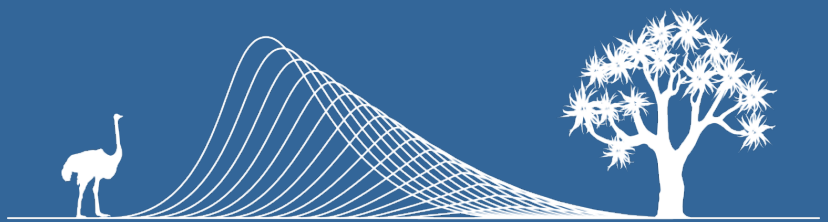
SEEC Toolbox seminars

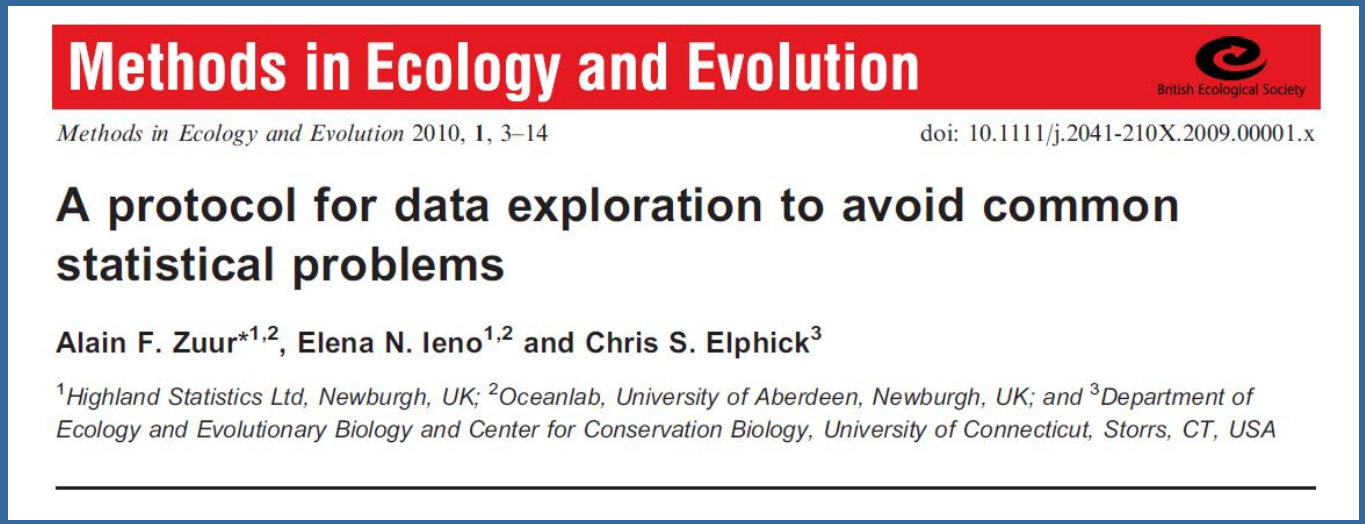
Data Exploration

Greg Distiller

29th November



Motivation



Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2010, **1**, 3–14 doi: 10.1111/j.2041-210X.2009.00001.x

A protocol for data exploration to avoid common statistical problems

Alain F. Zuur*[1,2], Elena N. Ieno[1,2] and Chris S. Elphick[3]

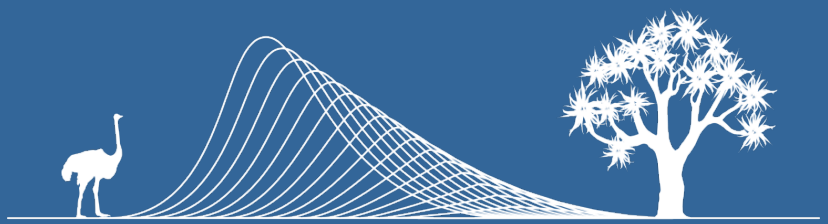
[1]*Highland Statistics Ltd, Newburgh, UK;* [2]*Oceanlab, University of Aberdeen, Newburgh, UK; and* [3]*Department of Ecology and Evolutionary Biology and Center for Conservation Biology, University of Connecticut, Storrs, CT, USA*

- The availability of sophisticated statistical tools has grown. But "rubbish in - rubbish out".
- "... even the well-known assumptions continue to be violated frequently".

Motivation

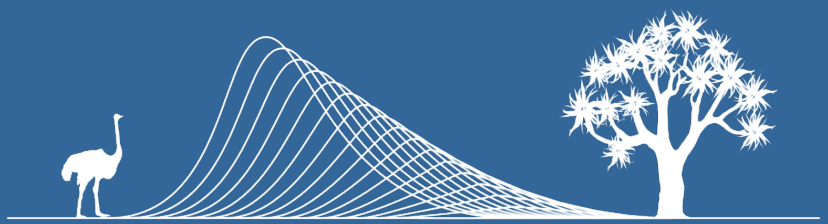
Examples of statistical pitfalls:

- Collinearity $\rightarrow$ Type II errors
- Zero inflation in GLMs may bias parameter estimates
- Spatio / temporal autocorrelation $\rightarrow$ Type I errors
- Linearity
- Ability to model interactions



