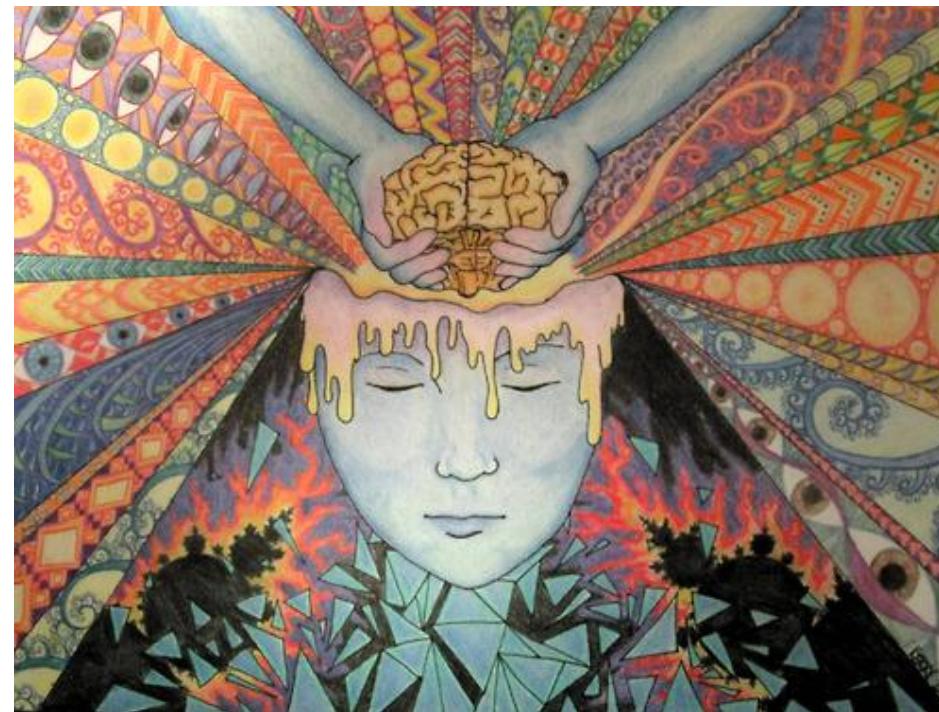
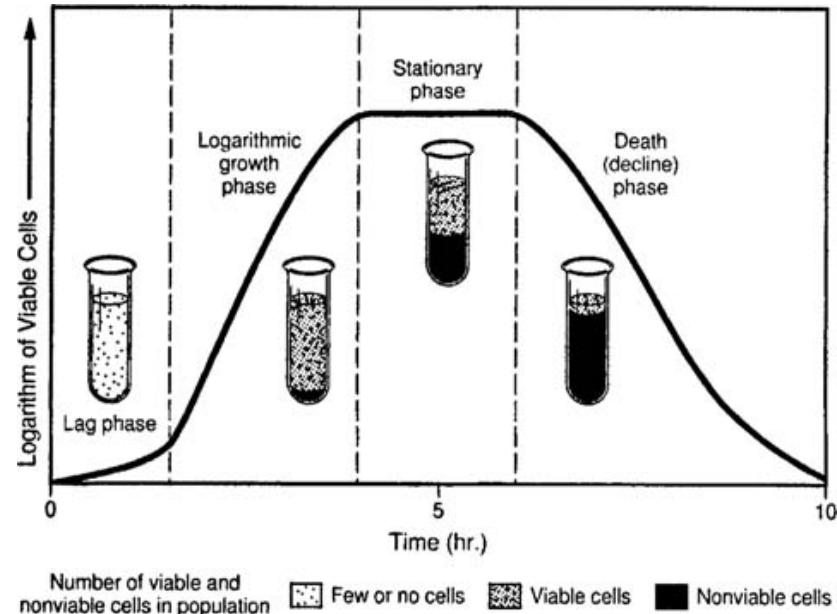


Expand your minds

Lecture 12



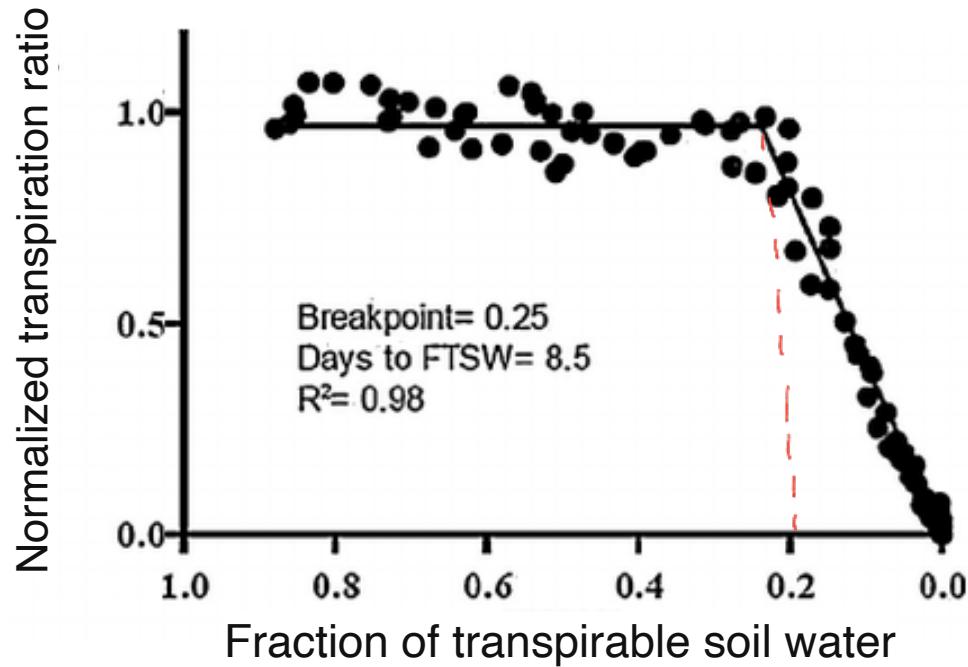
The relationship I would like to fit between two continuous variables is curvy!



