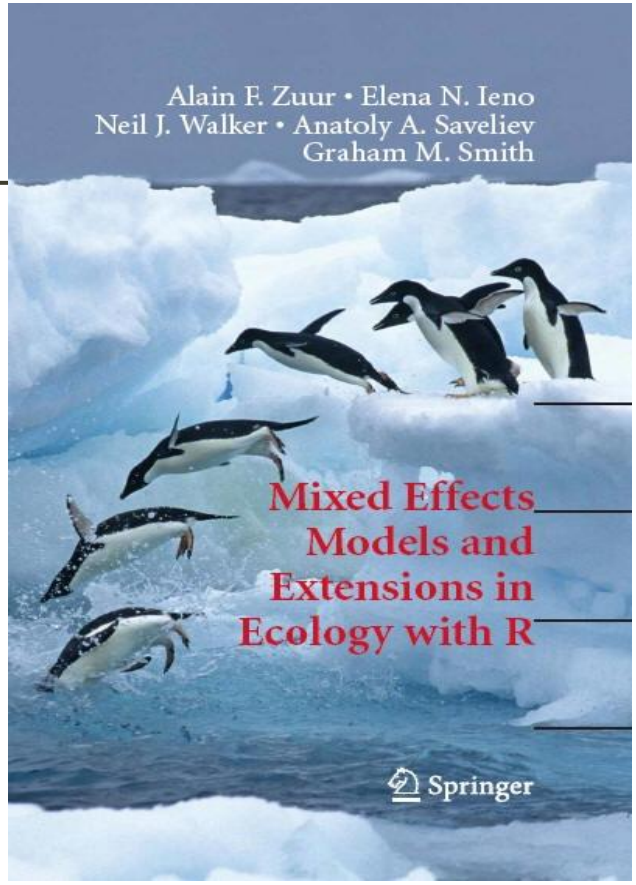
GENERAL AND GENERALIZED LINEAR MIXED MODELS

NICOLE MICHEL
Nicole.L.Michel1@gmail.com
17 OCTOBER, 2014



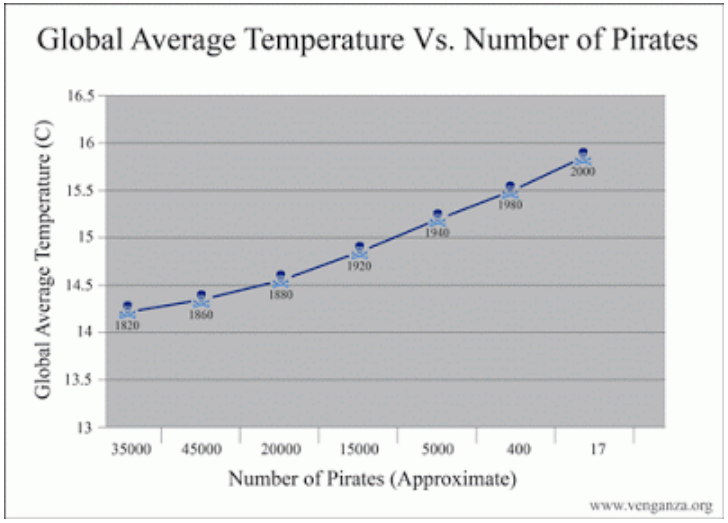
Zuur et al. 2009

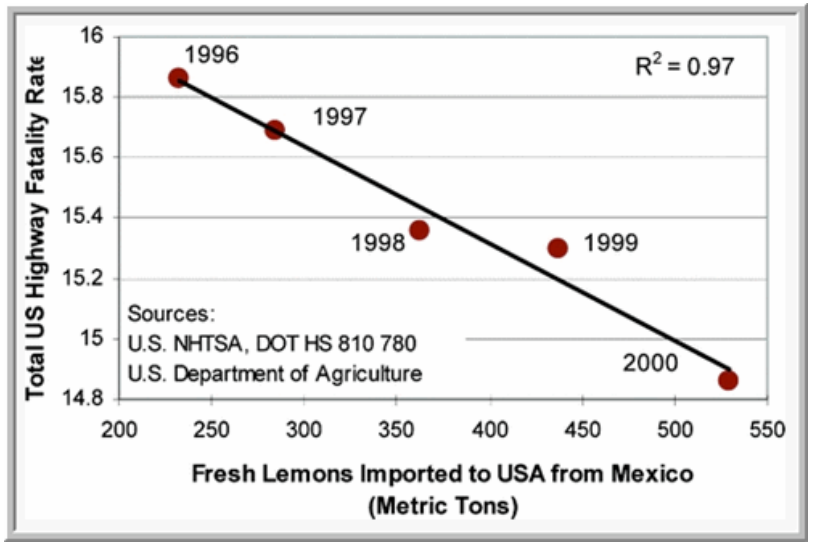
Outline

1. What are GLMs, GLMMs, and GzLMMs?
2. When to use mixed models?
3. Fixed and random effects
4. Common distributions
5. Transformations
6. Collinearity (and how to deal with it)
7. Repeated measures
8. Estimation techniques
9. R packages and commands
10. How to run, interpret, and report models in R
11. Customizing models for your dataset
12. When GLMMs aren’t enough: GAMs
13. Resources

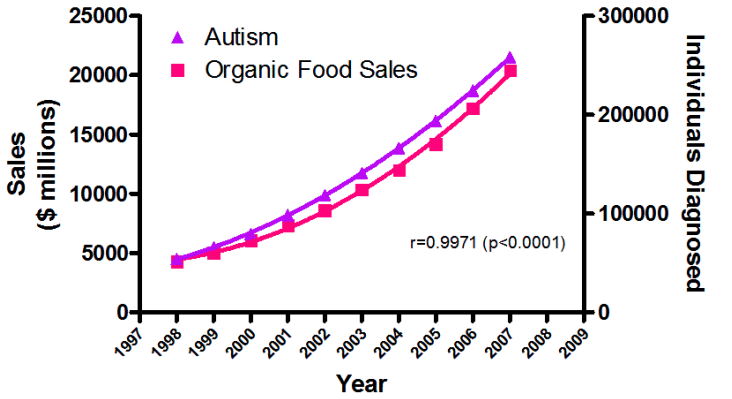
What are GLMs, GLMMs, & GzLMMs?

- GLM = General Linear Model
- GLMM = General Linear Mixed Model
- GzLMM (often just GLMM) = Generali**zed** Linear Mixed Model





General Linear Models (GLMs)



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

Equation:

$$Y_1 = \alpha + \beta X_1 + \varepsilon$$

- *t* tests
- Regression
- ANOVA/ANCOVA
- MANOVA/MANCOVA

- "linear" refers to the mathematical form of the equation; linear models do not necessarily fit straight lines

What are GLMs, GLMMs, & GzLMMs?

- GLM = General Linear Model (y = mx + b + error)
  - Also generalized linear models (see below)

- GLMM = General Linear Mixed Model
  - Mixed = includes both fixed and random effects (stay tuned)

- GzLMM (often just GLMM) = General**ized** Linear Mixed Model
  - Models that use distributions other than the normal (Gaussian), e.g., Poisson, negative binomial, binomial

Why not just use ANOVA?

- ANOVA was designed for analyzing experimental data.
- Tests means between groups
- ANOVA has assumptions that ecological data rarely, if ever, meet – and ANOVA is not very robust to violations:
  - Independence
    - Samples taken from the same individual/group, or spatially clustered individuals/groups are not independent
  - Normality (specifically: *normally distributed residuals* not *data*)
    - Ecological data is messy and very rarely meets this assumption, and normal residuals are essential for accurate F tests
  - Homogeneity of variances (homoscedascity)
    - Variances are equal in each group. Again, very rare!
  - Also: fixed X (that predictors are deterministic/known in advance)
    - Impossible with observational data

Enter mixed models

- Remember ANOVA assumes independence, normality, and homogeneity of variances
- What if you have:
  - Nested data (birds within nest boxes within sites – independence)
  - Temporal correlation (time 1 & 2 vs. time 1 & 5 – independence)
  - Spatial correlation (ponds in the same field – independence)
  - Heterogeneity (site 1 more similar than site 2 -homogeneity)
  - Repeated measurements (independence)
  - “Noise” that you need to account for but don’t care about

- **Bonus: using individual/block/site random effects allow you to generalize your results beyond your immediate study system!**

Fixed and random effects (Bolker et al. 2009)

- **Fixed effects**: “factors whose levels are experimentally determined or whose interest lies in the specific effects of each level, such as effects of covariates, differences among treatments and interactions.”
  - What you’re actually testing the effects of
- **Random effects**: “factors whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific effects of each level.”
  - Variables that represent a random subset of all possible levels
  - Variables that contain “noise” you need to control for, to ensure independence

Fixed and random effects example

- Main et al. 2014 PLoS ONE. Objective: assessing neonic concentrations over time as a function of crop type
- Independent variables. Which are fixed and random?:
  - Current crop type
  - Previous crop type
  - Time (month)
  - Pre-seeding concentration
  - Pond
  - Site (field containing multiple ponds)

Why do I need random effects?