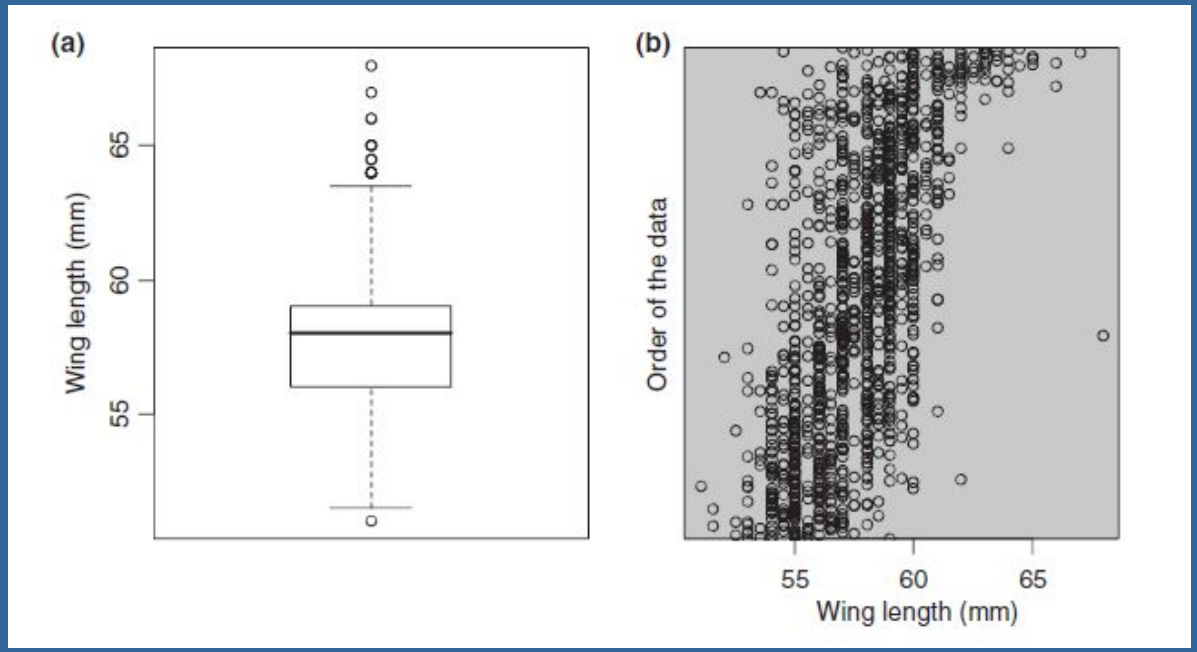
Motivation

- These various pitfalls can be avoided with proper data exploration.
- Exploration is separate from hypothesis testing.
- If a priori knowledge is very thin, exploration can be used as a hypothesis-generating exercise.
- Focus on graphical tools rather than testing.

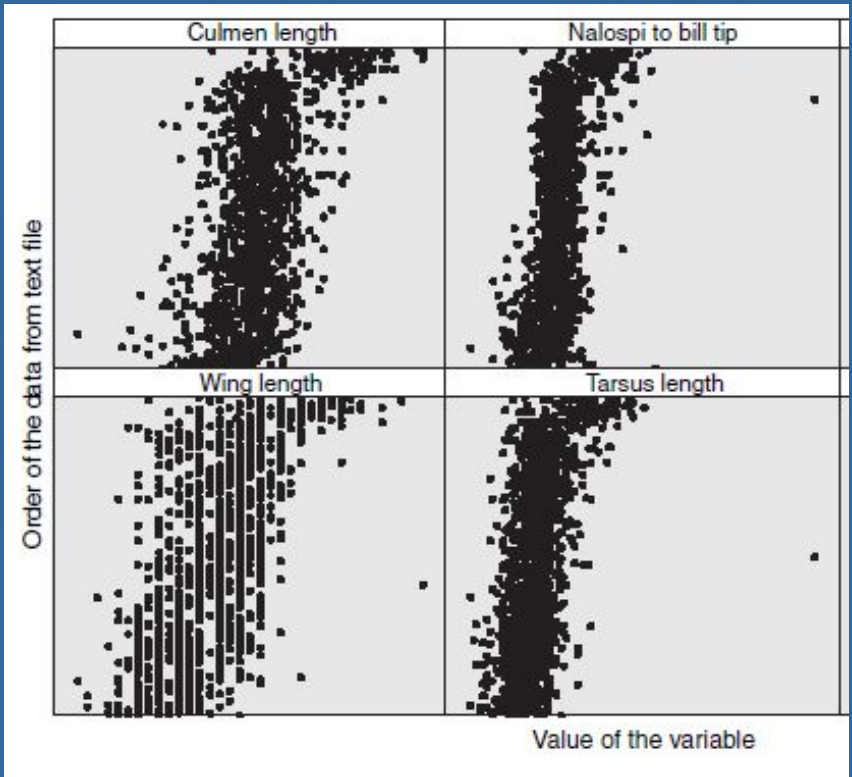


Outliers

- Some techniques can be dominated by outliers.
- They should not be removed as a matter of course!
- Boxplots are typically used.
- Cleveland dotplot (row nos vs observations) provides more information.



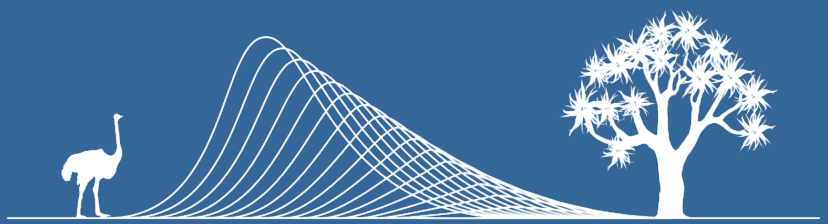
Outliers



- Extreme values by chance?
- Can simulate random observations.
- Can consider dropping if they seem to be measurement errors.

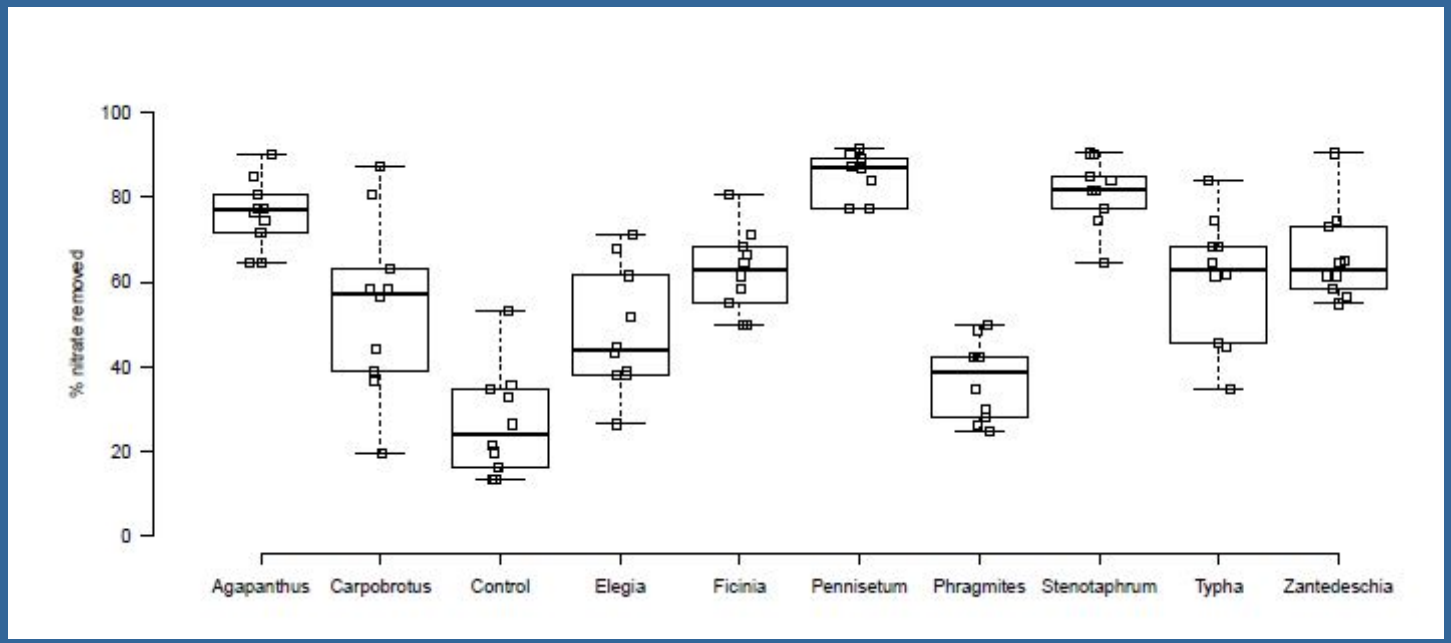
Outliers

- Outliers in X vs Y space.
  - influential observations are usually both!
- Transformations an option but really not ideal (especially for the response).
  - better to choose a more suitable model, or a more suitable metric if not modelling.



