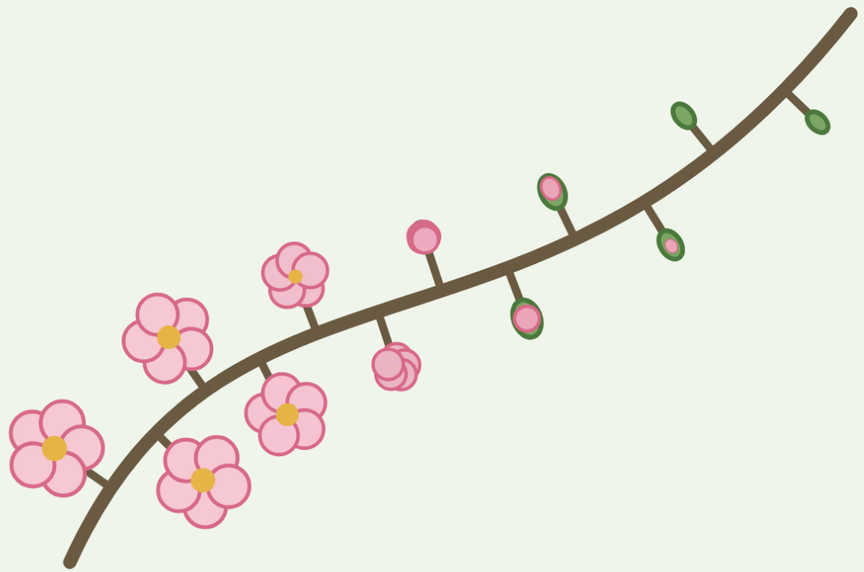
BIO144 · FS 2027 · UNIT 11, PART 2

Expand your mind: what next?

Data analysis beyond BIO144

Prof. Owen Petchey · University of Zurich



BIO144

LOOK HOW FAR YOU'VE COME

What you can already analyse

UNITS 3–7

Linear models

Data that meet their assumptions: regression, ANOVA, several explanatory variables, interactions.

UNITS 8–9

GLMs

Count data and binary data.

UNIT 11

Mixed models

Grouped or repeated-measures data.

UNIT 10

Ordination

Many variables at once.

That's a lot. And there is much more to learn.

THINK · PAIR · SHARE

What kind of data, or question, might you meet that BIO144 can't handle?

Klicker

BEYOND BIO144

Twelve situations, twelve families of methods

Tap one to jump to it

I have a series of values through time

1 · TIME SERIES ANALYSIS

My data points are in space, and I know their locations

2 · SPATIAL ANALYSIS

The relationship I want to fit is curvy

3 · NONLINEAR REGRESSION

The relationship may have a discontinuity

4 · BREAKPOINT ANALYSIS

I don't know what function, but I want one

5 · GENERALISED ADDITIVE MODELS

I have a network, a system of variables

6 · STRUCTURAL EQUATION MODELLING

I just want to predict…

7 · MACHINE LEARNING

I'm analysing the results of many studies

8 · META-ANALYSIS

My response is the time until something happened

9 · SURVIVAL ANALYSIS

I can't fix my residuals

10 · NONPARAMETRIC METHODS

I want my planned experiment to have enough power

11 · POWER ANALYSIS

I have some prior information

12 · BAYESIAN STATISTICS

Not an exhaustive list. For each: what's the situation, what's the method, and where to start in R.

