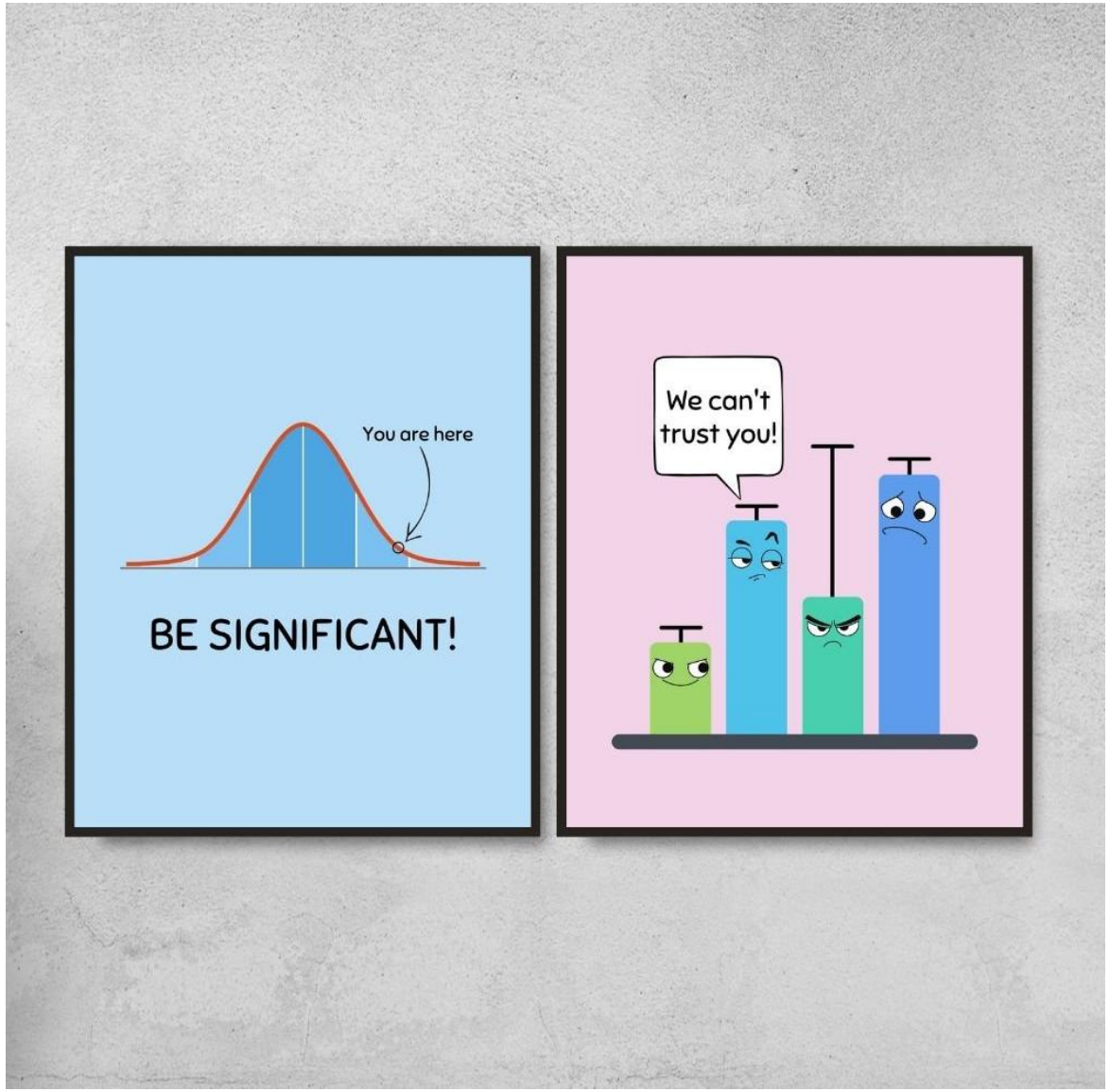


Generalized Estimating Equations & Generalized Linear Mixed Effect Models

Reminders/Updates

1. Tuesday, Nov 18: GLMM workshop (week 10) due: Nov 21
2. Thursday, Nov 21: What to write
3. Review: Nov 25 or Dec 2?
4. Final Project: Dec 5th w/ workshop
5. Anonymous Feedback – please ☺

My data are not independent – what do I do?