- Begging response of nestling Barn Owls in response to parental visits (Roulin and Bersier 2007, in Zuur et al. 2009). Food experimentally supplemented or removed.
- Response: number of calls in 15 mins before parent's arrival, divided by number of nestlings (2-7/nest)

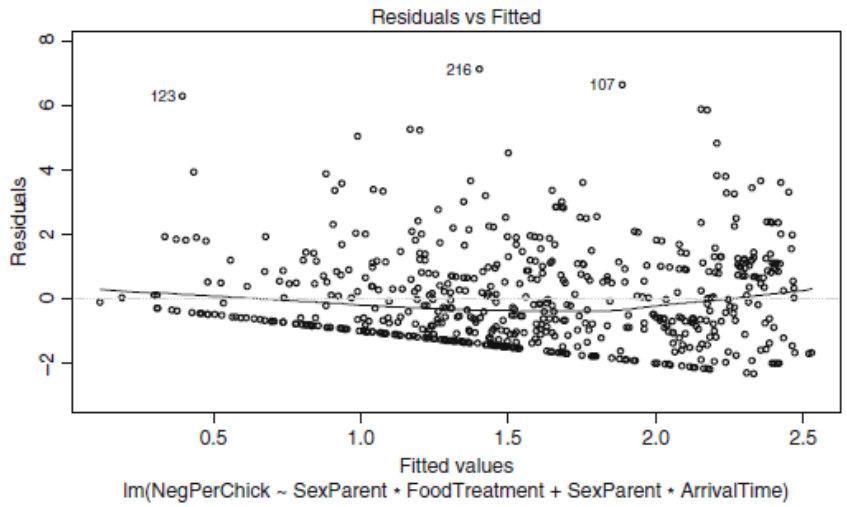


Fig. 5.4 Residuals versus fitted values for the linear regression model. Note that the residual spread increases for larger fitted values, indicating heterogeneity

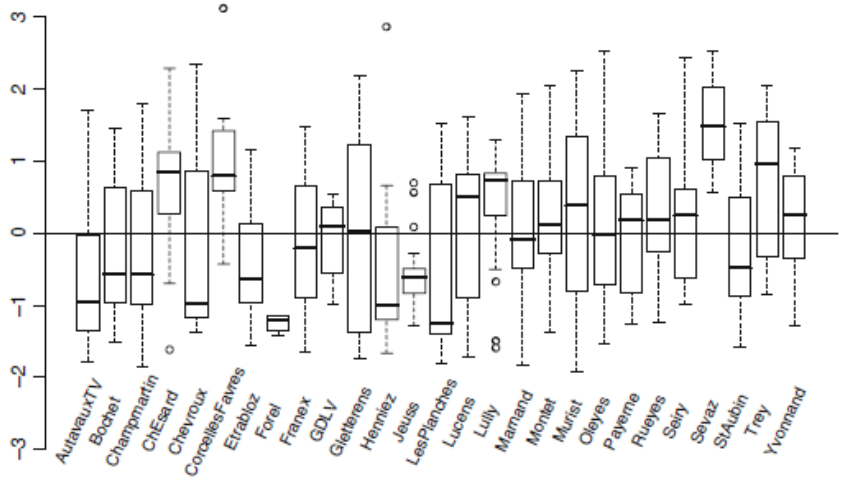
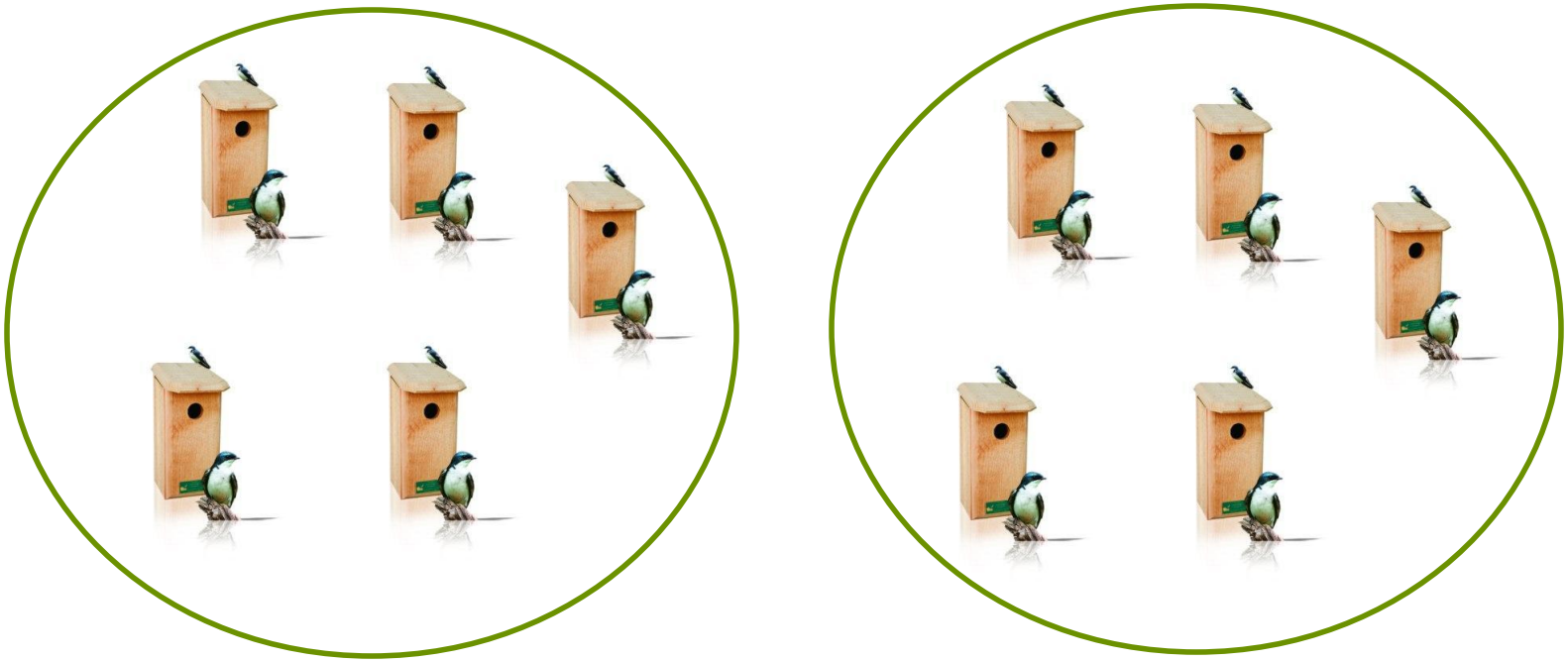


Fig. 5.5 Boxplot of standardised residuals obtained by a linear regression model applied on the log-transformed sibling negotiation data. The y-axis shows the values of the residuals and the horizontal axis the nests. Note that some nests have residuals that are above or below the zero line, indicating the need for a random effect

Nested and crossed random effects

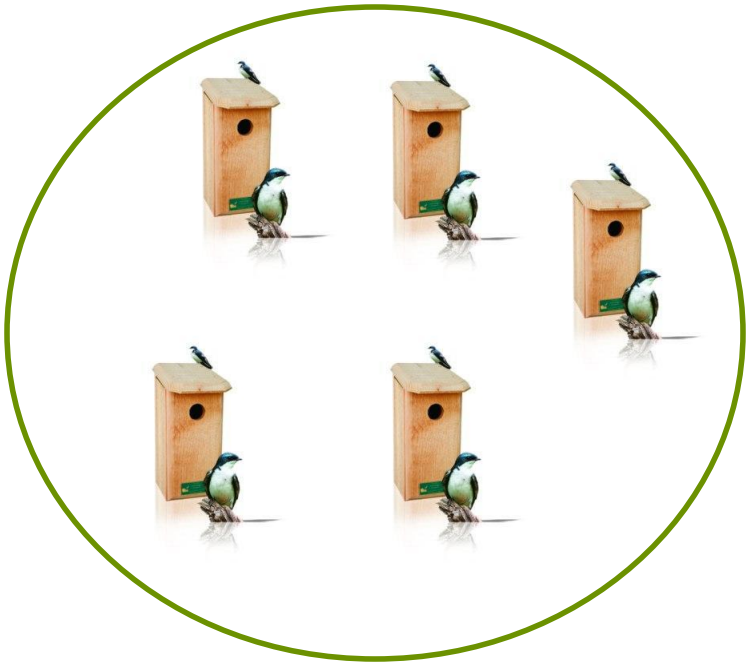
- **Nested random effects**: multiple random effects that are hierarchically structured



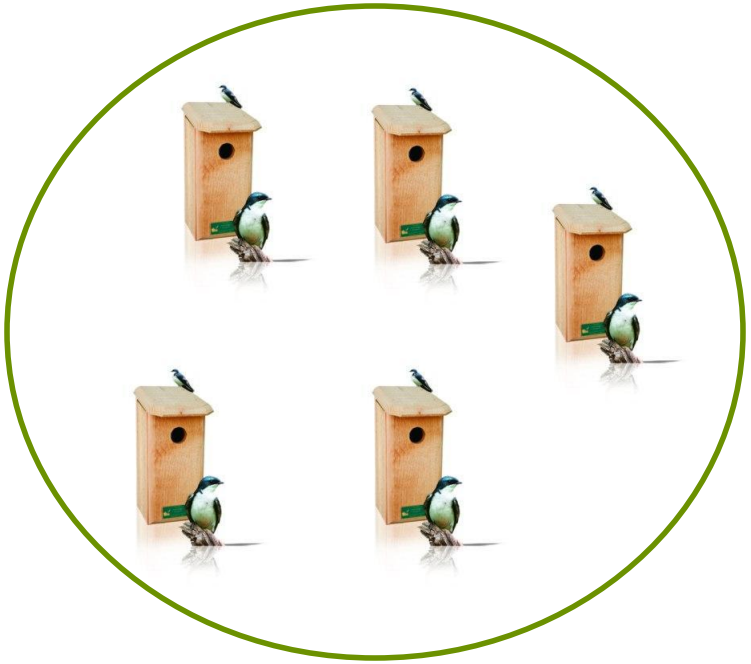
Site 1 **Site 2**

Nested and crossed random effects

- **Crossed random effects**: multiple random effects that apply independently to an individual, e.g. temporal and spatial blocks where time acts on all sites equally