SITUATION

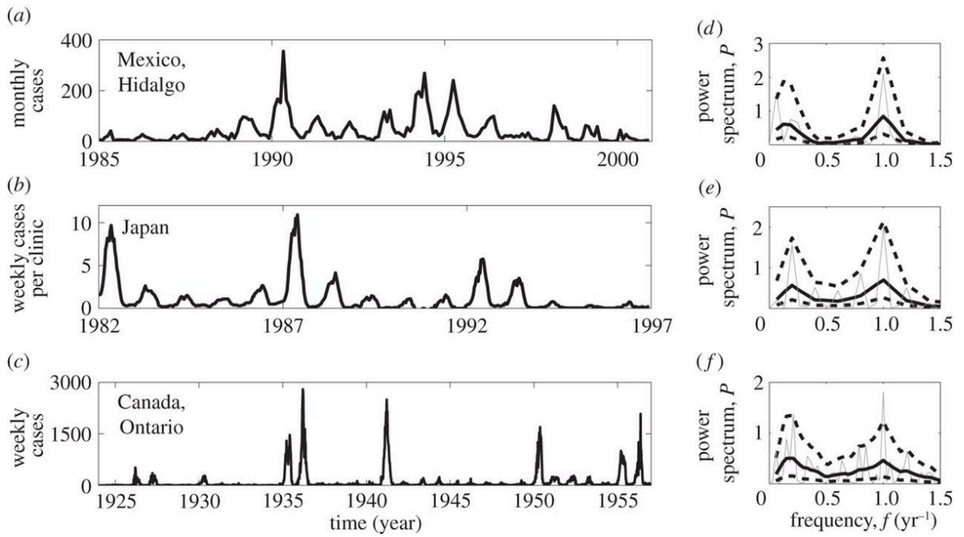
I have a series of values through time

Time series analysis

Observations close in time are usually similar: they are **not independent**, which breaks an assumption of most standard methods. Time series methods (e.g. ARIMA, seasonal decomposition, forecasting) handle this.

In R: `acf()` `forecast` `rEDM`

Example | **Try it**



Disease cases through time. Rozhnova, G., Metcalf, C.J.E. & Grenfell, B.T. (2013) Characterizing the dynamics of rubella relative to measles: the role of stochasticity. *J. R. Soc. Interface*, 10.

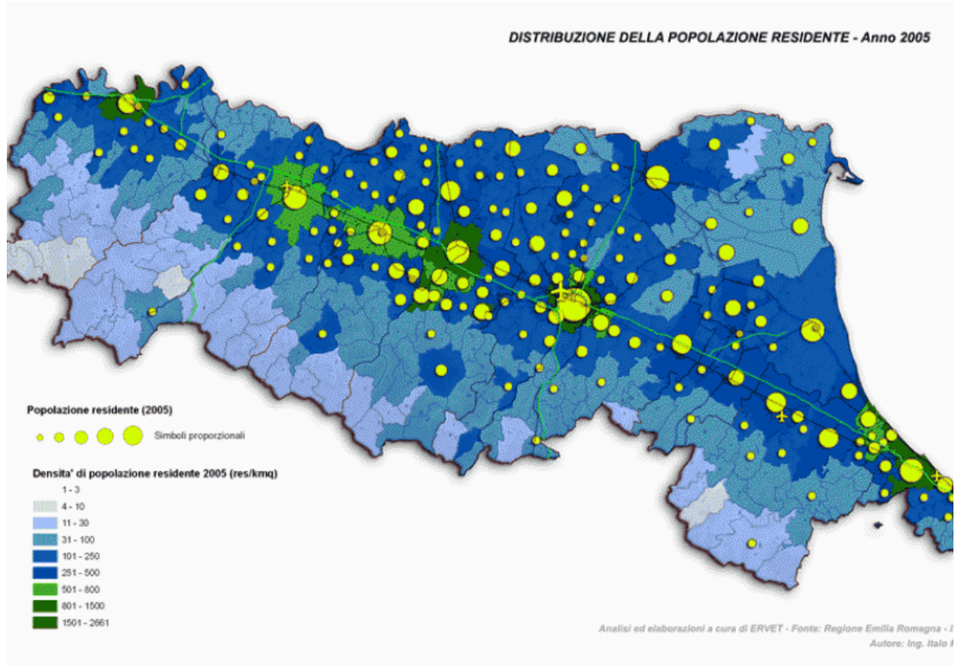
SITUATION

My data points are distributed in space, and I have their locations (e.g. latitude and longitude)

Spatial analysis

Nearby points tend to be similar (spatial autocorrelation), so again observations are **not independent**. Methods include spatial regression, kriging and geostatistics.

In R: the `Spatial` CRAN Task View



Points with locations: resident population across a region.

