.94865
```

The linear regression model

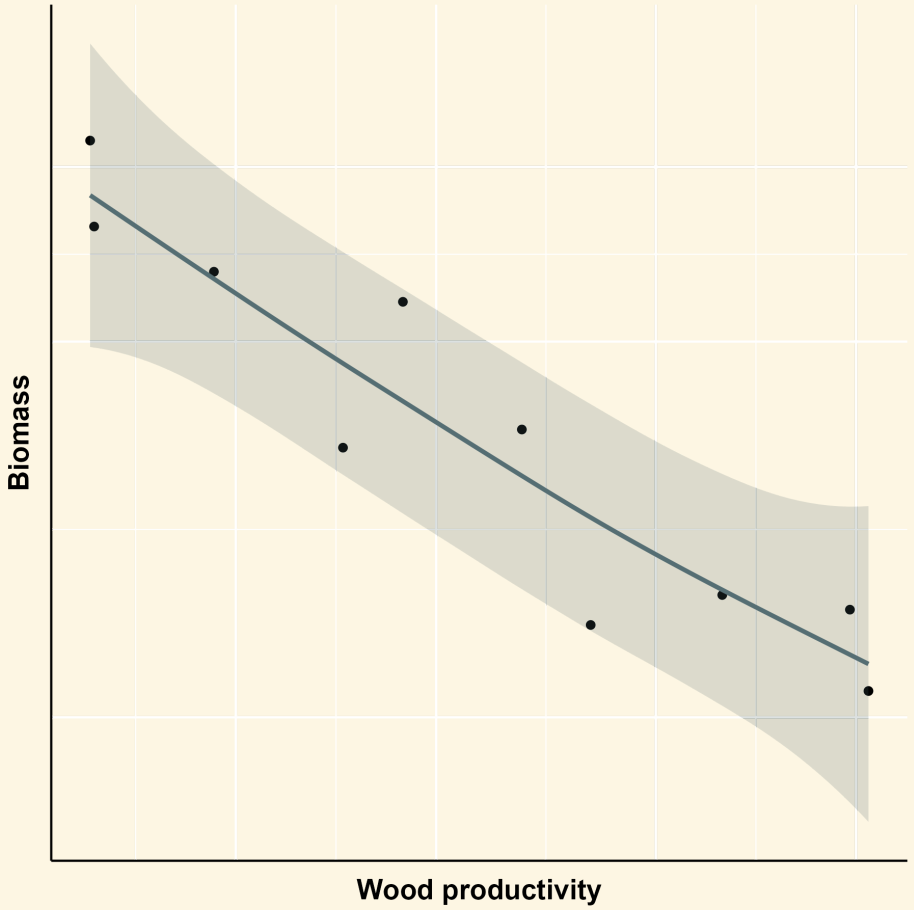
Our model:

$$Y_i \sim \mathcal{N}(\mu, \sigma)$$

$$\mu = \alpha + \beta \times AGCWP$$

$$\sigma \sim \mathcal{E}xponential(1)$$

Which reads: The biomass per basal area across plots $Y(i)$ follows a normal distribution with expected value μ that is a linear function of the coarse woody productivity of each plot, with a standard deviation σ



Implementing the model

Define the model

```
# Define a linear model of the biomass as a
# linear function of the WP
model <- alist(
  Biomass ~ dnorm(mu, sigma),
  mu <- a + b * AGCWP,
  a ~ dnorm(10, 5),  # Prior for intercept