For normal data, we can use:

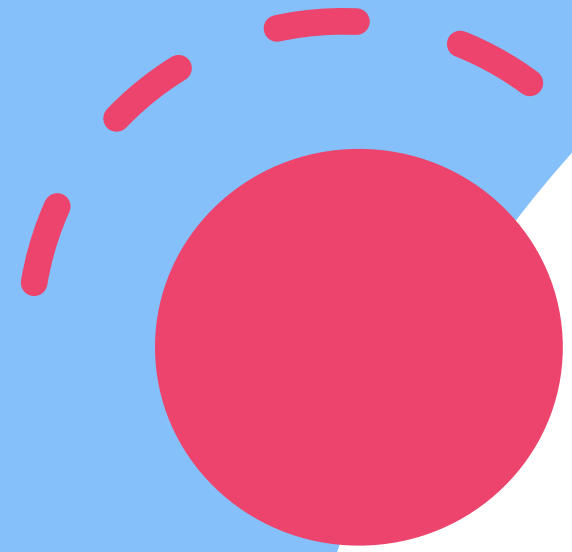
- temporal and spatial correlation structures (GLS)
- linear mixed effects models with random effects (LMM)

For non-normal data, there are two options:

- GEE: generalized estimating equations
- GLMM: generalized linear mixed effects models

Outline for Today

- Generalized Estimating Equations
  - What are GEEs?
  - When and how do we use GEE?
  - Example
- Generalized Linear Mixed Models
  - What are GLMMs?
  - Why are they more complicated than LMMs?
  - Example



Generalized Estimating Equations (GEE)

What are GEEs?

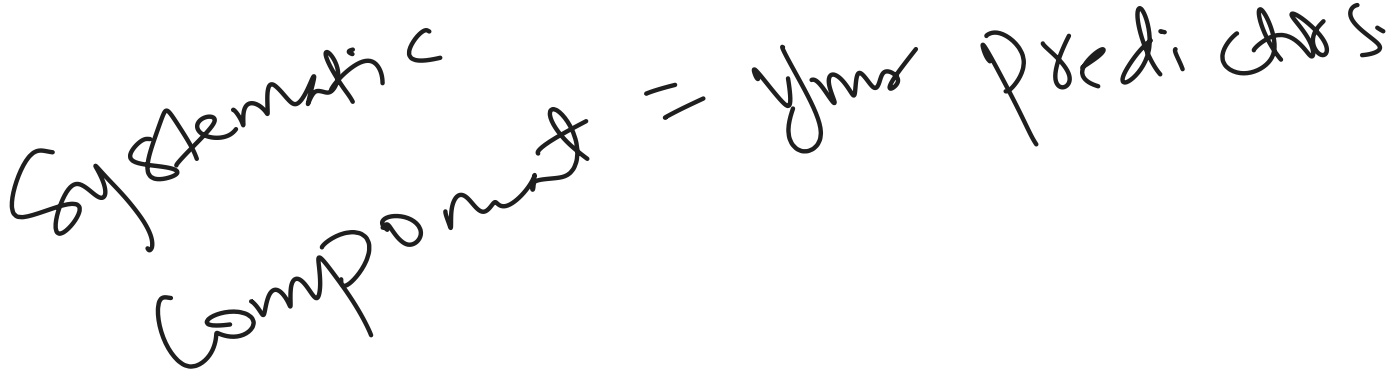
- Extension of GLMs that incorporate a correlation structure.
- The name refers to a set of equations that are solved to obtain parameter estimates (i.e., model coefficients).
- Designed for population-averaged inference: estimates the average effect of covariates across the entire population rather than subject-specific effects.

When to use GEEs

- GEE estimates population-averaged (marginal) effects. Use when you want to understand the overall effect of predictors on the entire population rather than individual or group-specific effects.
  - When you have non-normal outcomes: GEEs extend to various distributions (e.g., binomial, Poisson), making them flexible for binary, count, or other non-normally distributed outcomes.
  - When observations are correlated within groups
  - Longitudinal or repeated-measures data where individuals are measured multiple times over time.

How do GEEs work?

1. Specify systematic component and link function (just like in a GLM)
2. Choose a variance function (family)
3. Choose a correlation structure
4. Estimate model parameters (population-averaged effects) using quasi-likelihood
5. Estimate coefficients with robust standard errors using sandwich estimator



Step 1 and 2: Link and variance

1. Count data
    - Distribution: Poisson
    - Link: Log
2. Binary data
    - Distribution: Binomial
    - Link: Logit
3. Continuous data
    - Distribution: Gaussian, Gamma
    - Link: identity, Log

We are ***not*** assuming our data come from a poisson or binomial distribution.