SITUATION

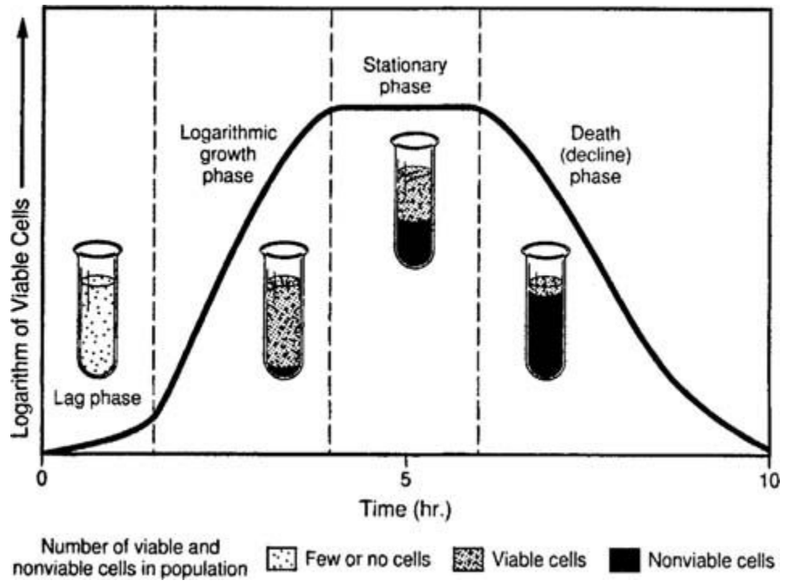
The relationship I would like to fit between two continuous variables is curvy!

Nonlinear regression

Estimate the parameters of a specific non-linear function, such as logistic growth, $f(x) = L / (1 + e^{-k(x-x_0)})$, and assess how well it fits.

In R: `nls()` (base R)

Example | **Fit it by eye**



Phases of growth of a bacterial population.

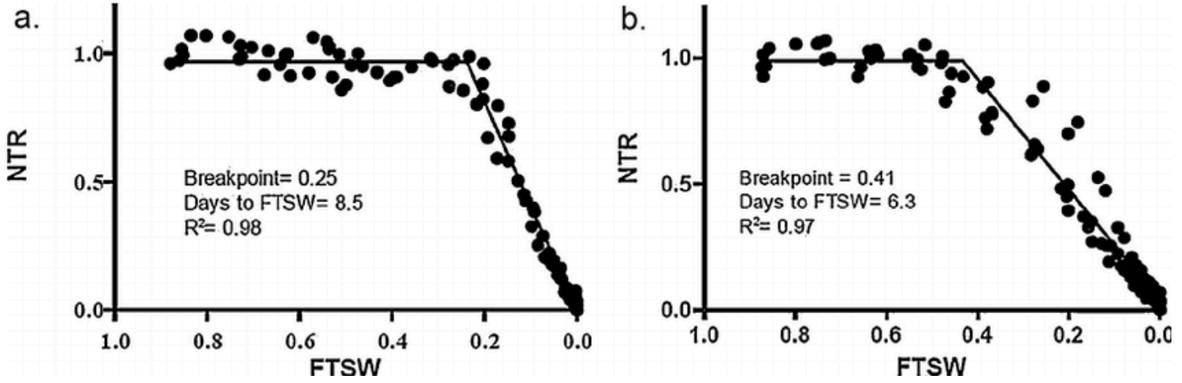
SITUATION

The relationship between two continuous variables may have a discontinuity!

Segmented, piecewise or breakpoint regression

The relationship changes at some point, e.g. before and after an intervention, or once a threshold is passed. Estimate where the break is, and the relationship on each side.

In R: `segmented`



Normalised transpiration ratio vs fraction of transpirable soil water. Fuentealba, M.P. et al. (2016) Transpiration responses of warm-season turfgrass in relation to progressive soil drying. *Scientia Horticulturae*, 198, 249–253.