Homogeneity of variance

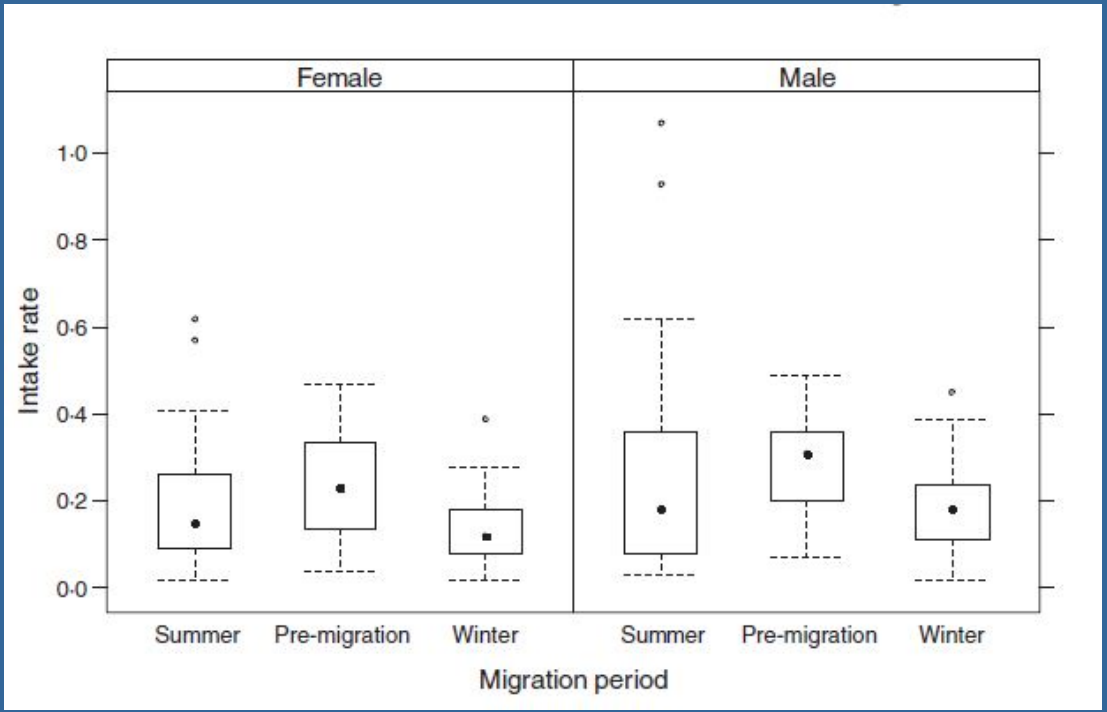
- Important assumption in regression type models (incl. ANOVA) and MV techniques like discriminant analysis.
- Regression-type models $\rightarrow$ verify with residuals.
- Can use conditional boxplots.



Source: Milandri S.G. et al (2012)

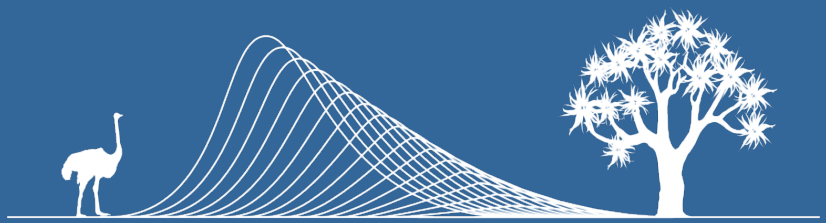
Homogeneity of variance

- Important assumption in regression type models (incl. ANOVA) and MV techniques like discriminant analysis.
- Regression-type models $\rightarrow$ verify with residuals.
- Can use conditional boxplots.



Normality

- Is normality required? What must be normal?



Normality

- Is normality required? What must be normal?

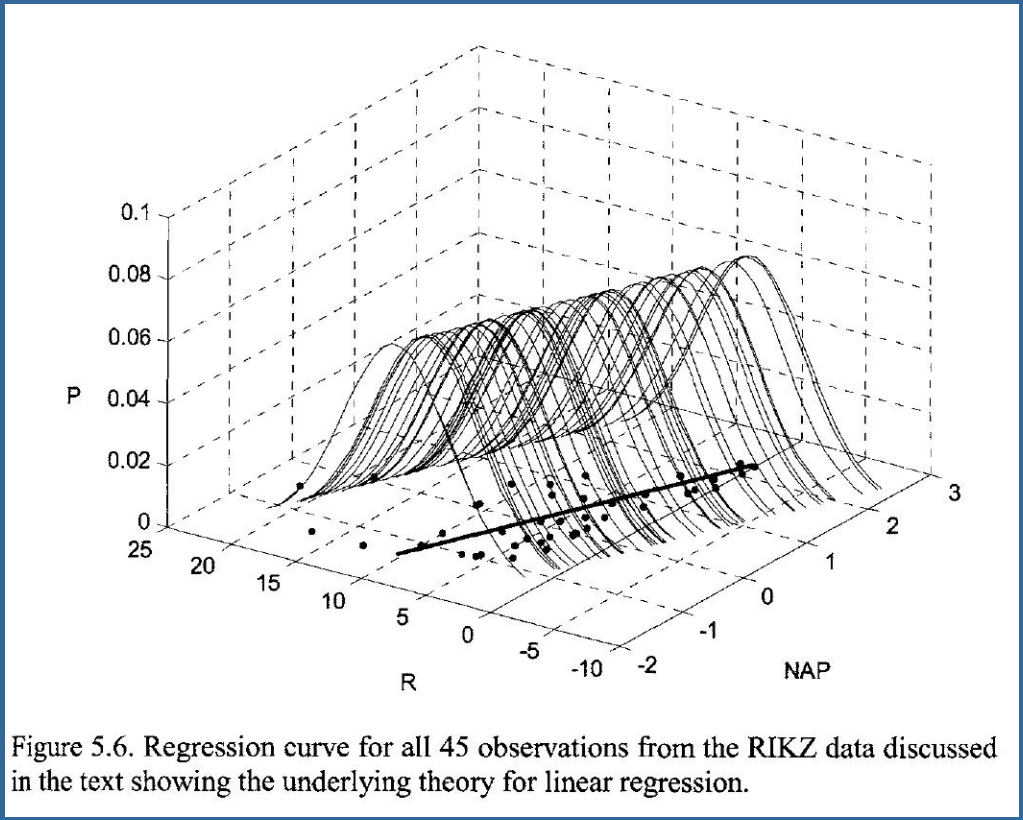


Figure 5.6. Regression curve for all 45 observations from the RIKZ data discussed in the text showing the underlying theory for linear regression.