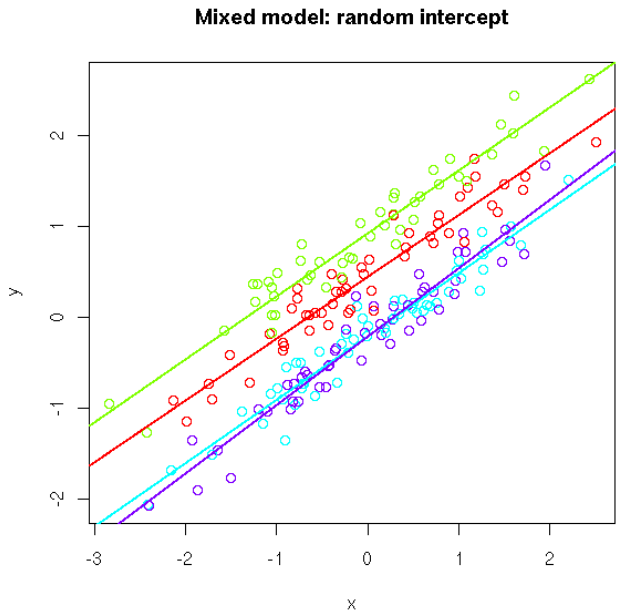
Site 1: June, July, August



Site 2: June, July, August

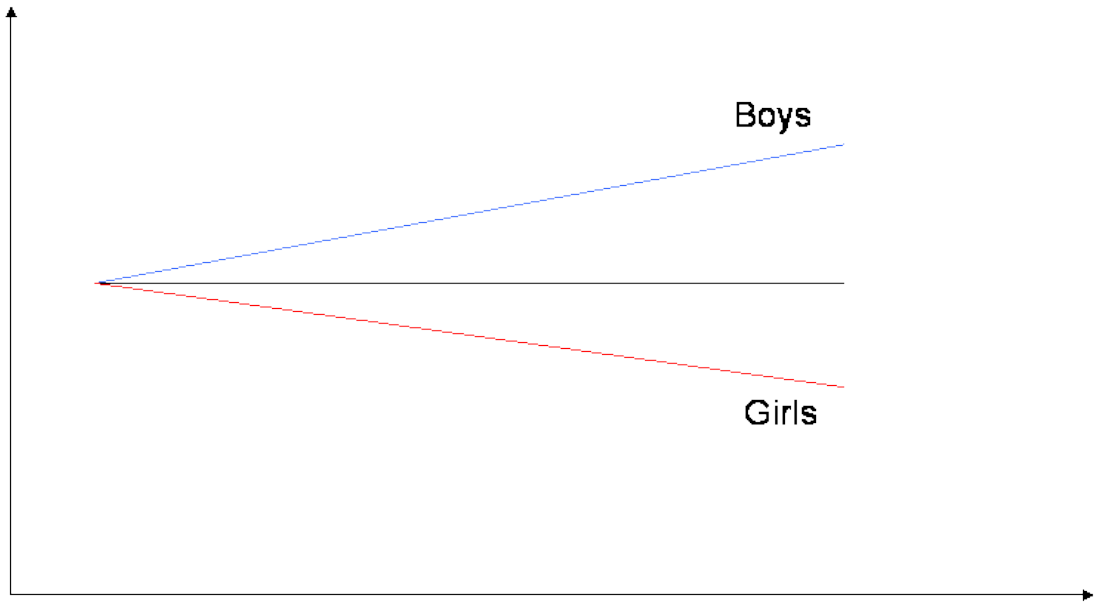
Intercept and slope random effects

- **Intercept random effect**: each level of the random effect (e.g., each individual, each pond) has a different intercept
- The relationship between the predictors and the response is the same for each individual, but the response values (levels) differ between individuals



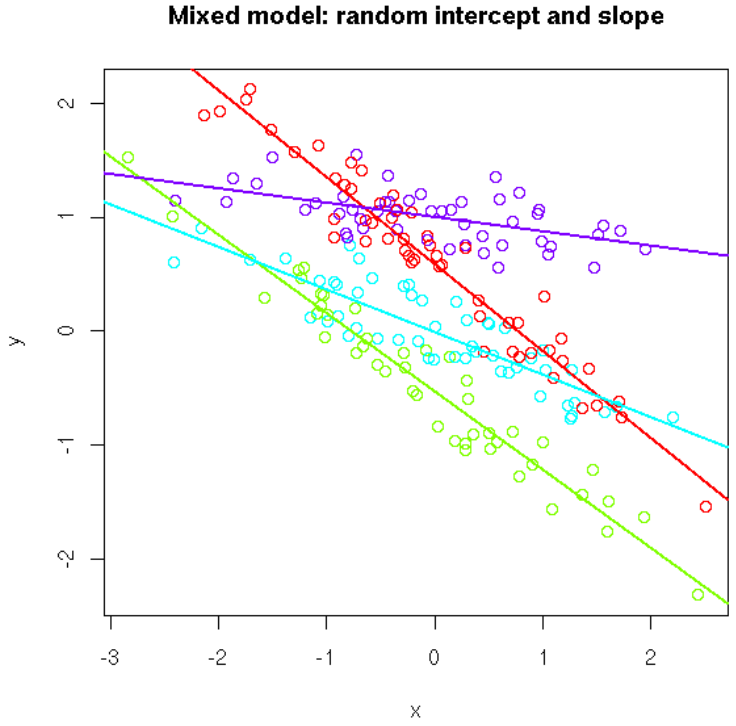
Intercept and slope random effects

- **Slope random effect**: each level of the random effect (e.g., each individual, each pond) starts at the same level of the response variable, but the relationship between the predictors and the response differs between individuals



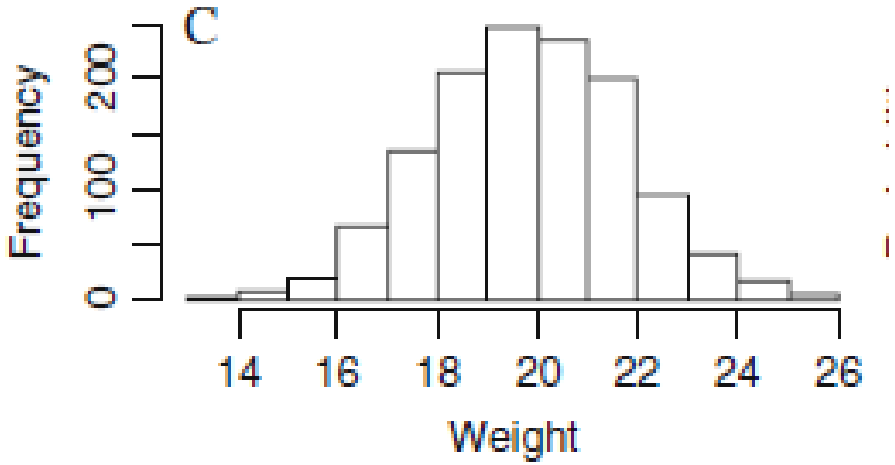
Intercept and slope random effects

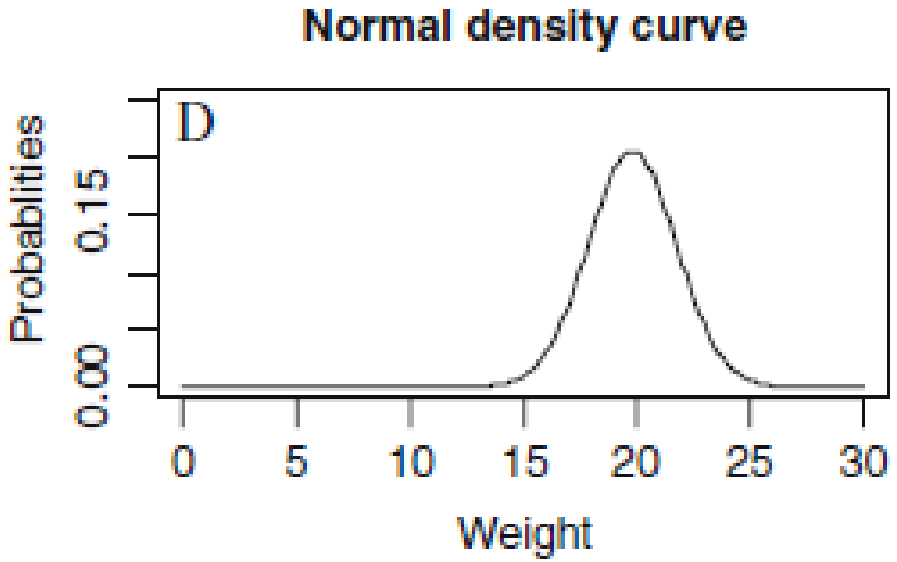
- **Intercept + slope random effects**
- This is where the “magic” comes in, correcting for lack of independence and allowing for extrapolation beyond your study system