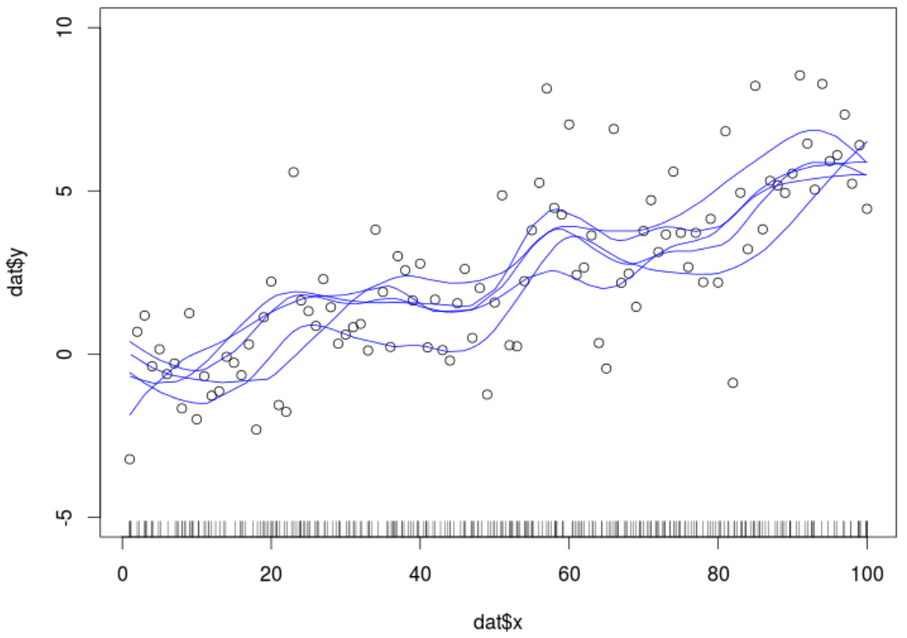
SITUATION

I don't know, and don't care, what function to fit between my two continuous variables, but I do want one!

Generalised additive models (GAMs)

Smooth functions capture non-linear effects, while the model stays interpretable. They extend the GLMs you know. Rather elegant!

In R: `mgcv`



A smooth fitted through noisy data, with its uncertainty.

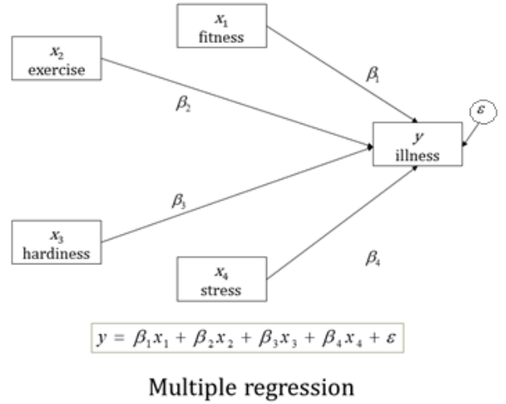
SITUATION

I have a network, a system, of variables

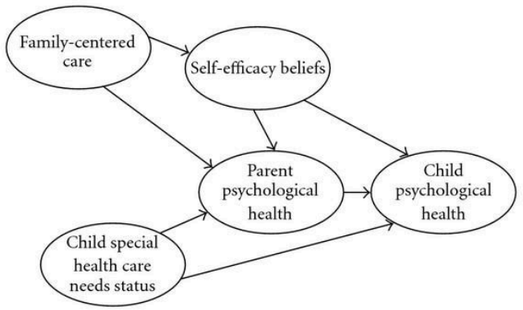
Path analysis and structural equation modelling (SEM)

Variables can be explanatory *and* response at the same time. Model direct and indirect effects, even latent variables: build and test causal models.

In R: `sem` `lavaan`



Multiple regression: every arrow goes to one response.



A path model: some variables are both causes and effects.

SITUATION

I just want to predict...

Machine learning

Decision trees, random forests, support vector machines, neural networks... They make predictions or classify, can handle large datasets and complex relationships, but may sacrifice interpretability.

```
install.packages(c(
  "ranger",   # random forests
  "xgboost",  # gradient boosting
  "glmnet",   # lasso/ridge/elastic net
  "kernlab",  # support vector machines
  "kknn",     # k-nearest neighbours
  "earth",    # adaptive regression splines
  "nnet"      # simple neural networks
))
```