$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Nonlinear regression
nls (base)

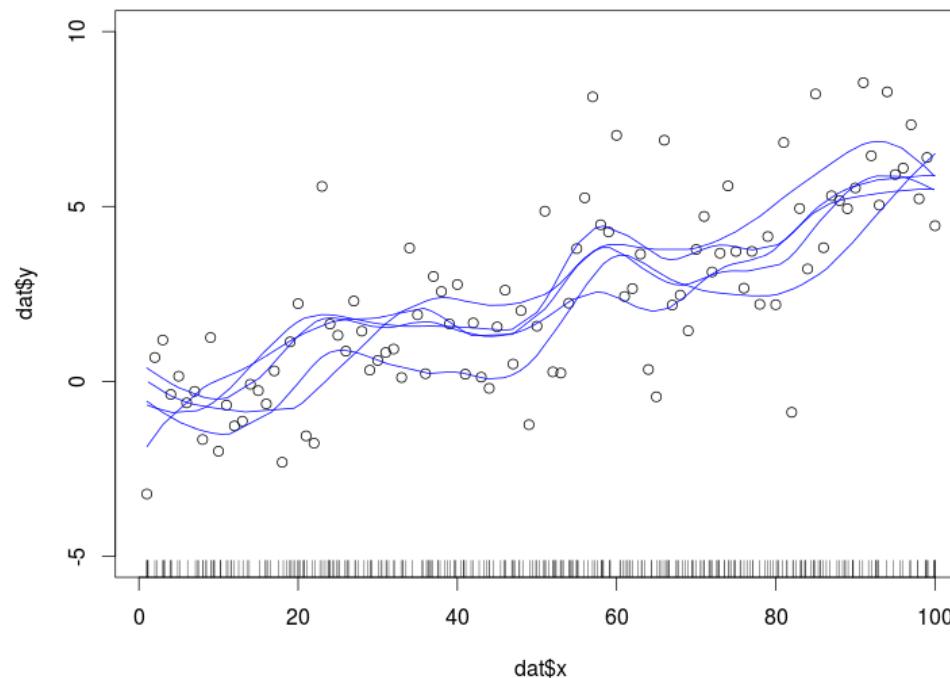
The relationship between two continuous may have a discontinuity!



Fuentealba, M.P., Zhang, J., Kenworthy, K., Erickson, J., Kruse, J. & Trenholm, L. (2016) Transpiration responses of warm-season turfgrass in relation to progressive soil drying. *Scientia Horticulturae*, 198, 249–253.

Segmented / piecewise regression / breakpoint analysis
segmented (a package)

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



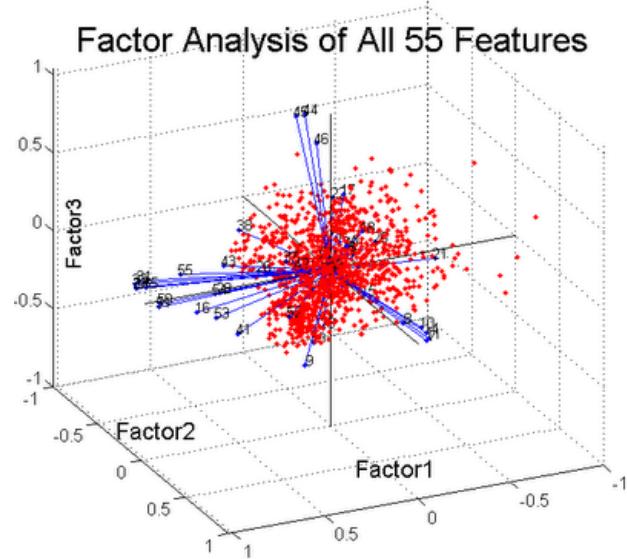
Generalised additive models
mgcv (a package)

I have multiple response variables!