Source: Zuur et al. Analysing Ecological Data

Normality

- Is normality required? What must be normal?

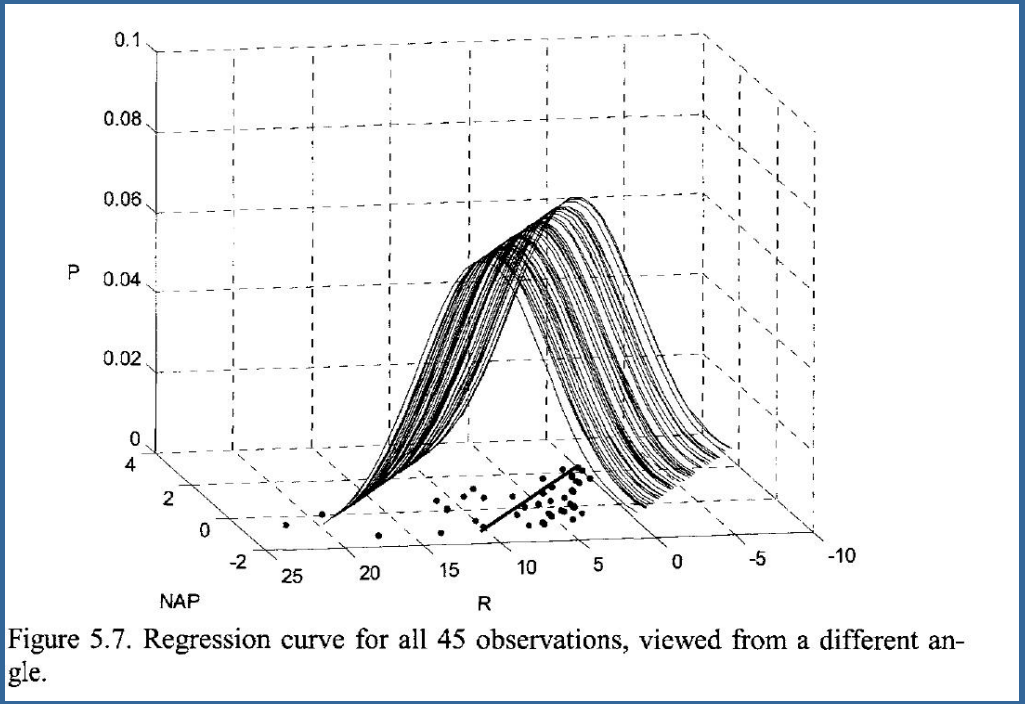


Figure 5.7. Regression curve for all 45 observations, viewed from a different angle.

Source: Zuur et al. Analysing Ecological Data

Normality

- Is normality required? What must be normal?

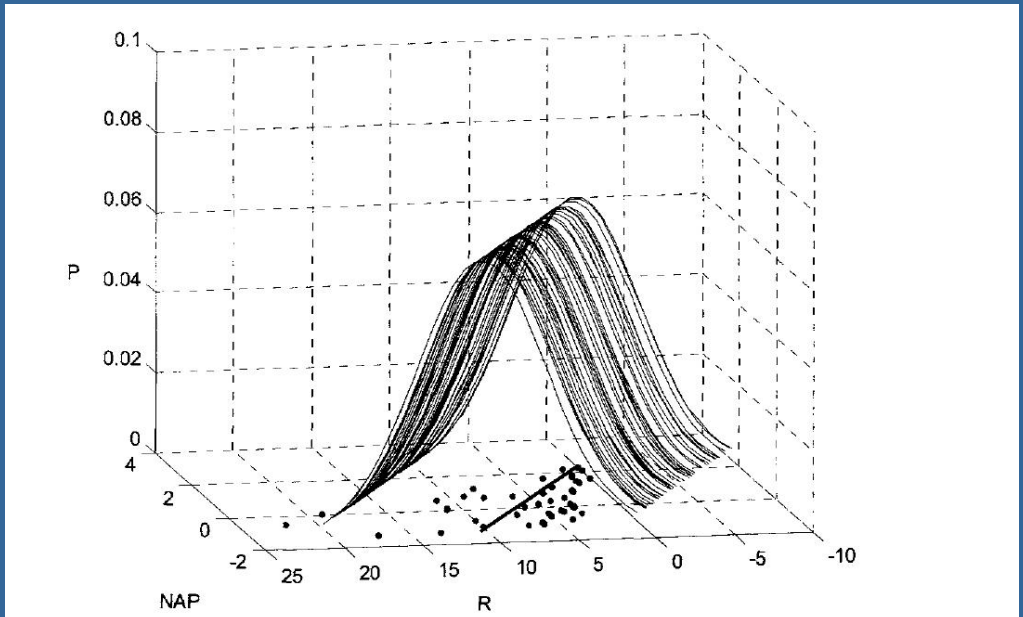


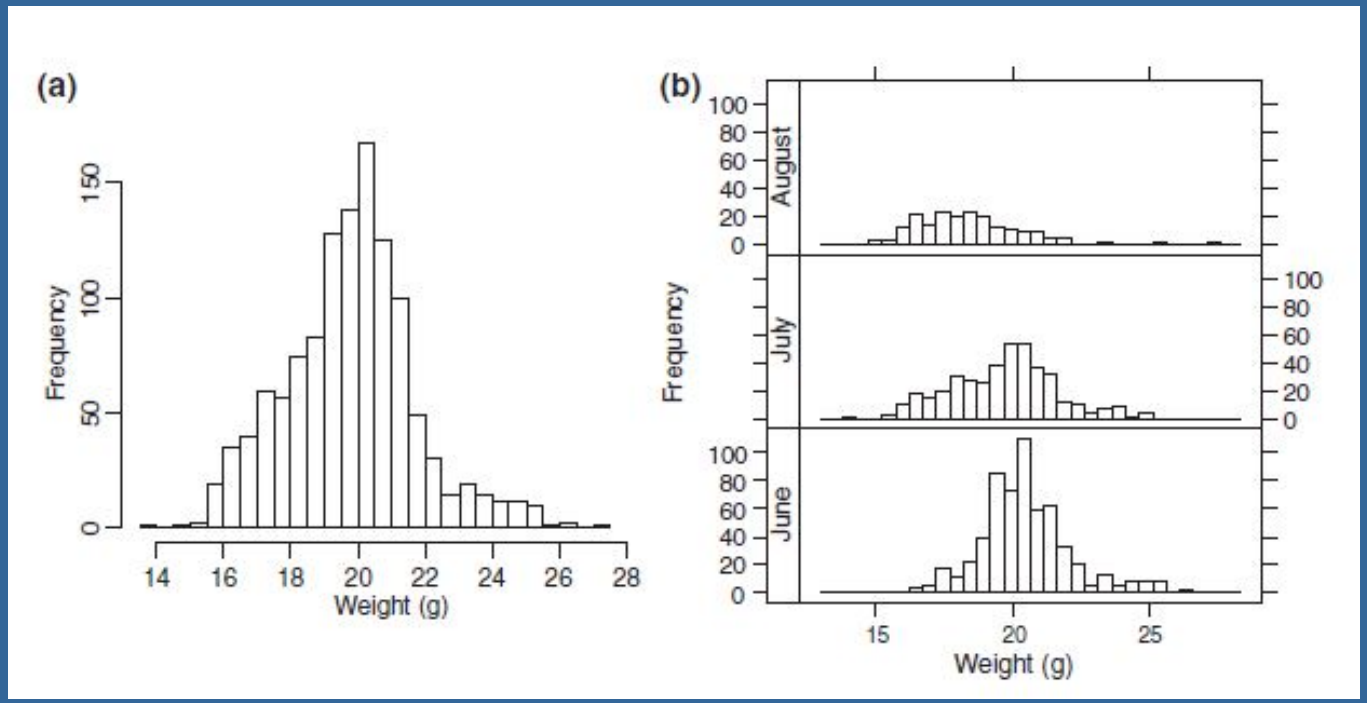
Figure 5.7. Regression curve for all 45 observations, viewed from a different angle.