Normal distribution

- Normal = Gaussian

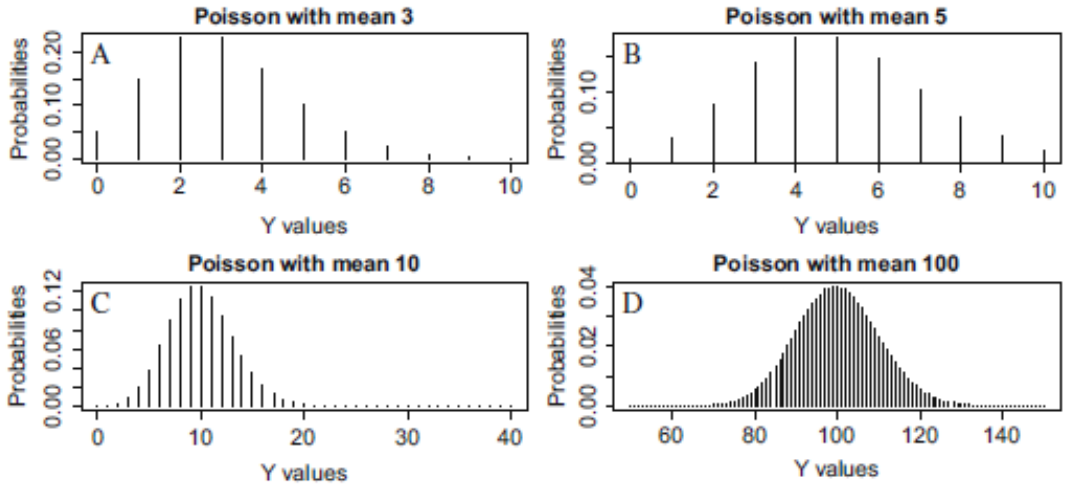




Zuur et al. 2009

Poisson distribution

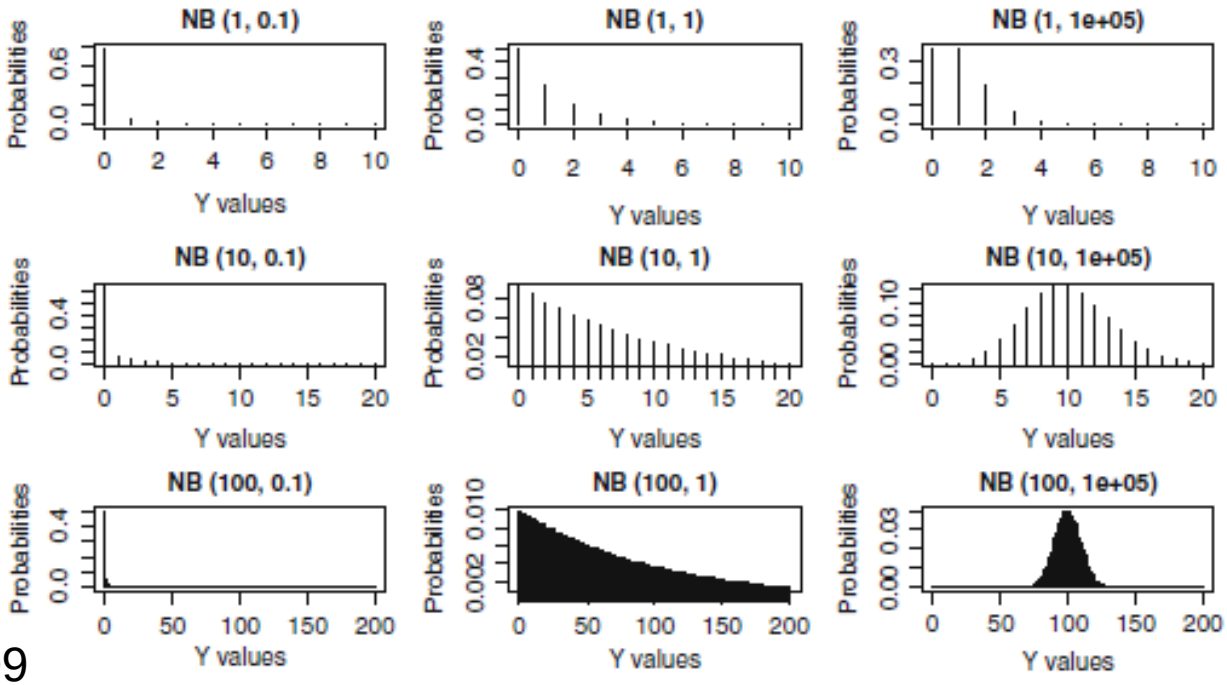
- Must be non-negative integers
- Mean = variance
- Typically recommended for count data
- However, count data is often *overdispersed*, meaning it has variance > mean (often due to lots of 0s or small values)
- Can be used to analyze density estimates with *offsets*



Zuur et al. 2009

Negative binomial distribution

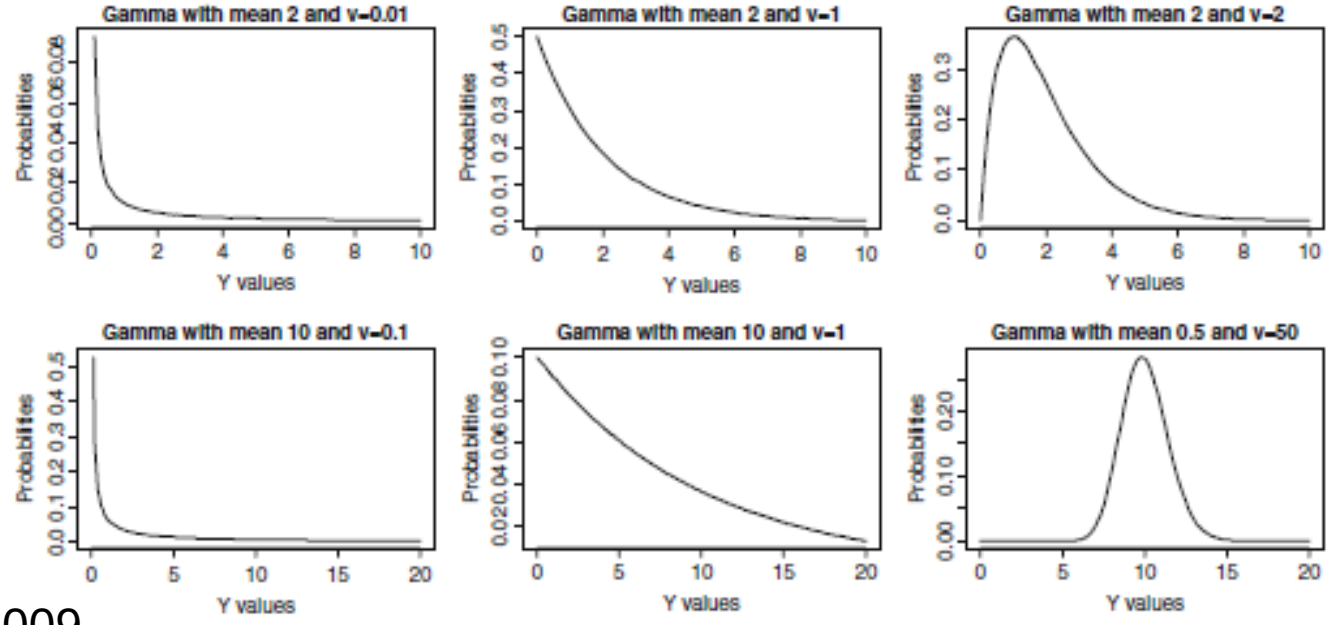
- Also frequently recommended for count data
- Accounts for overdispersion
- Combination of Poisson and gamma distributions
- Data must be non-negative integers



Zuur et al. 2009

Gamma distribution

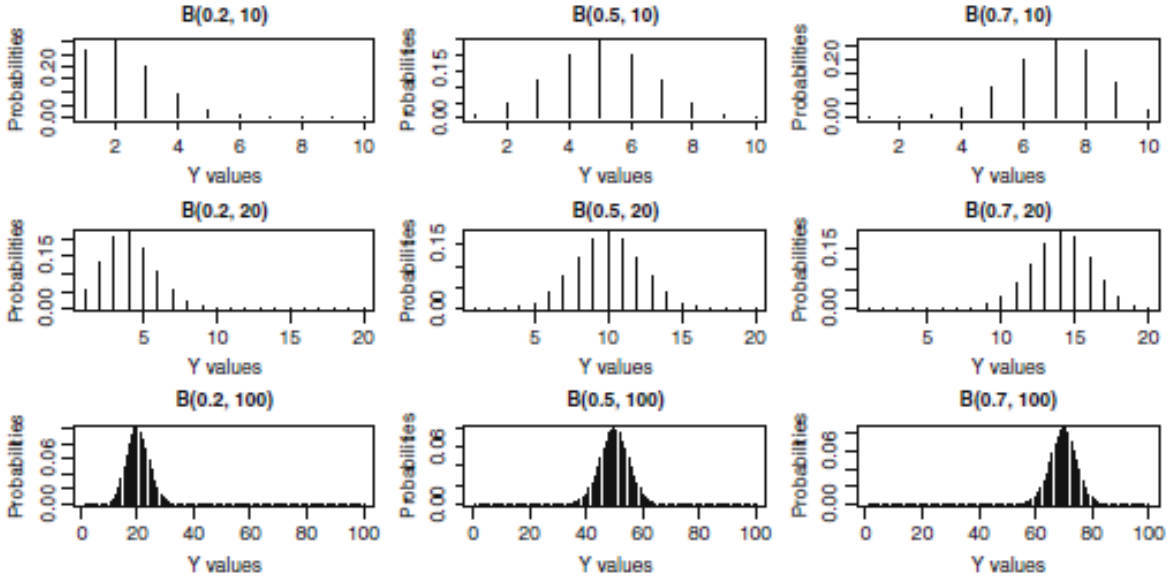
- For continuous data >0
- Uncommon, sometimes difficult to fit, but good for continuous data that can't be transformed



Zuur et al. 2009

Binomial distribution

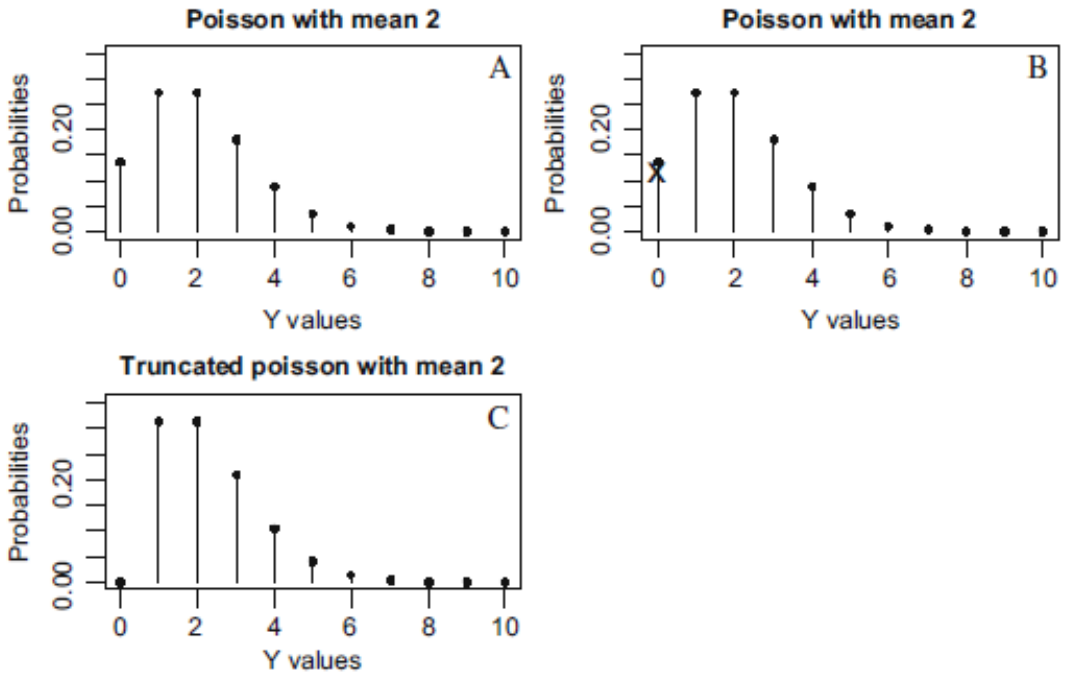
- Estimates probabilities
- For trials (0/1), e.g., presence/absence
- Also for proportions, if you can report them as successes/trials or success:failures (e.g., 10/15 samples contaminated; proportions can also be transformed)