We only adopt the mean and variance assumptions from the distribution.

Step 3: Correlation structure

Common structures include:

1. **Independence**: Assumes no correlation (often used as a benchmark).
2. **Exchangeable**: Assumes all measurements within a cluster have the same correlation.
3. **Autoregressive (AR)**: Assumes correlations decay over time.
4. **Unstructured**: Allows each pair of observations within a cluster to have its own correlation.

How do I choose?

Think-Pair-Share

Consider your own dataset (or a dataset you've recently worked with).

What link and variance function?

Which correlation structure would be most suitable?

- independence
- exchangeable,
- autoregressive (AR)
- unstructured.

How to choose a correlation structure?

Either choose the most general one your data can support (depending on sample size) or you can choose one you think suits the data best. Either way, don’t sweat it!

Step 4: Estimate model parameters

- Estimate regression coefficients for population-averaged effects.
- Parameters are estimated using **quasi-likelihood** methods, which don't require a fully specified likelihood; this is useful when the exact distribution of the data is unknown.
- Quasi-likelihood estimates are obtained by solving **estimating equations** that balance the expected value of the residuals for each predictor.

Step 5: Estimate Robust Standard Errors

- GEEs use a **sandwich estimator** (also called the robust or empirical variance estimator) to adjust the standard errors of the model parameters.
- Ensures unbiased standard errors even if correlation structure is misspecified by adjusting the variance-covariance matrix based on observed data
- Suitable for complex, unknown correlation structures common in ecology

The great thing about GEEs:
even if the variance-covariance matrix is mis-specified, the sandwich estimator converges to the true variance-covariance of the model parameters.