Source: Zuur et al. Analysing Ecological Data

- Implies normality of the residuals.

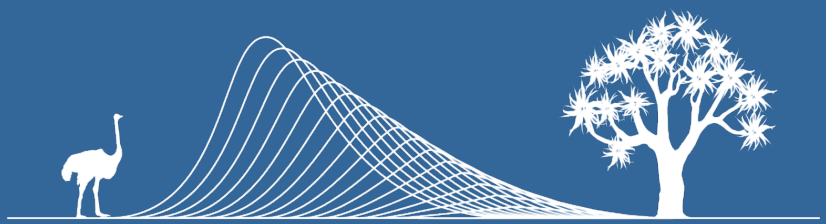
Normality

- There can be more to it than meets the eye:



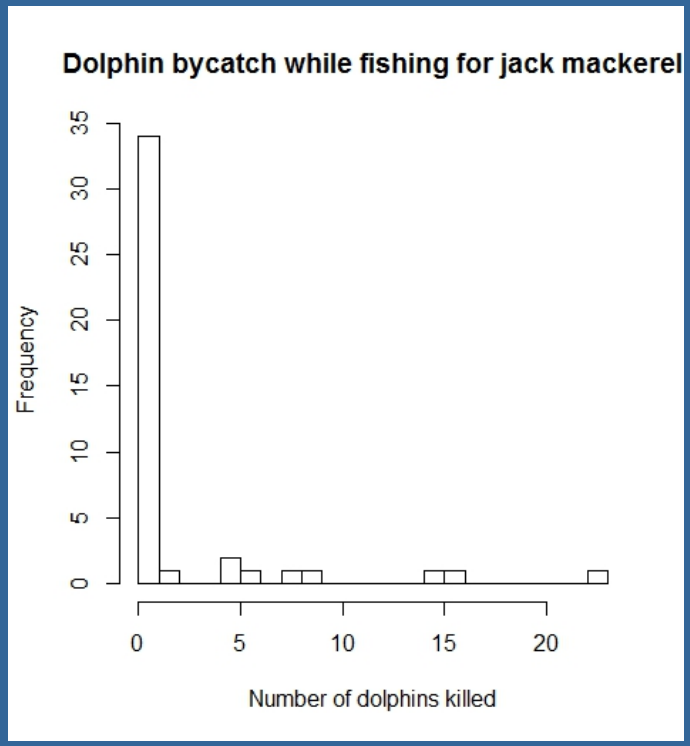
Excess Zeros

- Poisson GLM often used to model count data.



Excess Zeros

- Poisson GLM often used to model count data.



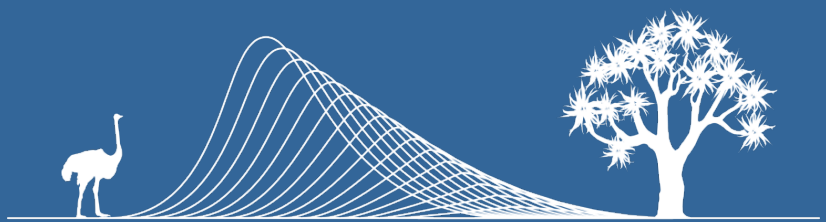
Source: Abraham, E. R., & Thompson, F. N. (2011)

- Zero-inflated models (ZIPs, ZINBs, Hurdle models . . .)

Excess Zeros

MV setting:

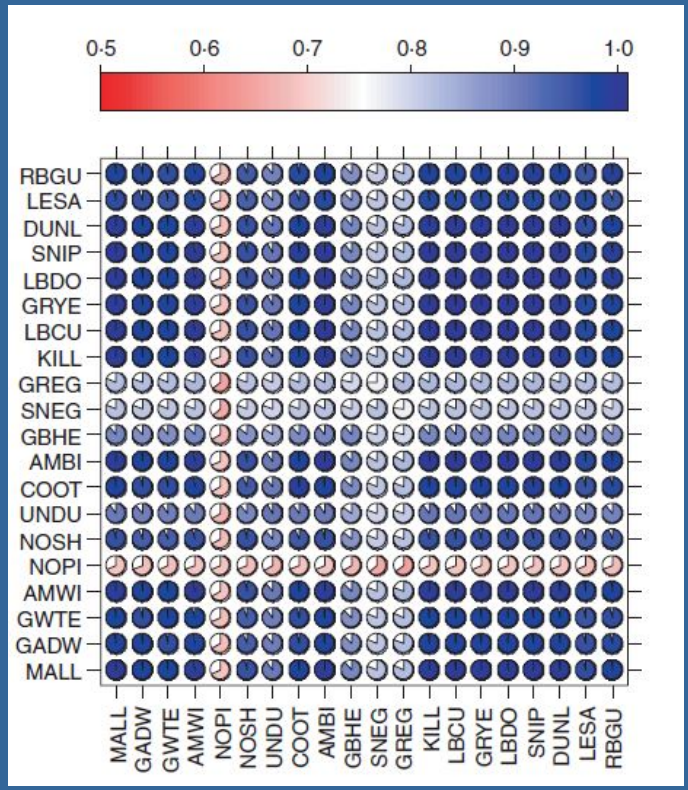
- is there useful information in joint zeros?