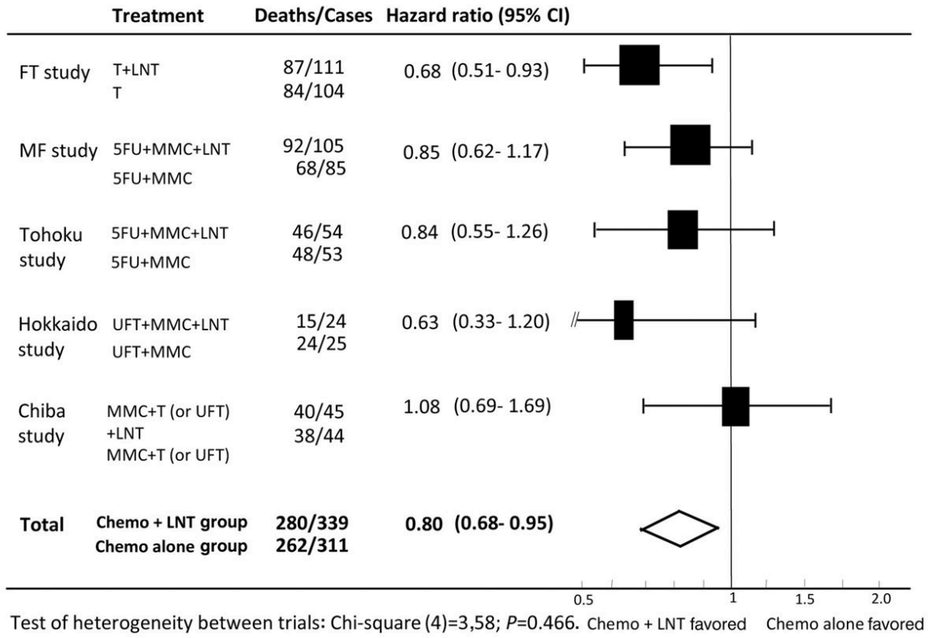
SITUATION

I'm reviewing and analysing the results of many studies

Meta-analysis

Combine the effect sizes from different studies to draw broader conclusions, and assess how much they differ (heterogeneity). Another reason to report effect sizes (Principle 3).

In R: `metafor` (and others)



A forest plot: one estimate per study, and the combined estimate (diamond). Oba, K. et al. (2009) Individual patient based meta-analysis of lentinan for unresectable/recurrent gastric cancer. *Anticancer Research*, 29, 2739–45.

SITUATION

My response variable is the amount of time until something happened

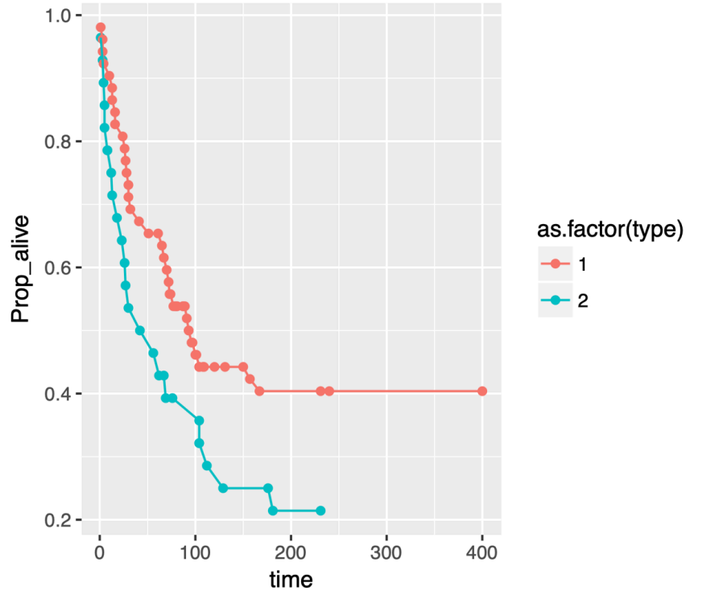
Survival analysis

Time-to-event data: time until death, failure, germination, flowering… Some individuals haven't had the event by the end of the study (**censoring**). Kaplan–Meier curves, Cox proportional hazards models.

In R: `survival` (and others)

type: tumour DNA profile (1 aneuploid, 2 diploid) · **time**: weeks to death or end of study · **delta**: 0 alive, 1 dead

type	time	delta
1	1	1
1	3	1
1	3	1
1	4	1
1	10	1
1	13	1
1	13	1
1	16	1
1	16	1
1	24	1
1	26	1
1	27	1
1	28	1
1	30	1
1	30	1
1	32	1
1	41	1



Proportion alive through time, for the two tumour types.

SITUATION

I can't fix my residuals

Nonparametric methods

Rank-based tests, bootstrapping and permutation tests make fewer assumptions about the distribution of the data.

In R: base R, e.g. `kruskal.test()` (and other packages)

Kruskal–Wallis is somewhat equivalent to a **one-way ANOVA**, on the ranks of the data.