  b ~ dnorm(0, 1),   # Prior for slope
  sigma ~ dexp(1)    # Prior for residual SD
)
```

Fit the model

```
fit <- quap(
  model,
  data = df
)
```

Estimated parameters

```
summary(fit)
```

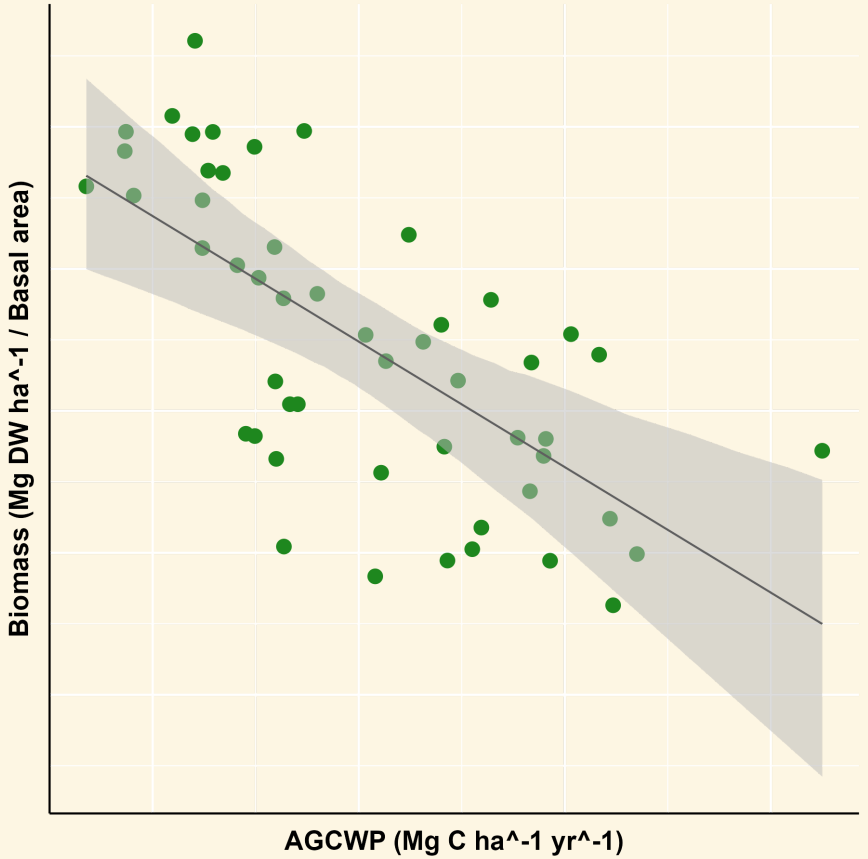
```
##              mean         sd       5.5%      94.5%
## a     13.1443395 0.41481859 12.4813792 13.8072997
## b     -0.8853512 0.13259524 -1.0972640 -0.6734384
## sigma  0.7524177 0.07372921  0.6345841  0.8702512
```

$$\alpha = 13.14$$

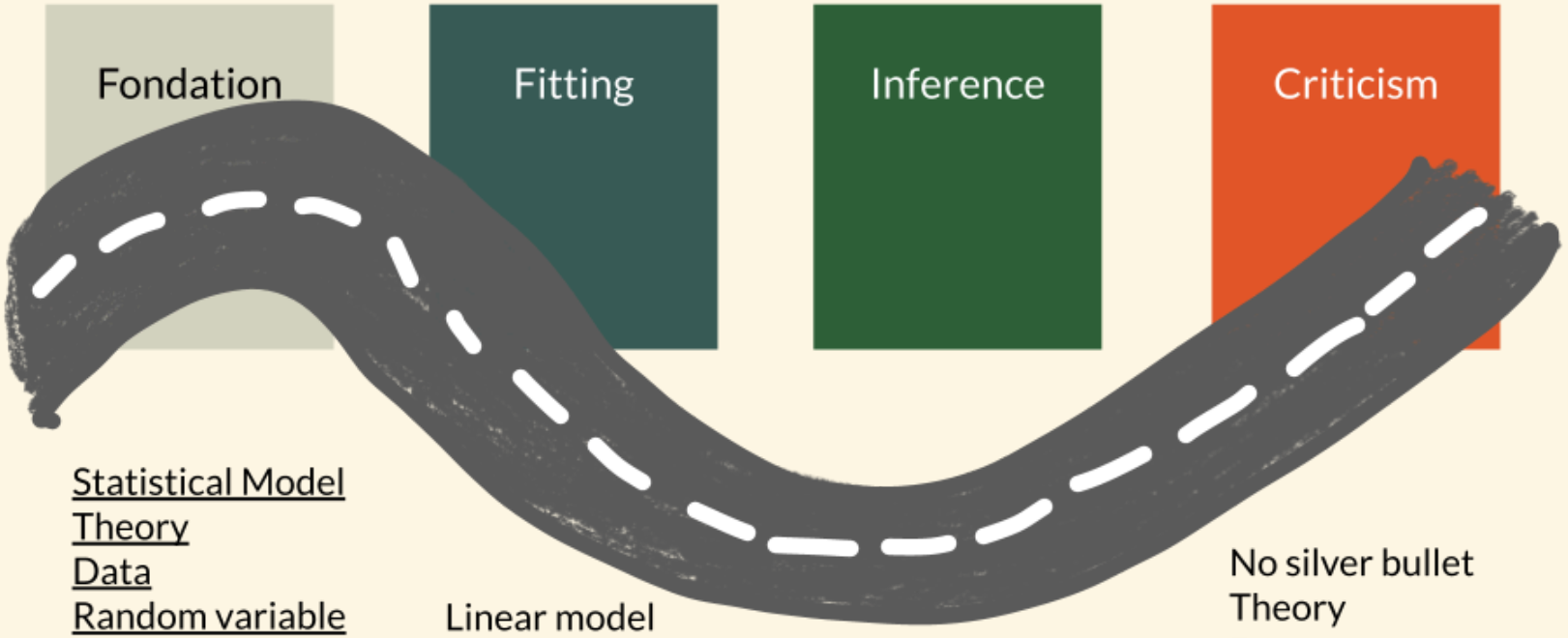
$$\beta = -0.89$$

$$Y_i = 13.14 + (-0.89 \times 3)$$

$$Y_i = 10.47$$



Where we are and what is next?