Excess Zeros

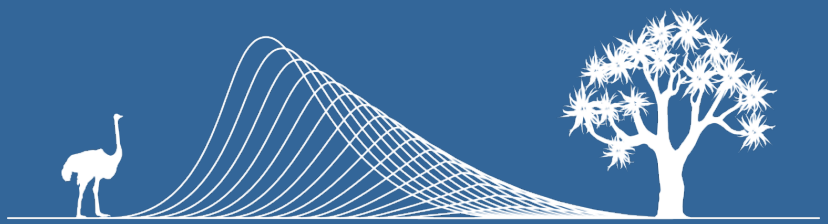
MV setting:

- is there useful information in joint zeros?
- corrgram can be used to assess prevalence of "joint absences":



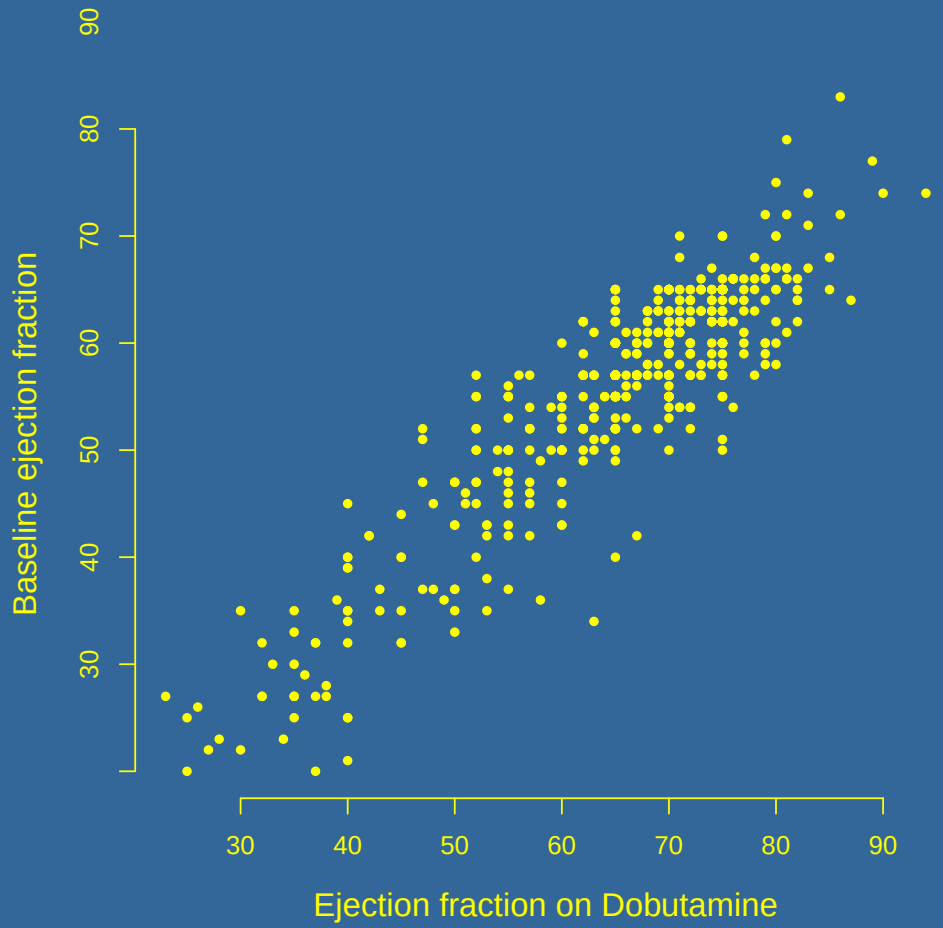
Collinearity

- NB when objective is to understand what covariates drive a system.
- Common examples: morphological measurements, water depth and distance to the shore ...
- Can lead to a confusing set of results.



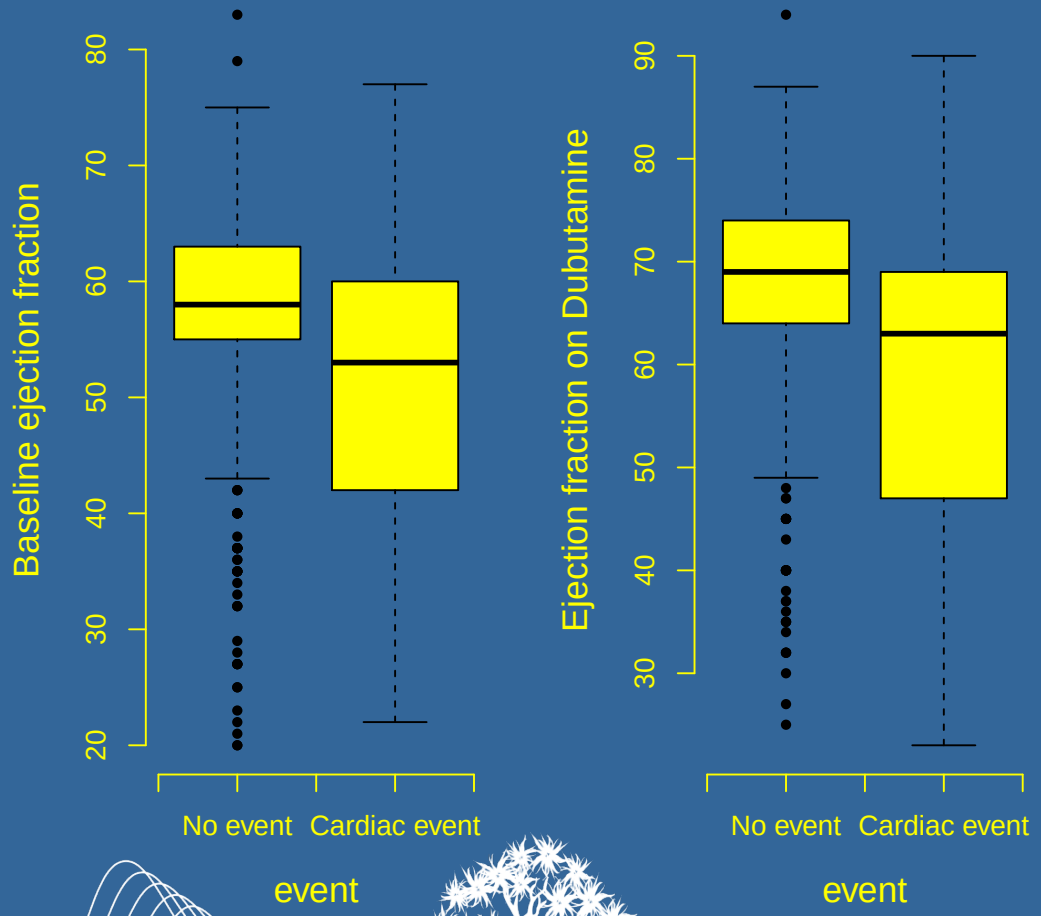
Collinearity

Source: Krivokapich J et al (1999)