But first: can you make a model that fits, e.g. a transformation, or a GLM with a suitable distribution (Principle 2)?

SITUATION

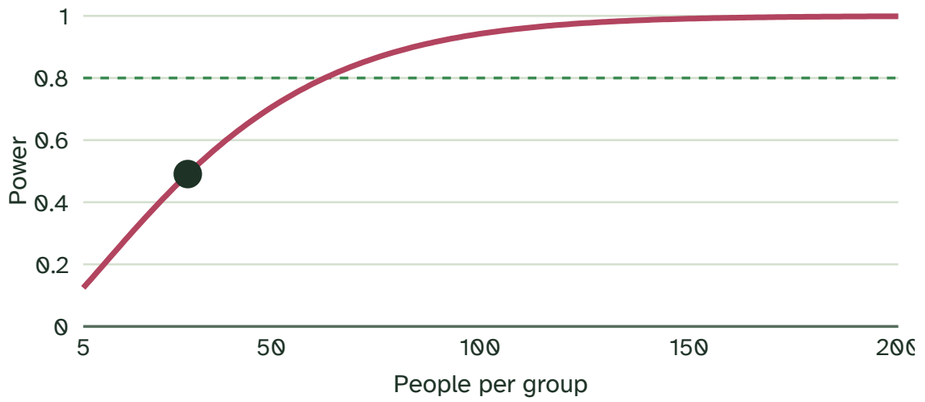
I want to make sure my planned experiment has adequate statistical power

Power analysis

Power is the probability of detecting an effect if it is really there. It depends on the **effect size**, the **variation**, the **number of observations** and the **significance level**. Use it to work out the sample size you need, before you collect data.

In R: `pwr`

TRY IT: COMPARING TWO GROUPS



Effect size (d) 0.50

People per group 30

α = 0.05 α = 0.01

POWER

49%

PER GROUP FOR 80% POWER

63

d = difference in means divided by the standard deviation.
Normal approximation to a two-sample t-test.

SITUATION

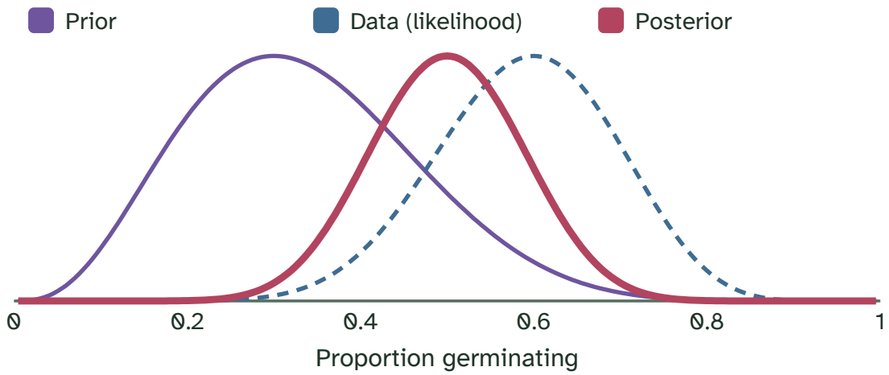
I have some prior information

Bayesian statistics

Combine prior knowledge with the data, using Bayes' theorem: `P(H | E) = P(E | H) P(H) / P(E)`. The result, the posterior, describes what we believe after seeing the data, with its uncertainty.

In R: `rstan` `R-INLA` `rjags` · Book: *Statistical Rethinking* (McElreath)

TRY IT: WHAT PROPORTION OF SEEDS GERMINATE?



Prior guess 0.30

Prior strength 10

Seeds sown 20

Germinated 12

Posterior mean **0.50**, 95% credible interval **0.33–0.67**.
Curves scaled to the same height.

Some methods are not mutually exclusive

Combined

Repeated measures through time: mixed models *and* time series. Counts across space: GLMs *and* spatial methods.

Extended

GAMs extend GLMs. Mixed models can be GLMs (GLMMs).

Bayesian versions

Bayesian methods can be used for most of them.

And there are multiple books about each.

YOUR TURN

Which family of methods?