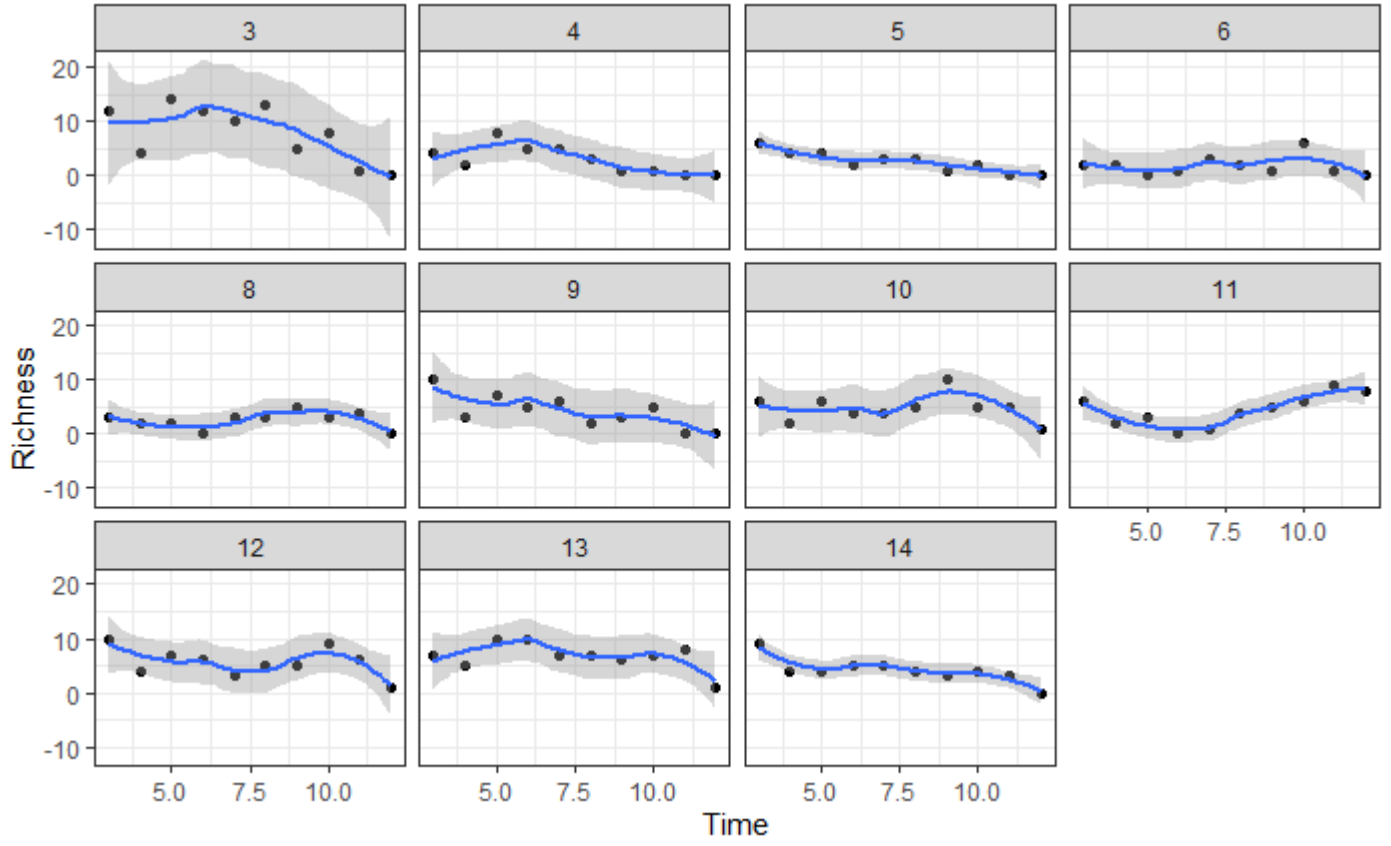
GEE - Bird Richness

Response: Aquatic Bird Richness

Predictors:

- Management Type
- Flooding Depth
- Size of Field (Area)

Distribution?
Correlation structure?
Grouping structure?



GEE - Bird Richness

```
library(geepack)

M.gee1<-geeglm(Richness~LA+fSptreat+DEPTH,
               data = RFBirds,
               family=poisson,
               id=fField,corstr = "ar1")

summary(M.gee1)
```

```
Coefficients:
             Estimate  Std.err  Wald Pr(>|W|)
(Intercept)   0.37484  0.26844  1.95    0.163
LA            0.56764  0.10984 26.71  2.4e-07 ***
fSptreatrlfld -0.31567  0.16751  3.55    0.059 .
DEPTH         0.00838  0.00544  2.37    0.123
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = ar1
Estimated Scale Parameters:

  Link = identity

Estimated Correlation Parameters:
Number of clusters:   11  Maximum cluster size: 10
```

1. specify a distribution with the family option, even though we are not assuming any distribution directly.
2. Grouping structure with the id option
3. Correlation structure with corstr option

Scale parameters

	Estimate <dbl>	Std.err <dbl>
(Intercept)	1.87	0.238

Correlation parameters

	Estimate <dbl>	Std.err <dbl>
alpha	0.378	0.0669

GEE - Bird Richness

- Use different correlation structures to confirm robustness of results

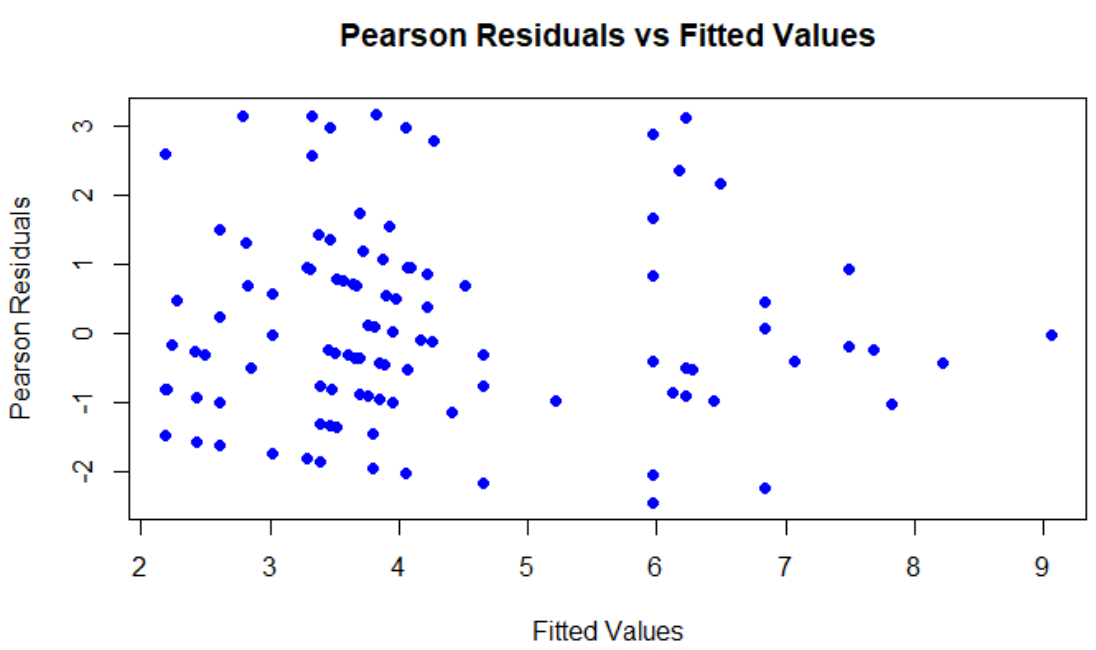
```
M.gee2<-geeglm(Richness~LA+fSptreat+DEPTH,
               data = RFBirds,
               family=poisson,
               id=fField,corstr = "exchangeable")

M.gee3<-geeglm(Richness~LA+fSptreat+DEPTH,
               data = RFBirds,
               family=poisson,
               id=fField,corstr = "independence")
```