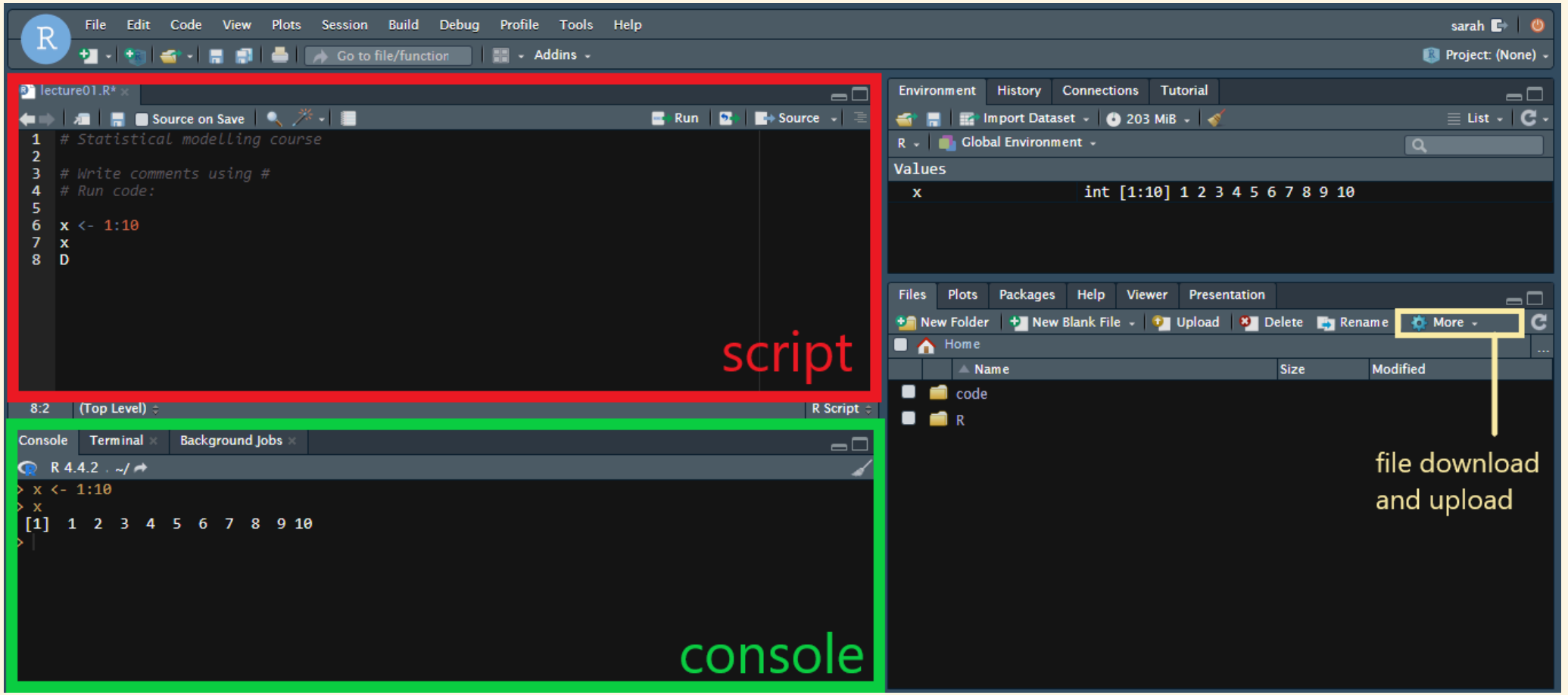
Statistical Model
Theory
Data
Random variable
Inference
Causality

Linear model
Multiple regression
GLMs
(G)LMMs
DAGs
Data simulation

Estimate uncertainty
Effect size
Causal effect
Significance
Model selection

No silver bullet
Theory
Data
Model diagnostics

Computational resources



Further reading

- Quantitative Reasoning in Ecology: Chapter 1 in Schneider, D. 2009. Quantitative Ecology: Measurement, Models, and Scaling. New York, Academic Press, 2nd Ed.
- Alternative views of scientific method and of modelling: Chapter 2 in Hilborn, R & Mangel, M. 1997. The Ecological Detective - Confronting Models with Data. Princeton, Princeton University Press.
- Approaches to ecological modelling: Chapter 1 in Ovaskeinen et al. 2016. Quantitative Ecology and Evolutionary Biology. Oxford, Oxford University Press.
- The golem of prague: Chapter 1 in McElreath. Statistical rethinking. Boca Ratol, FL, CRC Press.