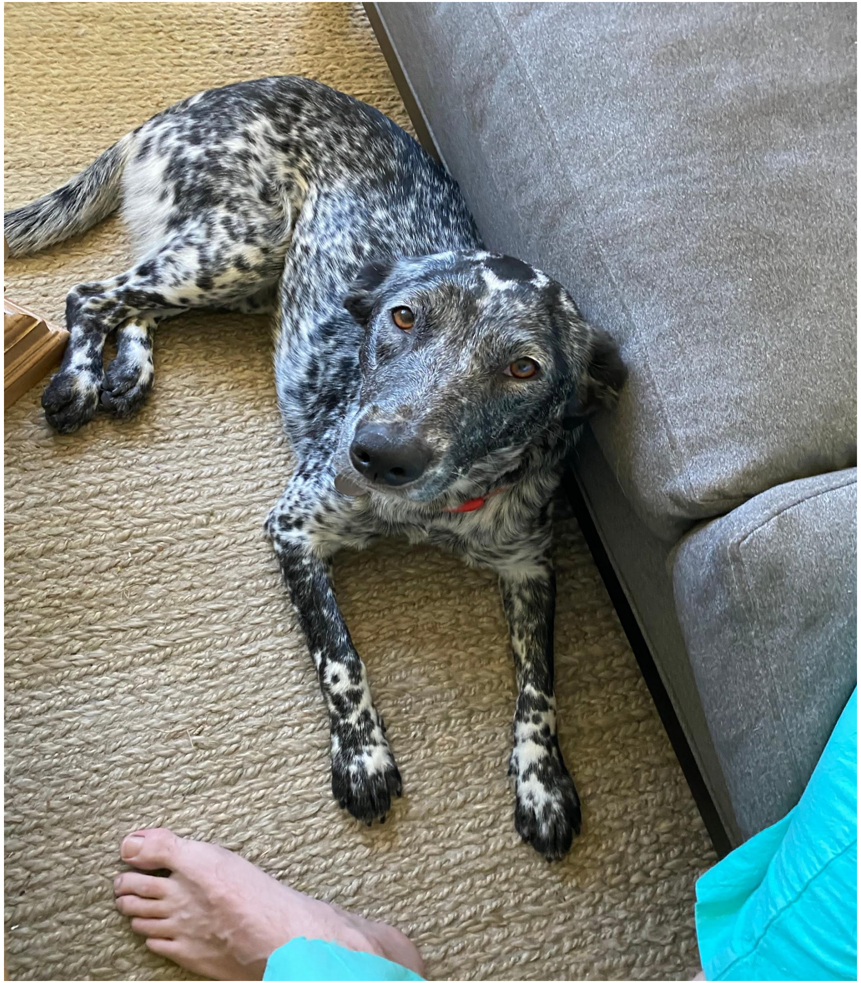
Generalized Linear Mixed Models (GLMM)

A step closer to actually analyzing ecological data

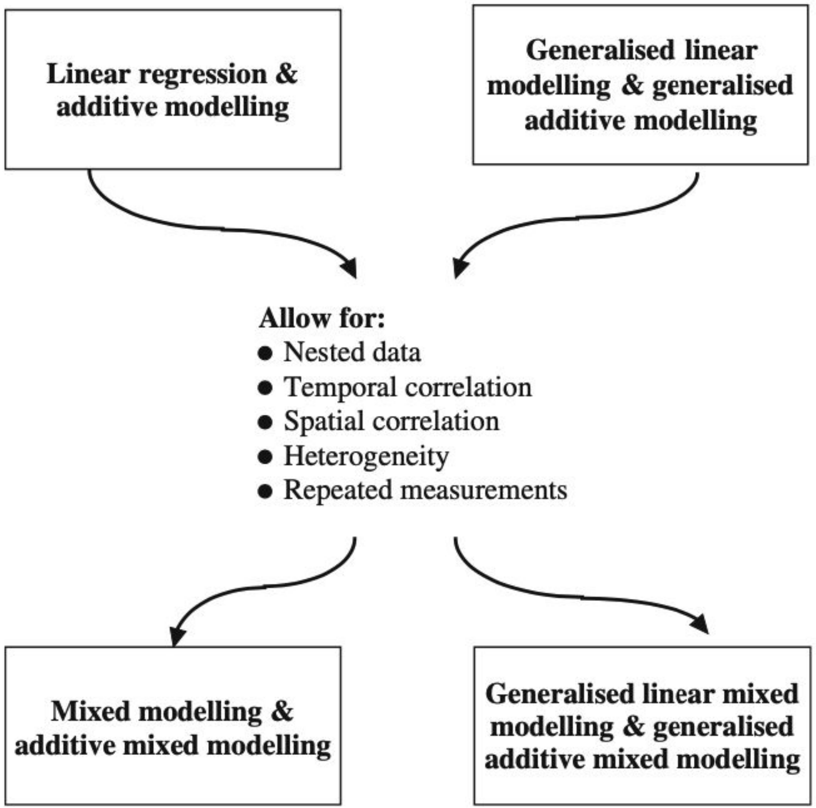
By: Zoë Zilz & Sam Sambado

Agenda

- Where does GLMM fit in GLAMM?
- What’s cool about GLMM
- Some math about GLMM
- Why tred cautiously using GLMMs
- R examples!