Zuur et al. 2009

Zero truncated distributions

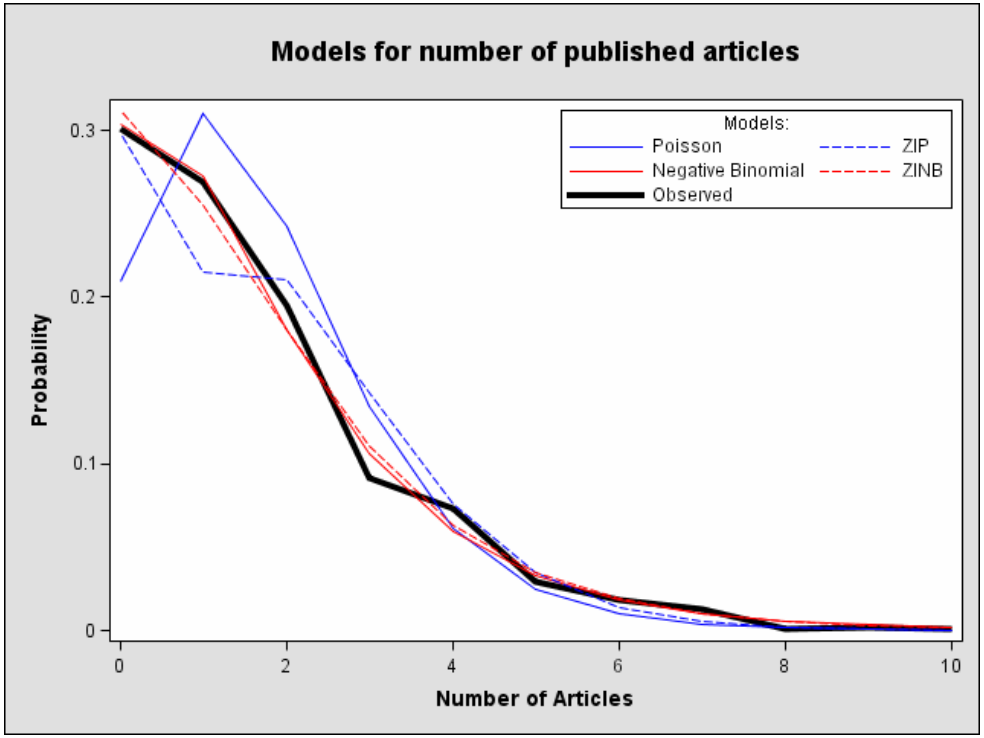
- Poisson and negative binomial distributions can be zero truncated, i.e., 0 values not allowed
- Rare in ecology, but could be used in toxicological studies where you have a lower detection limit but no absolute 0s



Zuur et al. 2009

Zero inflated distributions

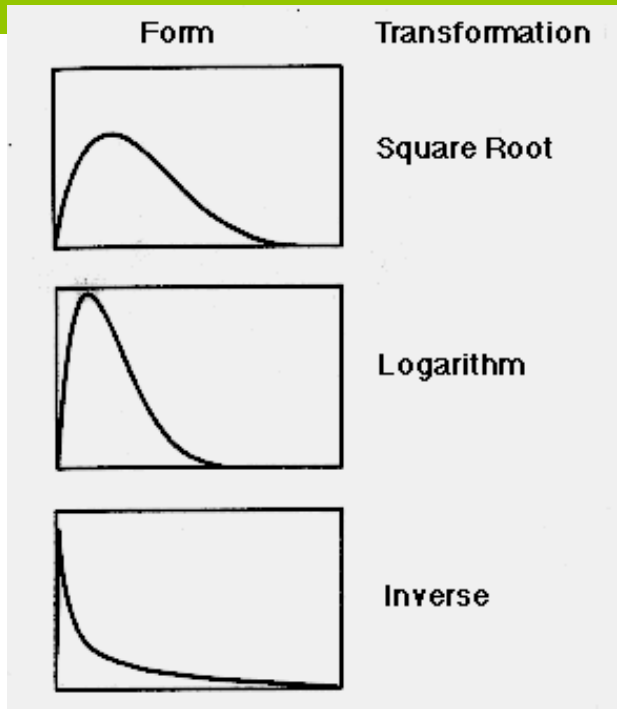
- Poisson and negative binomial distributions can also be zero inflated, better modeling data with many 0s
- Zero-inflated data common in ecology (e.g., bird count data)



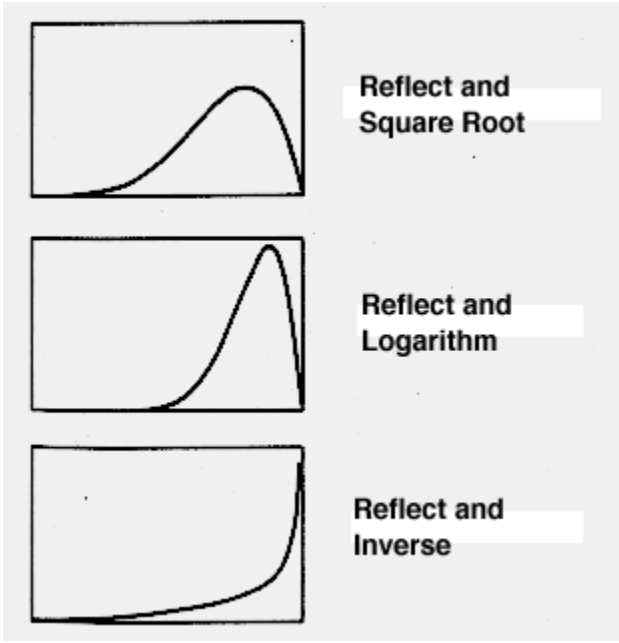
Transformations

Bio.georgiasouthern.edu

- Continuous data
  - Square root ($\geq$0)
  - Logarithm (ln, $\log_{10}$; >0)
  - Inverse (<>0)
  - Reflect = multiply by -1



- What if your data includes negative numbers or zeroes?
  - Add a fixed number to all data points
  - Old rule of thumb: add 0.1-0.5
  - New rule of thumb: add a value equivalent to your original value closest to 0
    - (-1.4, 0.15, 2.5) => add 1.4 = (0, 1.55, 3.9) => add 0.15 = (0.15, 1.7, 4.05)



Transformations - proportions

- Trials (successes/failures) should be analyzed with binomial distributions
- Sometimes proportions are not trials, e.g., % cover
- Previously: arcsine square root (arcsin(sqrt(y)).
  But:

Ecology, 92(1), 2011, pp. 3–10

The arcsine is asinine: the analysis of proportions in ecology

David I. Warton[1,2,3] and Francis K. C. Hui[1]

- Recommendation: logit transformation
  - Log(y/(1-y))

What about predictor variables?

- Distributions and transformations only apply to response variables
- However there's one major potential predictor variable problem to beware of: collinearity

EDITOR'S CHOICE

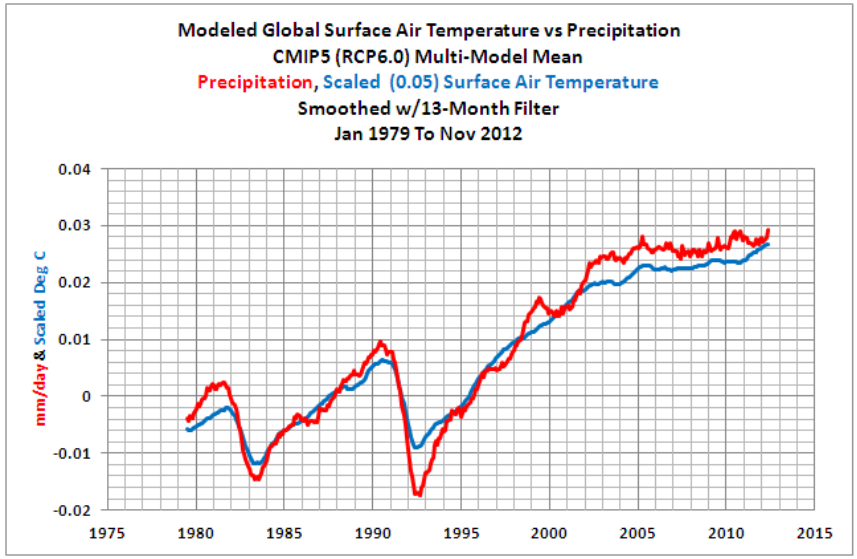
Ecography 36: 027–046, 2013
doi: 10.1111/j.1600-0587.2012.07348.x

Subject Editor: Marti Jane Anderson. Accepted 24 February 2012

Collinearity: a review of methods to deal with it and a simulation study evaluating their performance

Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell and Sven Lautenbach

Ecology, 84(11), 2003, pp. 2809–2815



CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION

MICHAEL H. GRAHAM[1]

When and why is collinearity a problem?