1. Daily counts of a bird species at one site over 20 years. You want to forecast next year.

- Time series analysis
- Meta-analysis
- Power analysis

2. Time until seeds germinate, under three treatments. Some seeds had not germinated when the experiment ended.

- Nonparametric methods
- Survival analysis
- Structural equation modelling

3. You collected 40 published estimates of the effect of warming on plant growth.

- Machine learning
- Spatial analysis
- Meta-analysis

4. Soil samples from 200 GPS locations. Samples close together are more similar.

- Spatial analysis
- Breakpoint analysis
- Bayesian statistics

5. A bacterial population grows in an S-shape. You want to estimate the maximum population size.

- Nonlinear regression
- Kruskal–Wallis test
- Meta-analysis

6. How many plots do you need to detect a 10% increase in yield?

- Survival analysis
- Power analysis
- Generalised additive model

7. Exercise affects fitness, fitness affects illness, and stress affects both.

- Structural equation modelling
- Time series analysis
- Ordination

8. Transpiration stays constant as the soil dries, then drops sharply below some level.

- Generalised additive model
- Breakpoint analysis
- Meta-analysis

Two pieces of advice

1 · PLAN AHEAD

Plan the analysis when you design the study

Including the sample size you need (power analysis). Then you collect the right data to answer your question. When something changes, explore the options and talk to experts.

2 · QUESTION FIRST

Be driven by your questions, not by techniques

We don't analyse data for its own sake, but to answer questions about the world. Let your questions guide what you learn next.

Principle 1 from Week 1: a focused question first, then plan sampling and analysis to answer it.

Where to start learning more

CRAN Task Views

Curated lists of R packages, by topic:

`TimeSeries` `Spatial` `MachineLearning` `MetaAnalysis` `Survival` `Bayesian` `Psychometrics` (SEM) `ExperimentalDesign` `Environmetrics`

cran.r-project.org/web/views

Statistical Rethinking

Richard McElreath's book on multilevel Bayesian modelling, with free lecture videos.

The course book

Chapter 11.2 lists all twelve, with what each is for.

People

Statisticians, and colleagues who use the methods. Ask early.

SITUATION

I want to work as a data scientist or analyst, and make lots of money

A recruiter in quantitative finance and technology told me: after a science Masters or PhD, the most common first job was **quantitative researcher**, **quantitative analyst** or **data scientist**. The skills asked for:

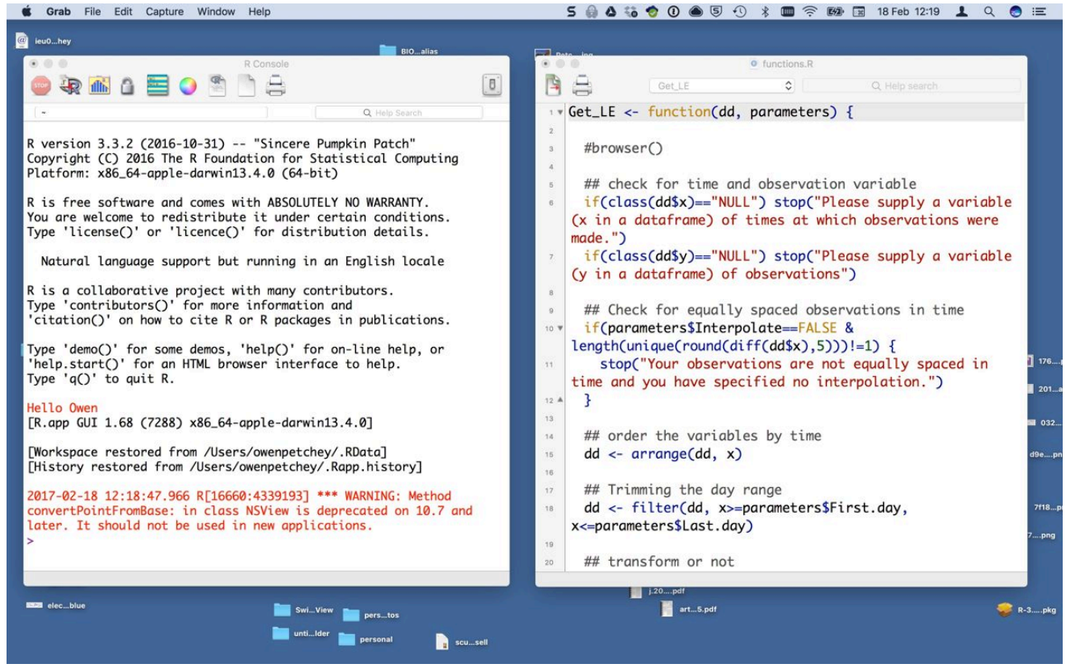
- machine learning (over the past few years, a must)
- working with large datasets: manipulation and visualisation
- programming, e.g. Python; research languages such as R
- time series analysis and forecasting