Table 3. Results of the physico-chemical parameters and metals analysed in natural waters of the upper Rio Doce River basin (Quadrilátero Ferrifero)

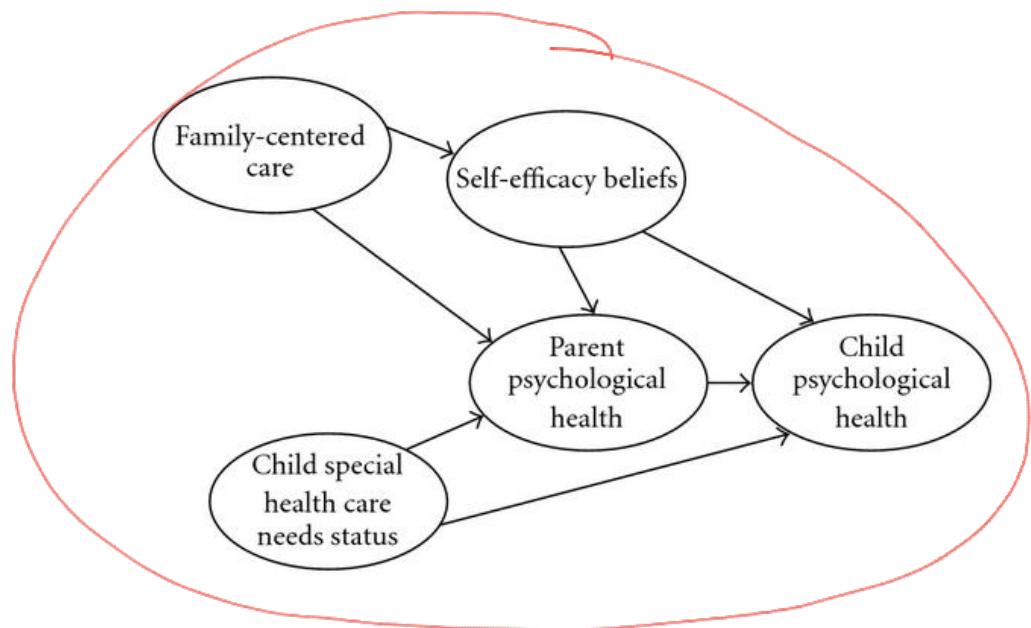
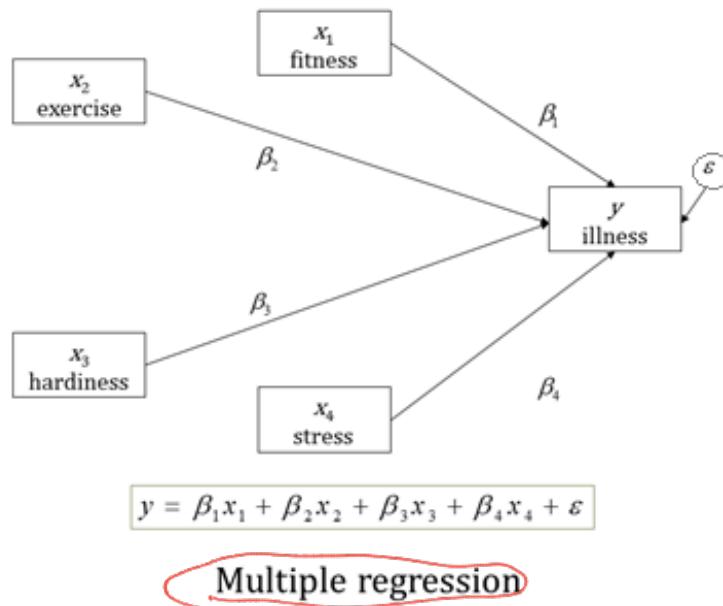
Sample	pH	DOC ^a / (mg L ⁻¹)	Temperature / °C	Alc ^b	Cond ^c / μS	ORP ^d / mV	Turb ^e / NTU	Resis ^f / kΩ	TDS ^g / (mg L ⁻¹)	Cl [−] / (μg L ⁻¹)	Ba ²⁺ / (μg L ⁻¹)	Ca ²⁺ / (μg L ⁻¹)	Fe ²⁺ / (μg L ⁻¹)	K ⁺ / (μg L ⁻¹)	Mg ²⁺ / (μg L ⁻¹)	Mn ²⁺ / (μg L ⁻¹)	Na ⁺ / (μg L ⁻¹)	S ²⁻ / (μg L ⁻¹)	Sr ²⁺ / (μg L ⁻¹)
S1A	5.92	2.79	24.0	ND ^h	10.8	155	6.57	90.4	6.9	0.98	8.7	1.22	4342.0	0.98	0.16	112.6	1.3	0.34	4.8
S1B	5.30	1.40	24.1	4.3	26.8	102	4.10	39.5	15.6	4.49	1.3	0.76	253.1	0.11	0.03	32.4	0.3	0.11	2.1
S2A	6.64	2.62	25.0	ND	28.6	108	7.29	35.0	17.9	2.41	14.7	1.35	11.5	0.32	0.65	31.6	0.3	0.09	6.4
S2B	6.16	1.17	27.4	11.9	19.5	58	32.60	48.6	12.5	4.49	54.3	2.05	36.9	0.32	0.79	4.4	0.5	0.13	7.7
S2C	7.70	1.51	22.4	5.0	11.2	268	64.80	87.6	7.2	0.50	6.9	1.54	77.6	0.06	0.45	35.7	0.3	< LOQ ⁱ	3.1
S3A	6.33	ND	22.7	7.6	19.8	109	2.81	50.3	12.4	ND	9.9	0.83	82.2	2.50	0.25	78.7	0.6	0.17	8.8
S3B	6.29	3.88	18.0	5.0	7.8	211	3.74	98.2	6.6	ND	8.1	2.00	10.6	0.45	0.33	72.6	1.0	0.45	12.7
S4A	7.45	2.71	19.1	17.5	41.2	59	36.80	24.1	26.1	ND	7.7	2.42	68.8	0.60	1.47	9.8	1.0	0.14	11.8
S4B	6.73	0.72	16.0	18.4	39.8	207	0.99	24.9	26.4	ND	9.4	4.06	65.2	0.36	1.99	6.4	1.5	0.27	16.7
S5	7.09	1.40	24.4	6.0	7.5	79	34.30	132.1	4.7	2.66	6.7	0.52	70.5	1.90	0.15	21.7	0.4	0.07	2.1
S6	7.25	1.48	25.7	36.4	127.3	45	267.00	0.0	80.7	2.33	27.3	6.27	680.0	0.95	2.40	105.3	12.3	3.15	11.0
S7	6.89	1.38	22.2	8.9	35.5	62	279.00	28.2	22.5	0.66	20.8	2.48	356.3	1.86	0.88	234.6	2.5	0.22	4.3
S8	7.30	< 0.50	21.4	8.9	20.5	61	9.67	48.0	13.1	4.66	10.4	1.37	71.7	0.18	0.62	2.0	0.4	0.07	3.9
S9A	5.40	3.83	15.6	2.0	8.0	247	1.24	119.7	5.2	0.50	5.9	1.65	469.4	0.48	0.14	8.3	1.5	0.19	2.1
S9B	6.22	2.20	16.8	3.0	4.6	183	ND	189.9	3.0	0.75	1.3	0.30	104.0	0.12	0.14	3.9	0.3	< LOQ	0.9
S10	5.75	3.63	13.9	2.5	5.0	218	0.79	172.7	4.0	ND	6.0	0.92	119.1	0.26	0.09	5.6	1.0	0.09	2.0