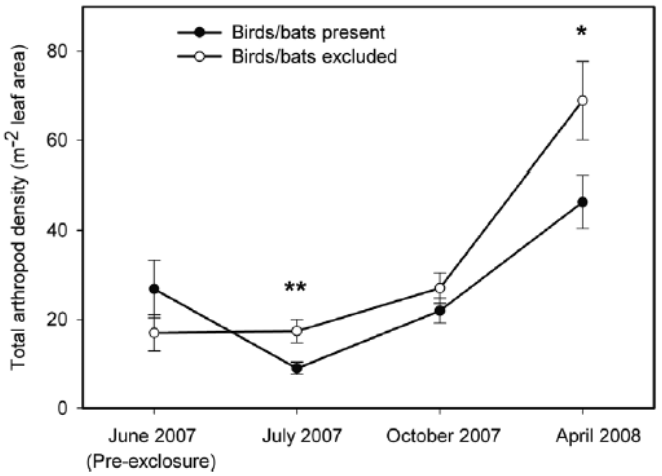
- When you have multiple correlated explanatory variables included in a model
- How to detect collinearity:
  - Calculate correlations. Dormann et al.: $r > 0.7$ problematic; Graham found $r \geq 0.28$ problematic
  - Calculate variance inflation factors (VIF). VIF > 5 or 10 considered problematic; Graham found VIF > 2 problematic
- Why it's a problem:
  - Inflates variance of regression parameters
  - Betas inaccurate
  - Reduces statistical power
  - Significant predictor variables may be excluded in favor of collinear variables

Dealing with collinearity

- Build small models *a priori* or remove collinear variables altogether (what if you miss important variables?)
- Enter collinear variables into a principal components analysis and use the PC axes as predictors (interpretation?)
- Regress one predictor on the other, and just include the residuals from the second predictor
  - PrecipResids <- residuals(lm(temp ~ precip, data=MyData))
- Use other modeling techniques
  - Sequential regression
  - Penalized regression techniques (ridge and LASSO regression)
  - Principal components regression
  - Structural equation modeling
  - Boosted regression trees

Repeated measures

- Often ecologists take repeated measurements on individual sampling units (e.g., birds, plants, ponds, tissues) over time.
- Repeated samples from the same individual are going to be more similar to one another than samples from another individual – they are *not independent*
  - A plant with 100 insects on it on week 1 is likely to have far more insects on it the next week than a plant with 10 insects on week 1
  - A deep pond will have a slower neonic degradation rate than a shallow pond
- Must analyze repeated measures in a single analysis using a random effect for the sampling unit

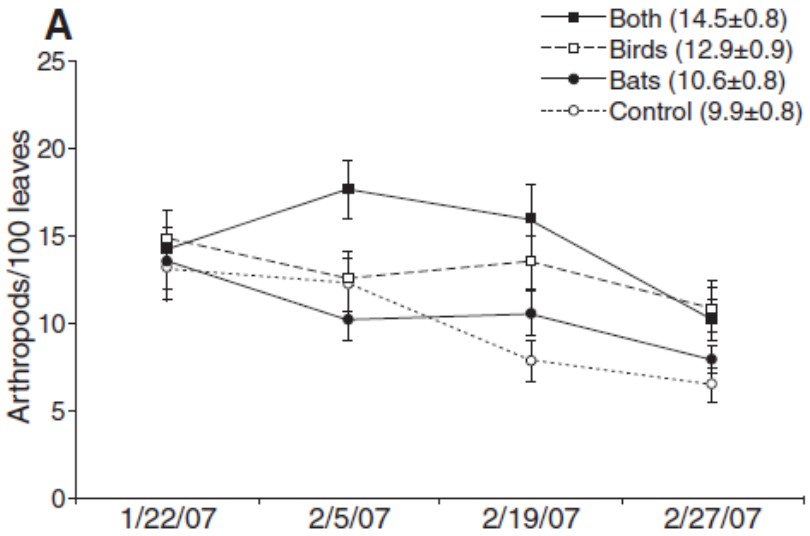


Michel et al. 2014 *J. Trop. Ecol*.

Should time be continuous or categorical?

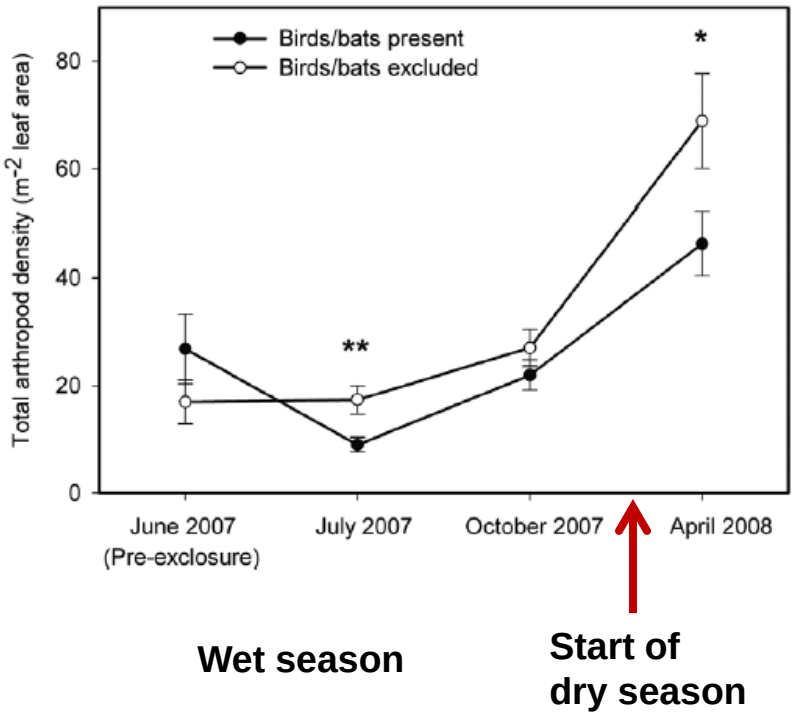
- Time can be analyzed as a continuous variable – giving you a slope/rate of change over time – or as a categorical variable – giving you absolute differences between points in time
- When to use continuous: data is sampled frequently at approximately equal intervals under similar conditions; you're interested in whether/how things changed
- When to use categorical: data is sampled infrequently, over unequal intervals, and under different conditions; you want to know the difference between 2 points rather than how it changed over the intervening time

Continuous



Williams-Guillen et al. 2008 *Science*
Sampled every 2 weeks during the dry season

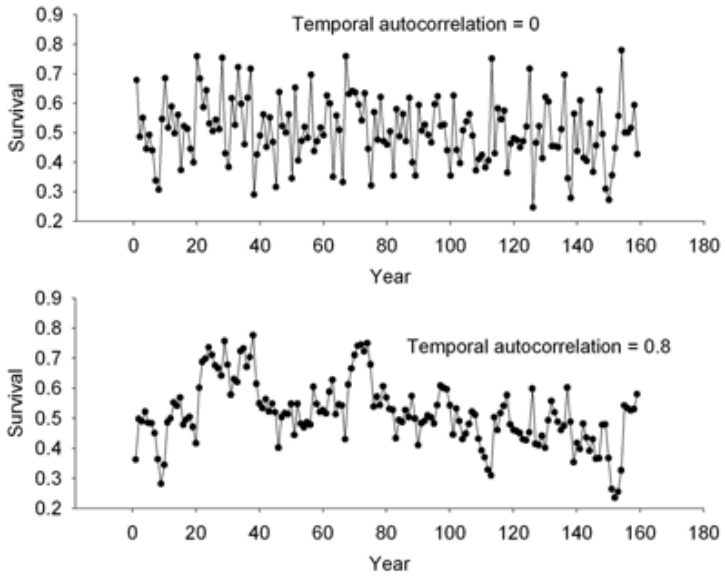
Categorical

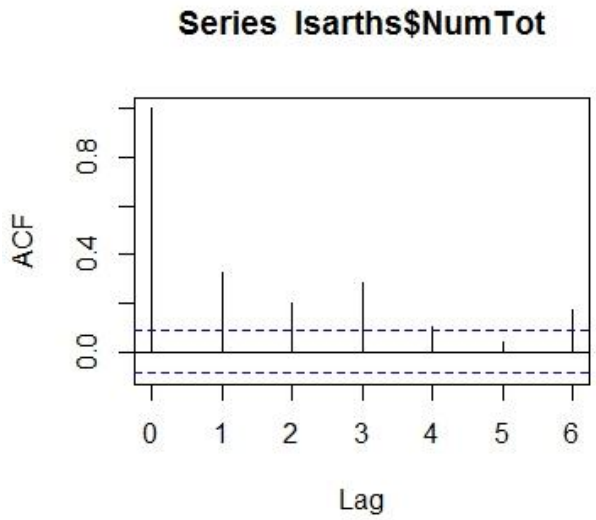


Michel et al. 2014 *J. Trop. Ecol.*
Sampled at uneven intervals
Spanned wet & dry seasons

Beware temporal autocorrelation!