The statistical journey of GLAMM



Why use GLMMs?

- Combines tools you know (GLMs & LMMs)
  - Follow similar steps used in mixed modelling
  - It’s a linear mixed model with Gaussian distribution
- Most ecological data is non-normal with random effects
- Common to understand random variation in space & time or among individuals
- However, challenging to estimate parameters & have statistical inference

Underlying mathematics in GLMMs

- Conditional vs. unconditional distributions
- Likelihood function for GLMMs
- How is it calculated
- You know most of this already

Linear mixed model:

$$Y_{ij} = \alpha + X_i \times \beta + Z_i \times b_i + a_i + \varepsilon_{ij}$$

Conditional Distributions

Poisson conditional

1. Conditional on the random effects b_i, the counts Y_{ij} are assumed to be **Poisson distributed** with mean $\mu_{ij}|b_i$
   - mean and variance of Y_{ij}: $E(Y_{ij}|b_i) = var(Y_{ij}|b_i)$.

2. The relationship between the conditional mean and the explanatory variables is determined by the **log link** : $\log(\mu|b_i)=\alpha+X_i\times\beta+Z_i\times b_i$.

3. The **random effects** $\boldsymbol{b_i}$ are assumed to be **normally distributed** with mean 0

*** specification of conditional distribution & random term unlike ordinary Poisson GLM***

Binomial conditional

1. Conditional on the random effects bi, the presence and absence data Yij are assumed to be **binomial distributed** with probability $pij|bi$.
   - Mean: $E(Yij|bi) = pij|bi$
   - Var : $(Yij|bi) = pij|bi \times (1 - pij|bi)$

2. The relationship between the conditional mean and the explanatory variables is determined by the **logistic link** $\text{logit}(pij|bi) = \alpha + Xi \times \beta + Zi \times bi$.

3. The **random effects** $\boldsymbol{bi}$ are assumed to be **normally distributed** with mean 0

Maximum likelihood

- For GLM, ML is specified & deviates with respect to parameters are calculated & set to 0
- For GLMM, parameters may not be calculated or contain a deviance & AIC
    - Some packages & methods can, but interpret with care

$$L(\beta, D) = \prod_i \int f(Y_{ij}|b_i) \times f(b_i) db_i$$

- Option: use numerical integration techniques & replace integral by a summation (i.e. Gaussian quadrature)
    - High computational burden

Proceed with caution.

311/537 (58%) GLMM analyses used inappropriately (Bolker 2009)

- Pitfalls of estimation, particularly with PQL
    - Poisson : < 5 counts/treatment
    - Binomial : < 5 # of successes/failures for each observation

GLMM touches on controversial stat issues

- Null hypothesis testing, validity of stepwise regression, use of Bayesian stats

Estimation of parameters

- Fixed-effect parameters
  - (i.e. effects of covariates, differences among treatments & interactions)
- Random-effect parameters
  - (i.e. sd of random effects)
- To find ML estimates for GLMM, integrate likelihoods over all possible values of random effects
  - **Penalized quasilikelihood (PQL)**, Laplace approximations, Gauss-Hermite quadrature (GHQ), MCMC algorithms

Inference for GLMMs

3 types of inference: hypothesis testing, model comparison, & Bayesian approaches