^aStandard deviation calculated by replicate analyses was less than 10%; ^bGiven in mg CaCO₃ L⁻¹; ^cND: Not determined; ^dLOQ: Limit of quantification; ^eDOC: dissolved organic carbon; ^fAlc: alkalinity; ^gCond: conductivity; ^hORP: redox potential; ⁱTurb: turbidity; ^jResis: resistivity; ^kTDS: total dissolved solids.



Multivariate analyses (PCA, NMDS, RDA, clustering, constrained ordination, ...)
vegan (a package); multivariate [taskview](#)

I have a network / system of variables



Path analysis / structural equation modelling
sem and lavaan (packages)

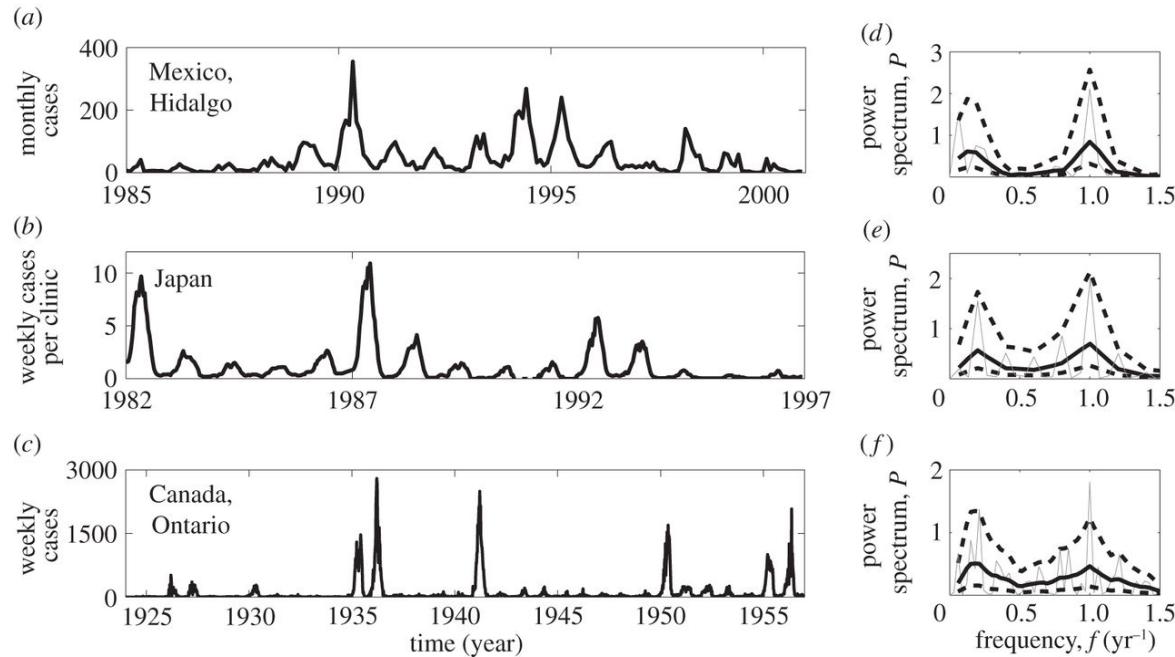
I have some prior information

The Posterior	The Evidence	The Prior
$P(H E)$	The probability of getting this evidence if this hypothesis were true	The probability of H being true, before gathering evidence
The probability that the hypothesis (H) is true given the evidence (E)	$\frac{P(E H) P(H)}{P(E)}$	The marginal probability of the evidence (Prob of E over all possibilities)