- Temporal autocorrelation = when a variable's value at one point in time is correlated with it's value at another time
- Very common in ecology: weather, population density, etc.
- Why it's a problem: lack of independence
- How to detect it: autocorrelation functions (ACF)





Bars above/below dotted lines = significant autocorrelation

Resolving temporal autocorrelation

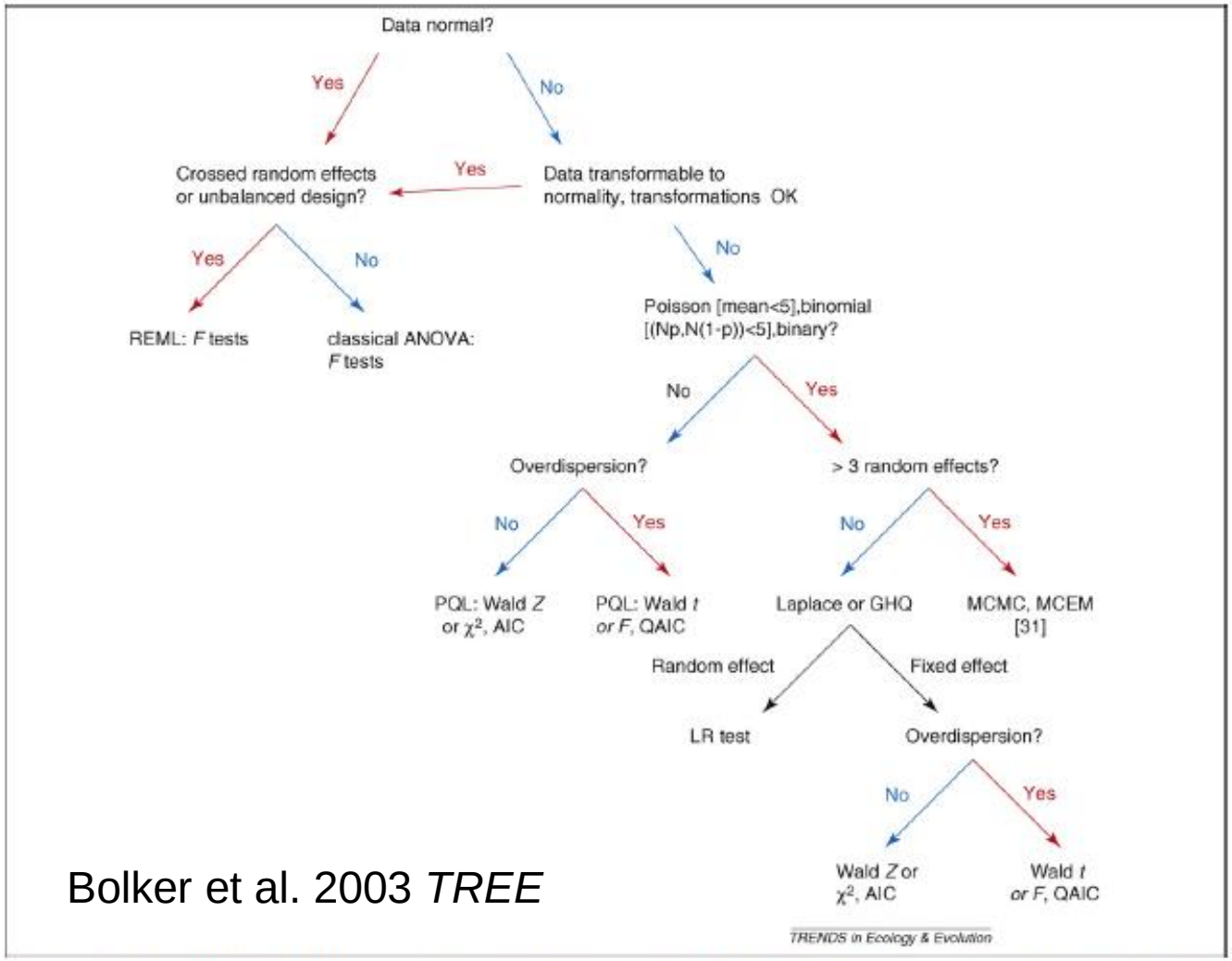
- Evaluate model residuals for autocorrelation using the pacf() function in R
- What to do about it: add correlation structures to your models (available in nlme, MASS, R_INLA)
- Common correlation structures (from nlme):
  - corAR1 – autoregressive of order 1 (lag=1 time period)
  - corARMA – autoregressive moving average process. Specify autoregressive (p) and moving average components (q). p and q usually between 0-2 (rarely 3)
  - corCAR1 – continuous autoregressive process
- For more information, see Zuur et al. 2009 chapters 6, 7, 14

Estimation techniques

- Simple, Gaussian models use maximum likelihood estimation based on sums of squares, like ANOVA
- Gets tricky with random effects, other distributions
- Alternative estimation techniques:
  - Pseudo- and penalized quasilikelihood
  - Laplace approximation
  - Gauss-Hermite quadrature
  - Markov Chain Monte Carlo (MCMC = Bayesian)
- Standard vs. restricted maximum likelihood estimation

- Don’t need to know the details! Just know that there are lots of methods, and it’s important to choose the right one for your data. Bolker et al. 2003 will help you (next slide)

Which estimation technique?



Bolker et al. 2003 *TREE*

Figure 1. Decision tree for GLMM fitting and inference. Conditions on the Poisson and binomial distributions along the right branch refer to penalized quasilikelihood (PQL) rules of thumb [30]: to use PQL, Poisson distributions should have mean > 5 and binomial distributions should have the minimum of the number of successes and failures > 5. MCEM = Monte Carlo expectation-maximization [40].

Which R package/command?

Table 1. Capabilities of different software packages for GLMM analysis: estimation methods, scope of statistical models that can be fitted and available inference methods

		Penalized quasilikelihood	Laplace	Gauss-Hermite quadrature	Crossed random effects	Wald χ^2 or Wald F tests	Degrees of freedom	MCMC sampling	Continuous spatial/ temporal correlation	Overdispersion	
SAS	PROC GLIMMIX	✓	✓*	✓*	✓	✓	BW, S, KR		✓	QL	
	PROC NLMIXED			✓		✓	BW, S, KR			Dist	
R MASS	glmmPQL	✓				✓	BW		✓	QL	B, P, QP, G
	glmmML		✓	✓							B, P
lme4	glmer		✓	(✓)	✓			(✓)		QL	B, P, QP
	glmmADMB		✓							Dist	B, P, NB, Z
	GLMM	✓			✓?	✓			✓	QL	B, P, NB, Z
GenStat/ ASREML			✓	✓	✓			✓		Dist	
AD Model Builder		✓	✓		✓					✓	
HLM				✓							
GLLAAMM (Stata)								✓		Dist	
WinBUGS					✓			✓			

Abbreviations: BW, between–within; dist, specified distribution (e.g. negative binomial); KR, Kenward–Roger; QL, quasilikelihood; S, Satterthwaite.
*Version 9.2 only.

Bolker et al. 2003 *TREE*

Also:

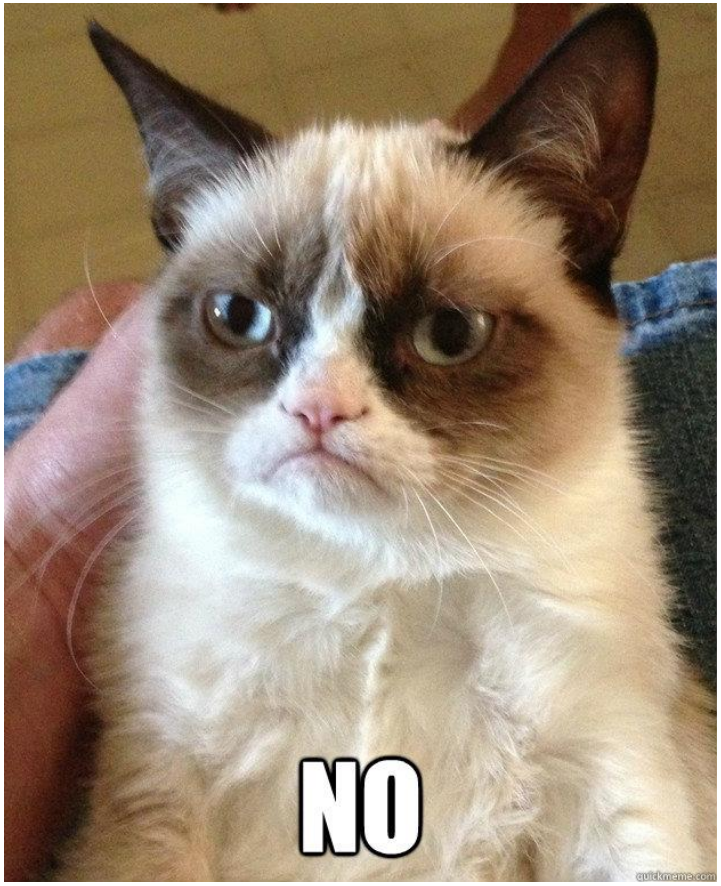
nlme: Gaussian linear and non-linear mixed models. Handles nested but not crossed random effects. Slower than lme4

betareg: beta regression for beta-distributed variables (rates, proportions)

MCMCglmm: MCMC methods for glmms, can include Bayesian priors

R-INLA: New. Nested Laplace, hierarchical models, Poisson, binomial, negative binom

But I found a script online / in a book / got one from a friend. Can't I just plug my data into that?



Why build your own models?