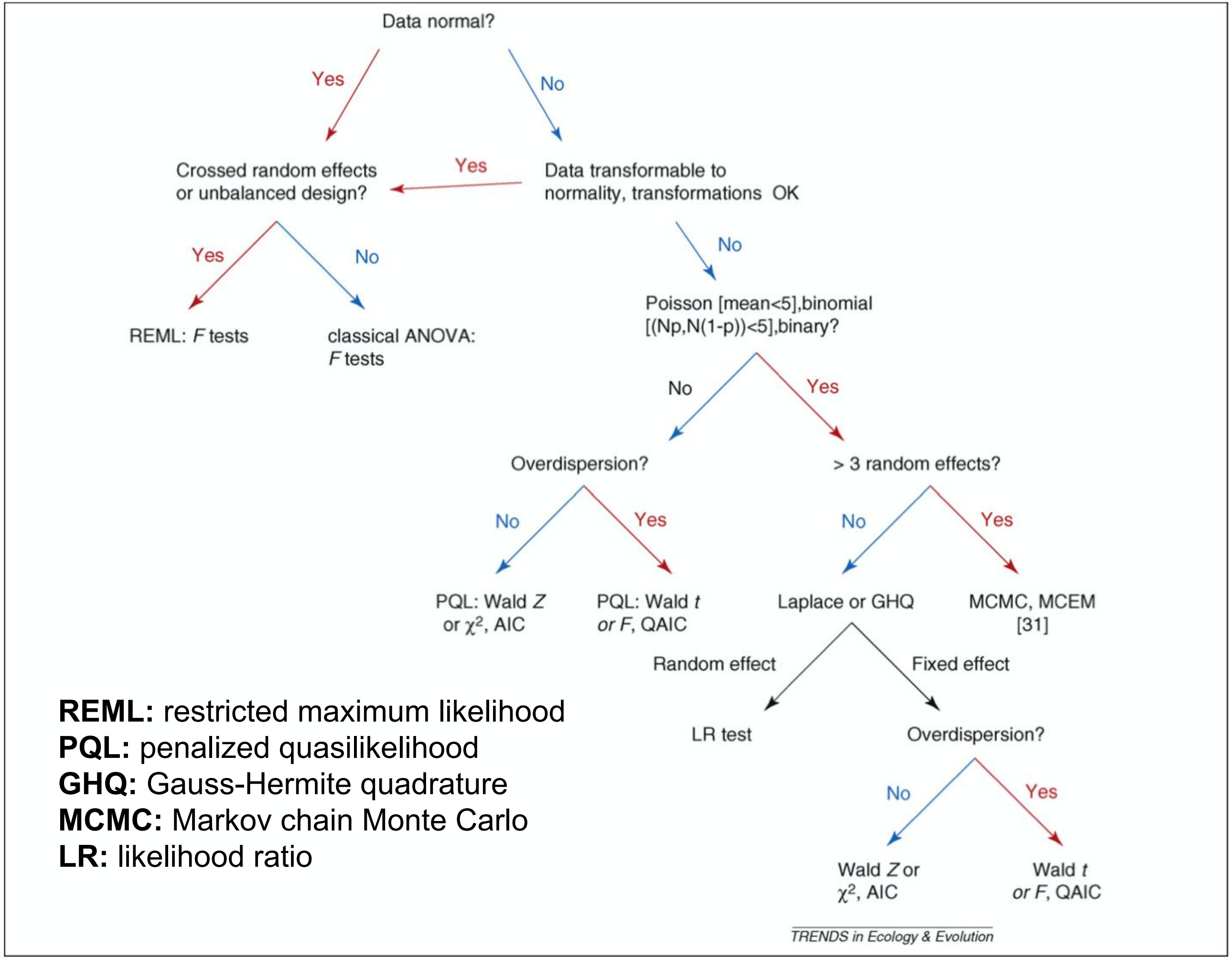
- **Hypothesis testing**: frequentists test statistics for expected distributions
  - Wald Z, chi-square (without overdispersion)
  - Wald t, F tests (with overdispersion)
- **Model selection**: compares fits of candidate models for predictions
  - Estimate magnitude of differences between models
  - AIC & IC
- **Bayesian**: general scope, yet different route
  - MCMC provides confidence intervals for fixed & random effects
  - BIC and DIC (weaker assumptions)

A more helpful table on inference.

Table I. Techniques for GLMM inferences, their advantages and disadvantages and the software packages that implement them

Method	Advantages	Disadvantages	Software
Wald tests (Z, χ^2, t, F)	Widely available, flexible, OK for quasilikelihood (QL)	Boundary issues; poor for random effects; t and F require residual df	GLIMMIX, NLMIXED (SAS), glmmPQL (R)
Likelihood ratio test	Better than Wald tests for random effects	Bad for fixed effects without large sample sizes; boundary effects; inappropriate for QL	NLMIXED (SAS), lme4 (R)
Information criteria	Avoids stepwise procedures; provides model weights and averaging; QAIC applies to overdispersed data	Boundary effects; no p value; requires residual df estimate for AIC_c	GLIMMIX, NLMIXED (SAS), lme4 (R)
Deviance information criterion	Automatically penalizes model complexity	Requires MCMC sampling	WinBUGS

Decision tree for GLMM fitting & inference



REML: restricted maximum likelihood
PQL: penalized quasilikelihood
GHQ: Gauss-Hermite quadrature
MCMC: Markov chain Monte Carlo
LR: likelihood ratio

Bolker 2009

Create a full model

1. Specify fixed & random effects
2. Choose error distribution & link function
3. Graphical checking
4. Fit fixed-effects GLMs to full data set & within each level of random factors
5. Fit the full GLMM
6. Recheck assumptions

Zoë (& June) are going to go over some code & graphical tricks!