Bayesian methods

RStan, R-INLA, rjags (packages)
(Book: Statistical Rethinking by McElreath)

I have a series of values through time

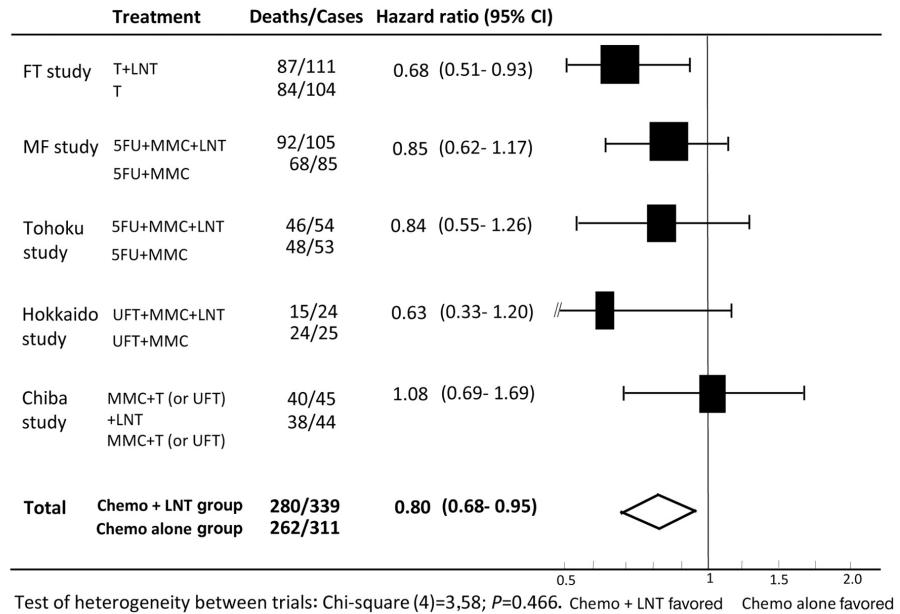


Rozhnova, G., Metcalf, C.J.E. & Grenfell, B.T.
(2013) Characterizing the dynamics of
rubella relative to measles: the role of
stochasticity. *Journal of The Royal Society
Interface*, 10.

Time series analyses

Many base functions (acf); forecast, rEDM packages

I'm reviewing and analysing the results of many studies



Oba, K., Kobayashi, M., Matsui, T., Kodera, Y. & Sakamoto, J. (2009) Individual patient based meta-analysis of lentinan for unresectable/recurrent gastric cancer. Anticancer research, 29, 2739-45.

Meta-analysis metafor package (and others)

My response variable is the amount of time until something happened

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

type

Tumor DNA profile

(1=Aneuploid Tumor, 2=Diploid Tumor)

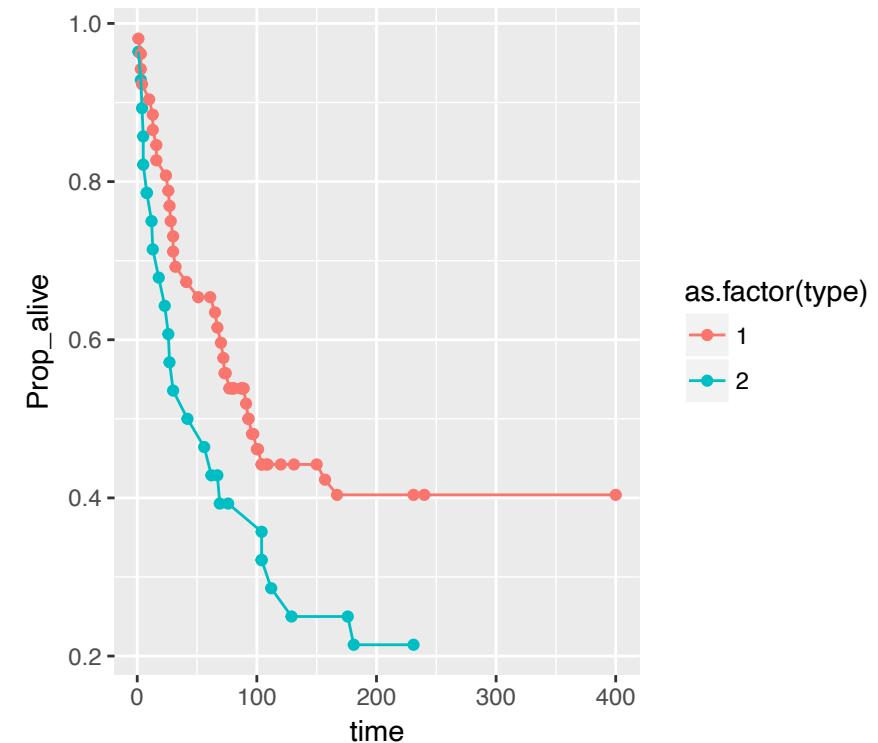
time

Time to death or on-study time, weeks

delta

Death indicator (0=alive, 1=dead)

Censoring



Survival analysis survival package (and others)