- As should be apparent by now, GLMMs are powerful tools that correct for multiple sources of bias, heterogeneity, and lack of independence

- But they're also complex: "GLMMs are surprisingly challenging to use even for statisticians" – Bolker et al. 2003 (Benjamin Bolker is a statistician, GLMM expert, and coauthor of lme4)

- GLMMs need to be customized to correct for lack of independence and heterogeneity in *your specific dataset*

- You can use other scripts as examples or templates, but they're not cookie-cutters like, e.g., ANOVA or t-tests

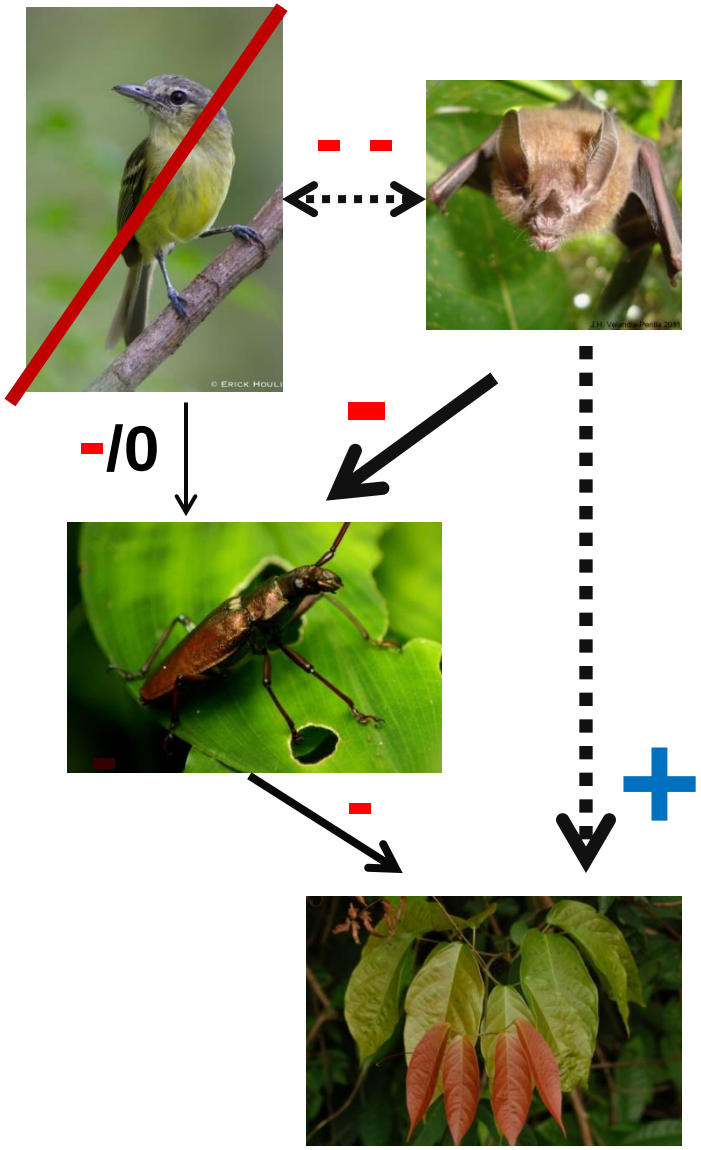
GLMM steps

1. Think carefully about your data and the question you're asking! Describe or diagram your experiment. Identify fixed and random effects.
2. Inspect your data. Look for entry errors, outliers, missing data, collinearity, etc.
3. Identify the appropriate distribution(s)
4. Explore responses to treatments/fixed effects
5. If repeated measures: test for temporal autocorrelation
6. Select package and command
7. Fit full model with all fixed effects, both slope and intercept random effects (where applicable)
8. Inspect residuals, test for non-normality and autocorrelation or overdispersion (where applicable)
9. Add correlation structures, random effects, or polynomial terms. Where >1 distribution or estimation technique is possible, try alternate methods and use AIC to select
10. Repeat 8-9 until model fits (or close to it)
11. Remove slope and intercept random effects, use AIC to select model
12. Remove fixed effects, use AIC to select model
13. Evaluate residuals of final model and report.

About the data

- Part of a larger study during my PhD investigating relative effects of birds and bats in limiting arthropods and herbivory at a site with an intact bird community vs. a site experiencing insectivorous bird declines

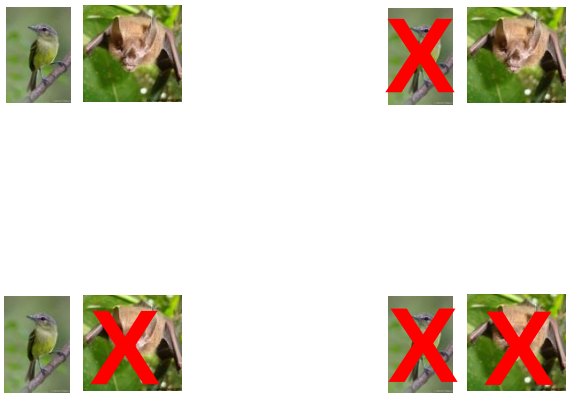
- 80 plants of 3 families with 4 treatments (control, birds excluded, bats excluded, birds & bats excluded), sampled every 10 days for 60 days (=7 sampling periods)



About the data

- Response variables: arthropod counts (density when offset by leaf area), herbivory (% damage), Leaf Damage Index (final LDI > initial LDI; binary 0/1)
- Predictor variables: treatment, time (sampling period), leaf area, plant family, block, plant ID, pre-treatment count

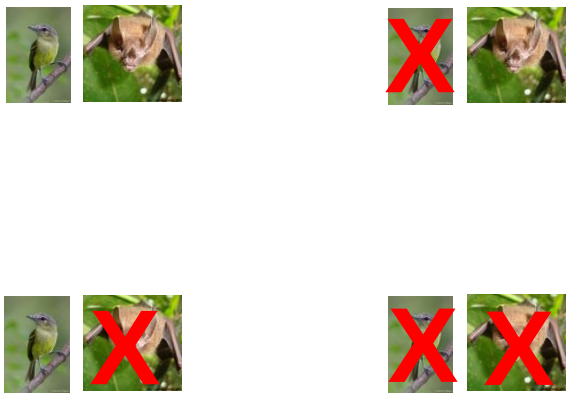
1 Block (n=20)



Describing the data pre-analysis

- Fixed effects: treatment, time, pre-treatment count
- Random effects: plant nested within block nested within plant family
- Time: continuous, may have temporal autocorrelation, may be nonlinear responses
- Collinearity: not a concern here

1 Block (n=20)



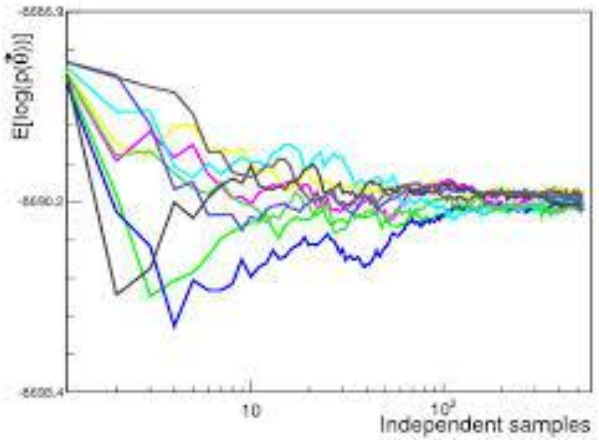
Other comments

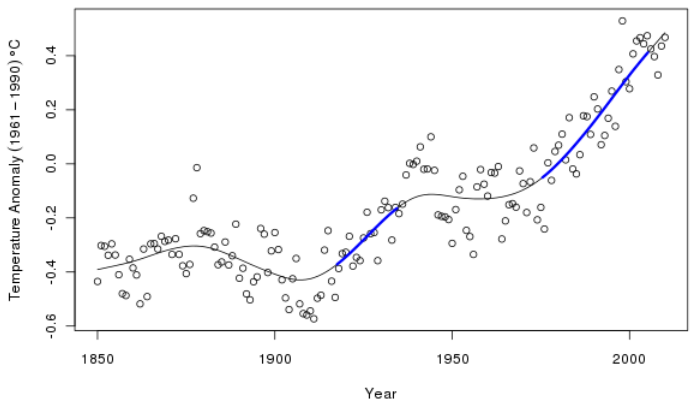
- Contrasts
  - If you have significant interactions, or categorical predictors with >2 levels, you need to run post-hoc tests
  - If predictors are categorical use package phia to build and run custom contrasts
  - If one or more predictors are continuous, use command glht in package multcomp (runs Tukey tests)

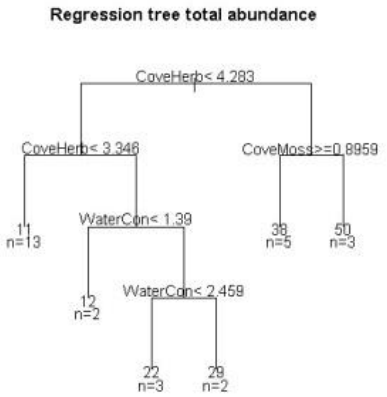
- R^2 values (model fits: variance explained by predictors)
  - lme: use function saved in “rsquared.lme.R” script
  - lm: displayed in model output
  - Binomial GLMs: use command Rsq in package binomTools
  - GLMMs: use command rsquared.GLMM in package MuMIn

What if GLMMs aren't enough?

- Welcome to the wonderful worlds of Bayesian statistics (MCMC, hierarchical modelling), general additive modelling (GAMs/GAMMs), and/or machine learning (boosted regression/classification trees)
- Highly flexible – and highly complex – techniques that account for even more sources of non-independence than GLMs/GLMMs
- See Zuur et al. 2009 (GAM/GAMM); Kéry 2010, Kéry & Schaub 2012 for Bayesian methods, Elith et al. 2008 for BRTs







Some resources

- Bates, D. lme4: Mixed-effects modeling with R. http://lme4.r-forge.r-project.org/book/
- Bolker, B.M., M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H.H. Stevens, and J.-S.S. White. 2009. Generalized linear mixed models: A practical guide for ecology and evolution. Trends in Ecology & Evolution, 24:127–135. (and supps.)
- Elith, J., J.R. Leathwick, and T. Hastie 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802-813.
- Kéry, M. 2010. Introduction to WinBUGS for ecologists. Academic Press, MA.
- Kéry, M., and M. Schaub. 2012. Bayesian population analysis with WinBUGS. Academic Press, Burlington, MA.
- Pinheiro, J.C., and D.M. Bates. 2000. Mixed-effects models in S and S-PLUS. Springer, New York.
- Zuur, A.F., E.N. Ieno, N.J. Walker, A.A. Saveliev, and G.M. Smith. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, New York.

- R-sig-mixed-models FAQ (comprehensive review of everything you ever wanted to know – and more – about running GLMMs in R): http://glmm.wikidot.com/faq & http://glmm.wikidot.com/pkg-comparison
- Questions? Stuck with your code? Google is your friend! There are tutorials, lecture slides, blogs, and message boards with tons of helpful info, especially Cross Validated (http://stats.stackexchange.com/) and http://R-bloggers.com. See also Andrew Gelman's blog (http://andrewgelman.com/) and Benjamin Bolker's webpage (http://ms.mcmaster.ca/~bolker/). Or contact me: nicole.l.michel1@gmail.com.