Collinearity

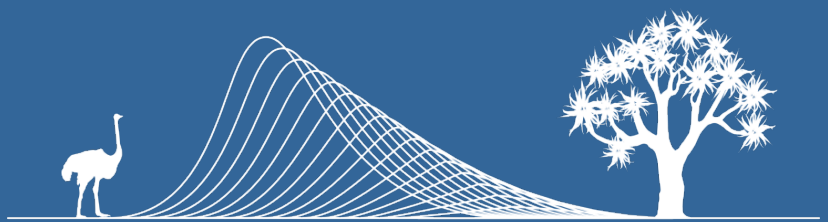
Source: Krivokapich J et al (1999)



Collinearity

```
> m1 <- glm(event ~ baseef + dobef, family = binomial)
> summary(m1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.66404    0.56674   2.936 0.003323 **
baseef       0.02878    0.02605   1.105 0.269212
dobef       -0.07770    0.02333  -3.331 0.000865 ***
```

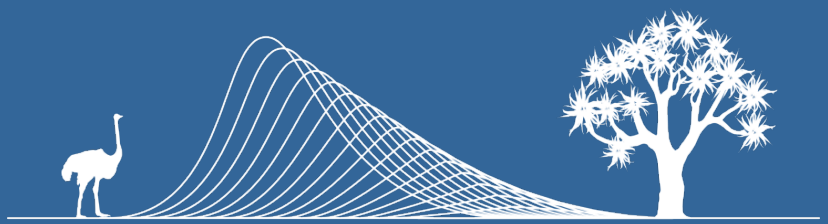


Collinearity

- There are some diagnostic tools:
  - VIF

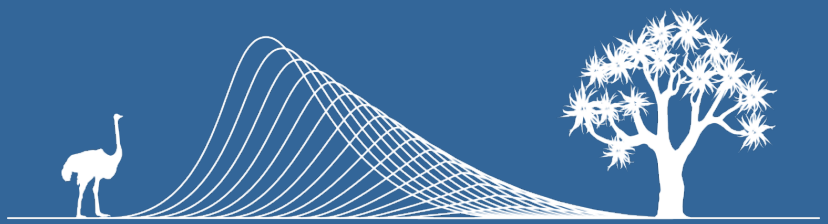
$$\frac{1}{1 - R_j^2}$$

  - scatterplots: plot covariates against any temporal / spatial variables.
- High or even moderate collinearity is problematic when the signal is weak.



Nature of the Relationship

- Use scatterplots of the response variable vs each covariate.
- Note that the absence of two-way relationships does not necessarily mean that there are no relationships.



Nature of the Relationship

Source: Gary Witt. Using data from climate science to teach introductory statistics. Journal of Statistics Education, 21(1), 2013.

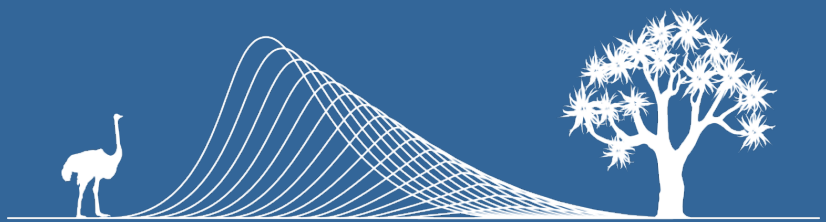
Nature of the Relationship

Source: Gary Witt. Using data from climate science to teach introductory statistics. Journal of Statistics Education, 21(1), 2013.

Nature of the Relationship

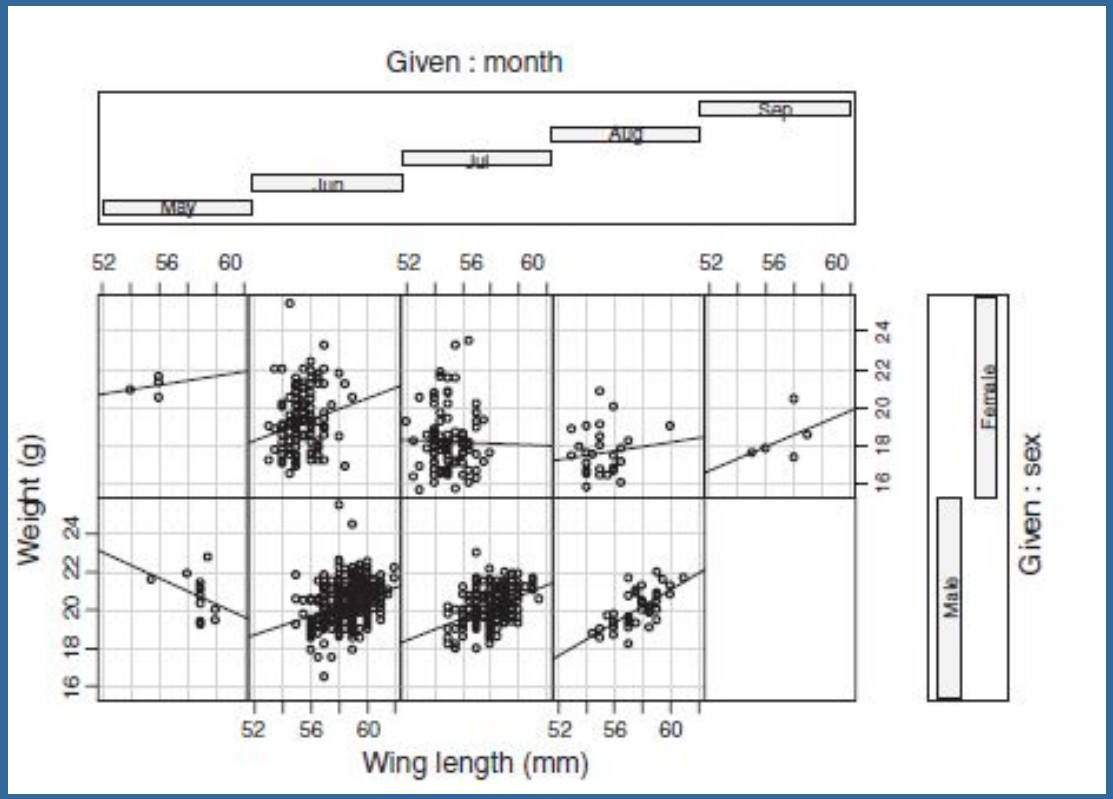
```
> m1 <- lm(ice ~ year, data = dat)
> m2 <- lm(ice ~ year + I(year^2), data = dat)
> anova(m1,m2)
Analysis of Variance Table

Model 1: ice ~ year
Model 2: ice ~ year + I(year^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     32 10.529
2     31  6.822  1    3.7072 16.846 0.0002731 ***
```



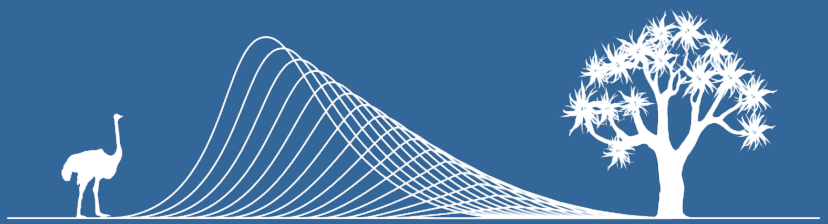
Including interactions

- A coplot is a useful plot to visualise potential interactions.
- Can reveal data sparsity.