I can't fix my residuals

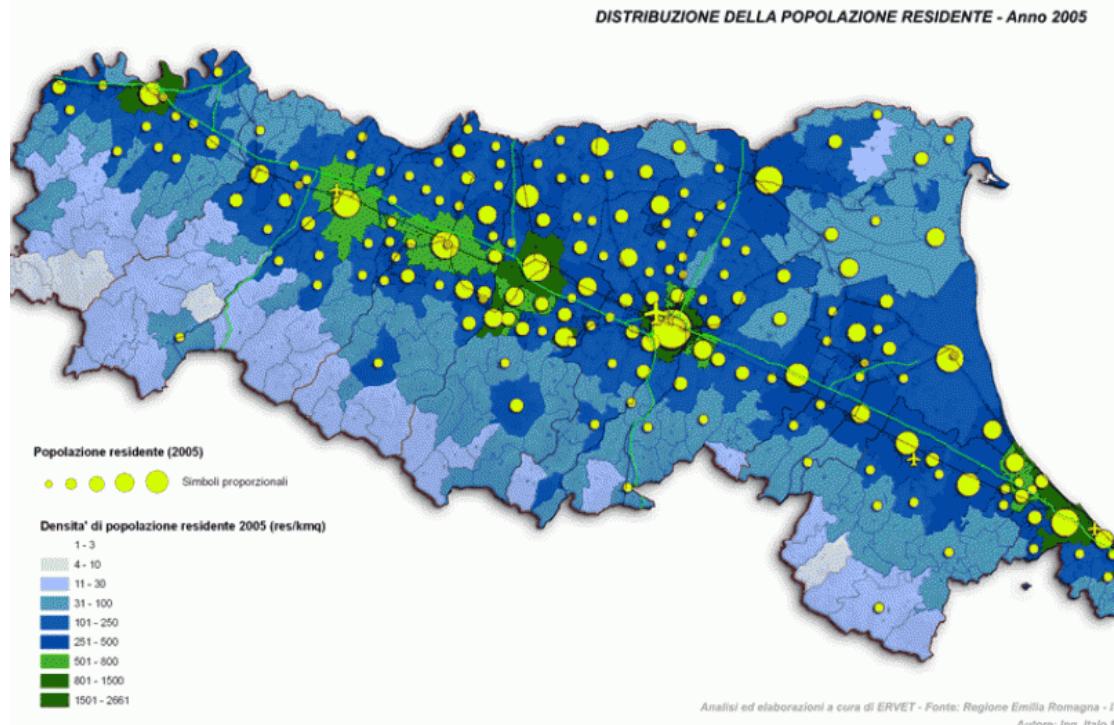
Kruskal–Wallis

Somewhat equivalent to

one-way analysis of variance

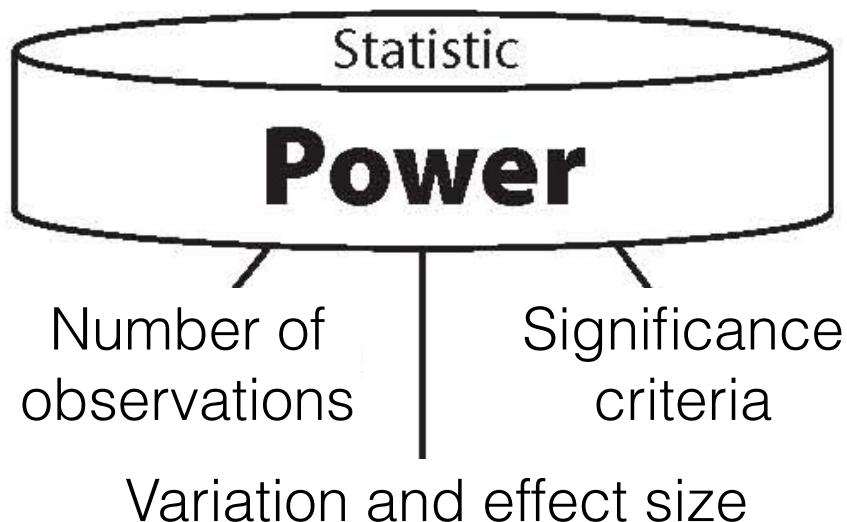
Nonparametric analysis
base package (and others)

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



Spatial analysis
spatial taskview

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



Power analysis
pwr package

Multiple books about each

Some methods are not mutually
exclusive

E.g. Bayesian methods can be used for most of them.