We do all of these in my ecology group :)

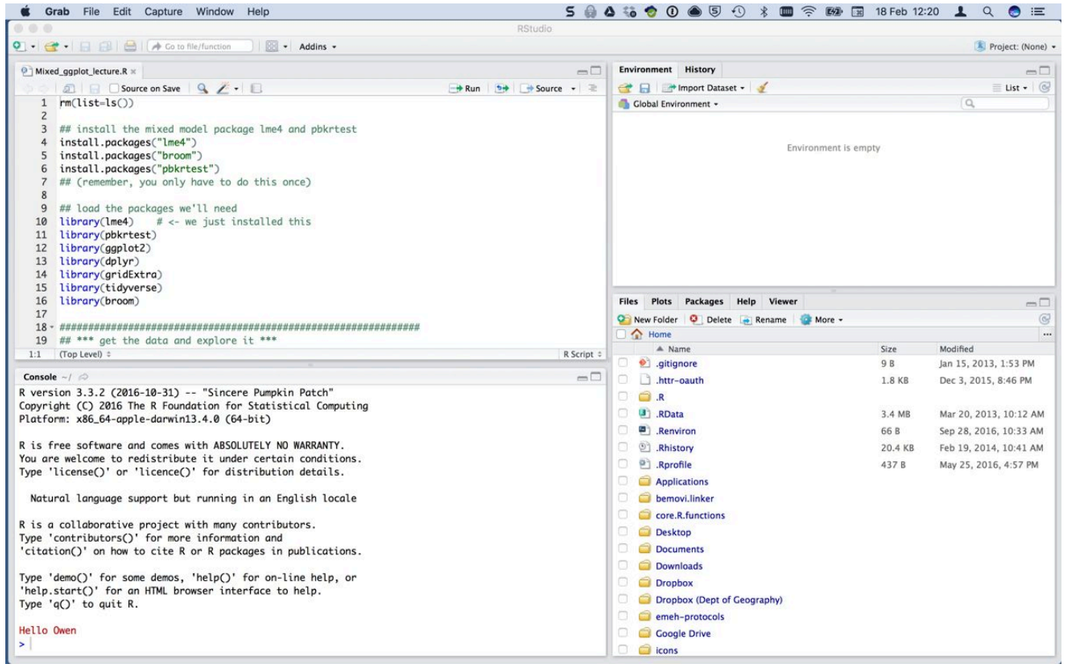
> **Machine learning, time series analysis, forecasting**

Listening: *Data Analytics in the Liberal Arts*; *How to Build Analytics Technical Skills* (How to Get an Analytics Job podcast).

R and RStudio evolve



Then: R on its own.



Now: RStudio, the tidyverse, the native pipe |>, Quarto...

So do the books: *Getting Started with R* (Beckerman & Petchey 2012; 2nd ed. with Childs, 2017), *Insights from Data with R* (Petchey et al. 2021). Keep learning, and keep updating.

Statistical methods and thinking evolve too

> **Ordinary multiple regression was once considered cutting edge: something only experts would fiddle with.**

Paraphrasing Richard McElreath, *Statistical Rethinking*.

Today's specialist method (e.g. multilevel Bayesian modelling) may be tomorrow's standard tool.

FINALLY

Can all problems be solved with data and analyses?

Necessary, but **not sufficient**. Think of the UN Sustainable Development Goals: poverty, hunger, health, climate, life on land and in water.

And “we need more data before we act” is often no longer a good argument.

Until next week…

BEFORE MONDAY

Tell us

Post on the OLAT forum what you'd like to go over in the review.

MON 31 MAY

Unit 12: Review

Planning and implementing statistical analyses.

FRI 4 JUNE

Last practical

The only practical that week: the Thursday group comes on Friday.

FRI 18 JUNE

Final exam

10:00–12:00. Mandatory Probeprüfung before then.