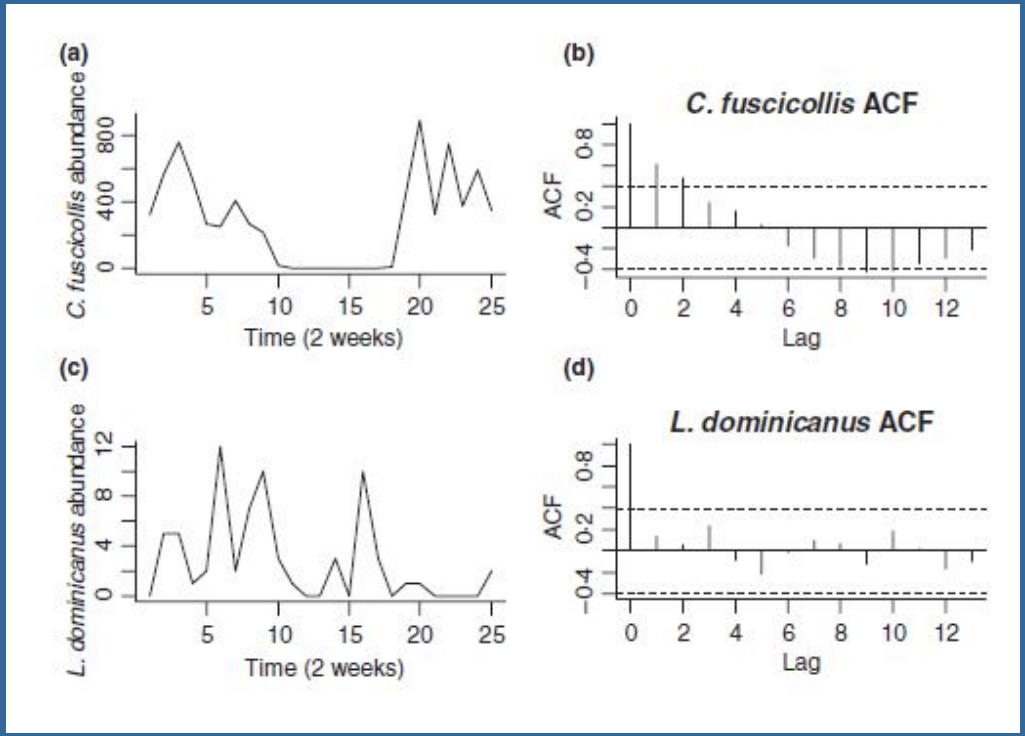
Independence

- A crucial assumption for many techniques.
- Violation increases type I errors.
- Some examples:
  - Data from different locations $\rightarrow$ are birds from locations close to each other more similar than birds from locations further away?
  - Individuals from one family may tend to be similar due to shared genes.
  - Repeated measurements.



Independence

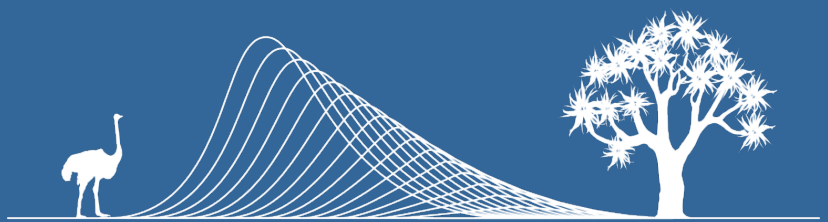
- Plot the response against time or space, any pattern suggests dependence.
- More formally, plot auto-correlation functions (ACF's).



Independence

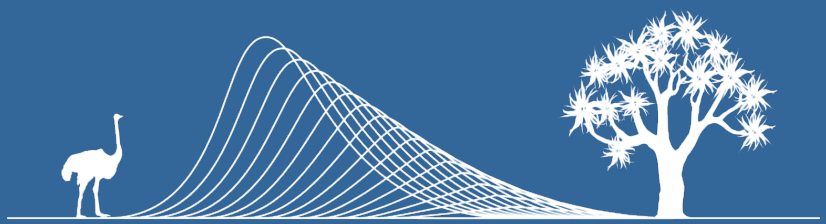
When the data are not independent:

- need a model that can account for the dependence: mixed-effects models, AR models, generalised least squares (modelling the correlation structure).
- can sometimes model the dependence with suitable covariates.
- residuals should display no dependence.



Final thoughts

- Not all steps are relevant for all data sets.
- Some techniques require analysis first (i.e. to produce residuals)
- Transformations are not ideal.



Final thoughts

- Not all steps are relevant for all data sets.
- Some techniques require analysis first (i.e. to produce residuals)
- Transformations are not ideal.

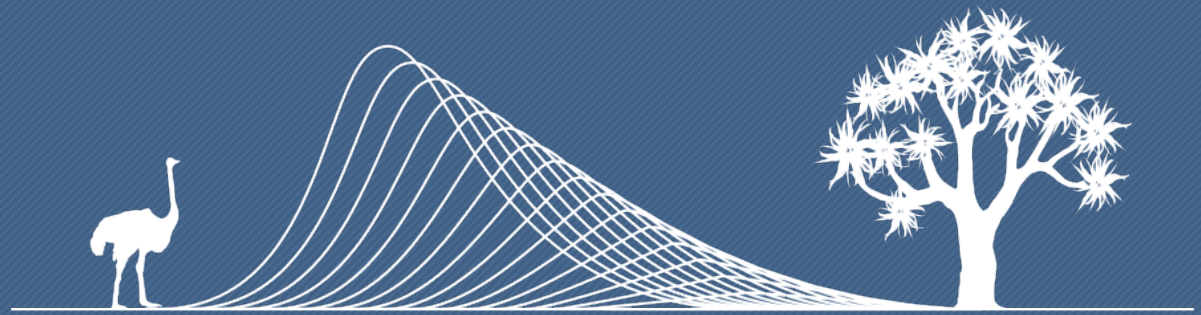
"...the routine use and transparent reporting of systematic data exploration would improve the quality of ecological research and any applied recommendations that it produces."

SEEC Stats Toolbox

See you next year!

Want to broaden your stats knowledge? Unsure of what you can do with your data? Still developing your proposal?

Join us for the monthly **SEEC Stats Toolbox** seminars where we introduce you to statistical methods that are useful for ecologists, environmental and conservation scientists.



SEEC - Statistics in Ecology, Environment and Conservation