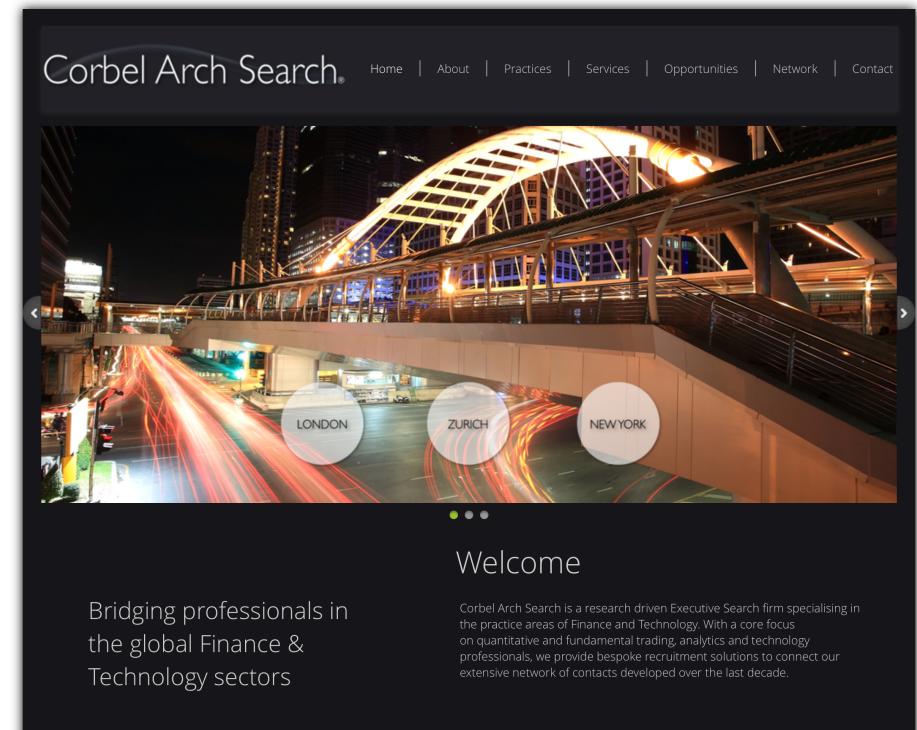
I want to work as a data scientist / analyst and makes ~~/~~ lots of money

The most common first job (within our recruitment area of quant trading and finance technology), after completing a Masters / PhD in a science was “quantitative researcher” or “quantitative analyst” or “data scientist”.

The most common skills required for this (and therefore possibly something to think of covering / including in the course are):

- Understanding or exposure to **Machine Learning Techniques** (over the past 3 years this has become a must)
- Experience working with large data sets (**data manipulation and visualization**)
- Experience with numerical **programming** in an object-oriented language is useful
- Languages: Java / C / **Python** (or at least one scripting language or some exposure / project with a scripting language such as Perl / Python).
- Research languages: **R** or Matlab
- **Time series analysis and forecasting**, this is very relevant for finance

We do all of these in my ecology group :)



Machine learning, time series analysis, forecasting

I want to work as a data scientist / analyst
and makes lots of money

Data Analytics in the Liberal Arts

How to Build Analytics Technical Skills
(*How to Get an Analytics Job* Podcast)

Machine learning, time series analysis, forecasting

R / RStudio evolves

This screenshot shows the R console window from 2013. The title bar reads "R Console". The main area displays R version 3.3.2 code, including a function named "Get_LE" and a warning message about deprecated methods. The interface is minimalist with a dark background and light text.

```

R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Hello Owen
[R.app GUI 1.68 (7288) x86_64-apple-darwin13.4.0]
[Workspace restored from /Users/owenpetchey/.RData]
[History restored from /Users/owenpetchey/.Rapp.history]

2017-02-18 12:18:47.966 R[16660:439193] *** WARNING: Method
convertPointFromBase: in class NSView is deprecated on 10.7 and
later. It should not be used in new applications.
>

```

This screenshot shows the RStudio interface from 2016. The title bar reads "RStudio". The main area displays the same R version 3.3.2 code as the previous screenshot. The interface is more modern, featuring a sidebar with tabs for "Environment", "History", and "Global Environment". The file browser on the right shows a list of files and folders, including ".gitignore", ".httr-cauth", ".RData", ".Rprofile", and various PDF and PPTX files. The overall design is cleaner and more integrated than the original R console.

```

R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Hello Owen
>

```

R / RStudio evolves

BECKERMAN & PETCHY

GETTING STARTED WITH R

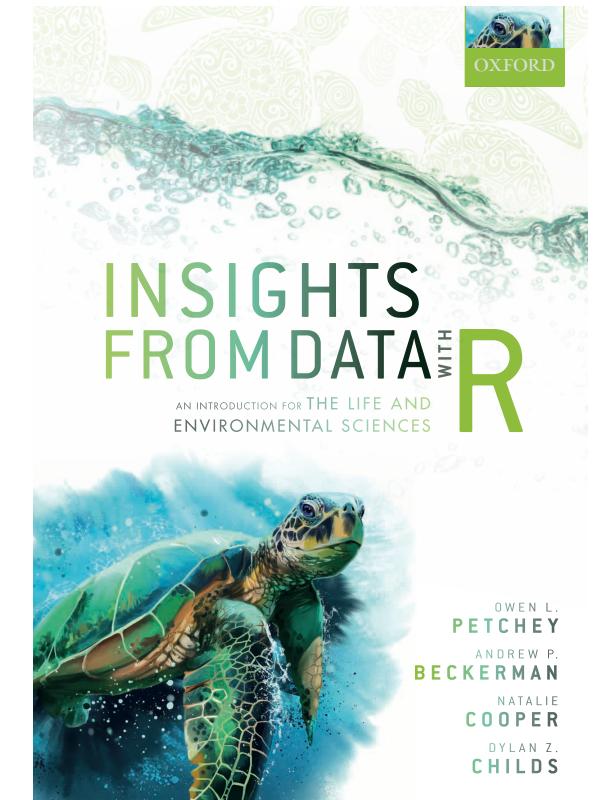
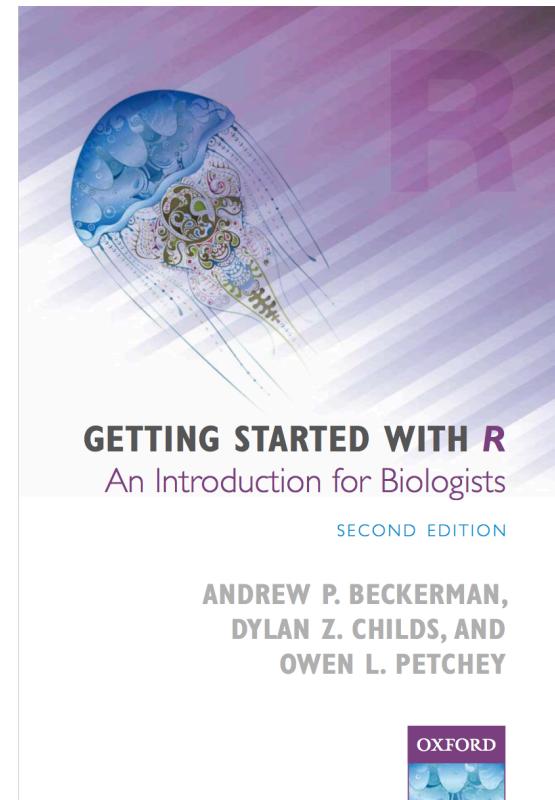
OXFORD



GETTING STARTED WITH R

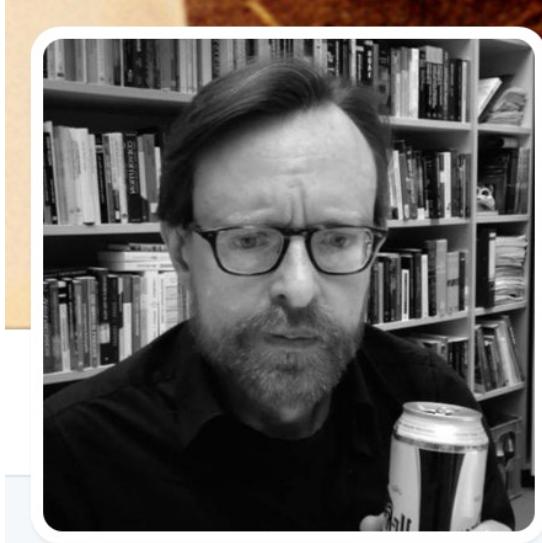
An Introduction for Biologists

ANDREW P. BECKERMAN
& OWEN L. PETCHY



Statistical methods and thinking evolves

“There was a time in applied statistics when even ordinary multiple regression was considered cutting edge, something for only experts to fiddle with.”



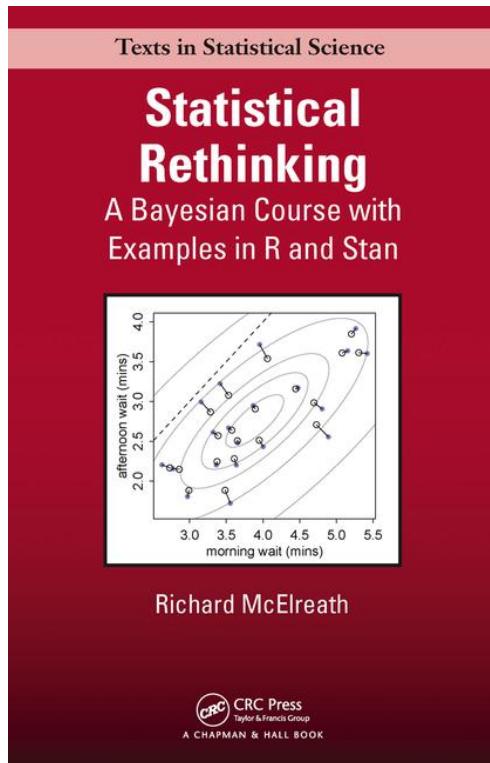
A black and white photograph of Richard McElreath, a man with glasses and a beard, holding a can of beer. He is standing in front of a bookshelf filled with books. To his right is a small orange square image of a brain scan.

Richard McElreath
@rlmcelreath

Evolutionary Anthropology, behavioral ecology, Bayesian statistics. Bayes stats course: xcelab.net/rm/statistical...

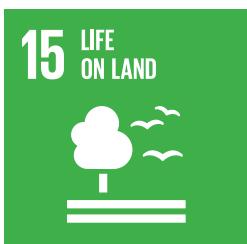
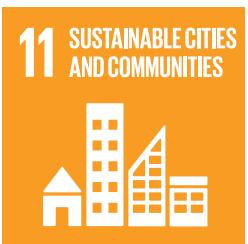
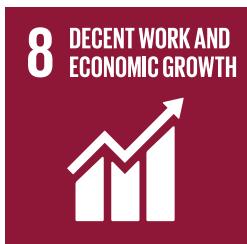
📍 MPI-EVA Leipzig
🔗 xcelab.net/rm/

Multilevel Bayesian Modelling



Can all problems be solved with data and analyses?

SUSTAINABLE DEVELOPMENT GOALS



Necessary but not sufficient

And arguments to not act until we have more data may sometimes be in bad faith.