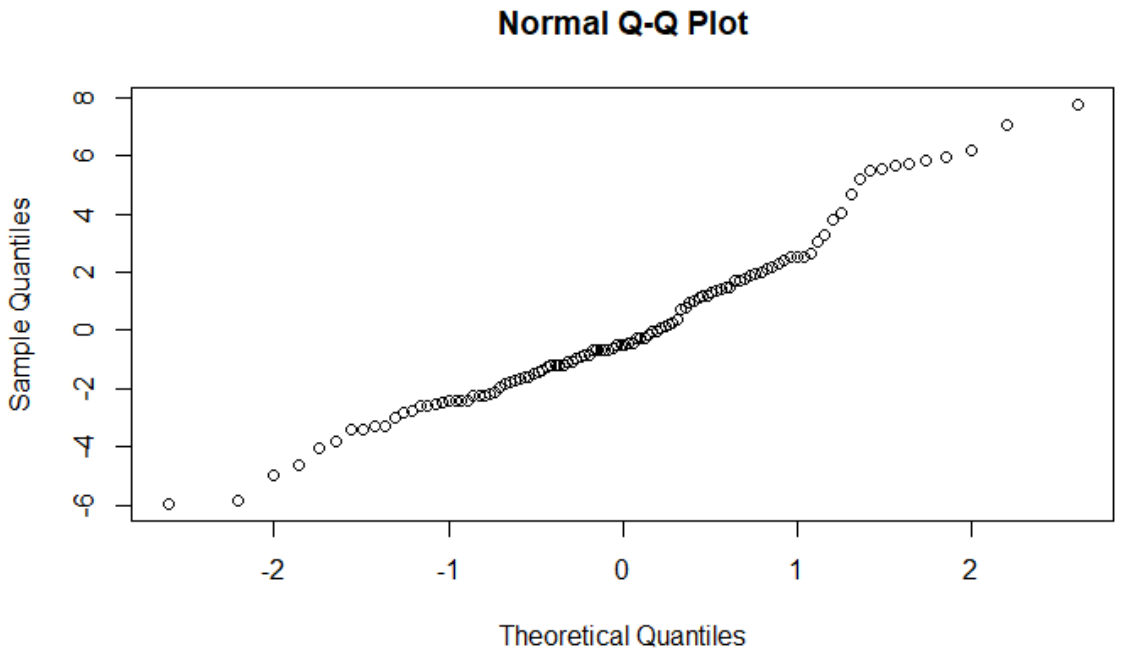
	AR1			Exchangeable			Independence		
Predictors	*Incidence Rate Ratios*	*CI*	*p*	*Incidence Rate Ratios*	*CI*	*p*	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	1.45	0.86 – 2.46	0.163	1.39	0.83 – 2.34	0.212	1.44	0.85 – 2.43	0.172
LA	1.76	1.42 – 2.19	**<0.001**	1.79	1.47 – 2.20	**<0.001**	1.78	1.44 – 2.20	**<0.001**
fSptreat [rlfld]	0.73	0.53 – 1.01	0.059	0.75	0.54 – 1.03	0.077	0.74	0.54 – 1.02	0.065
DEPTH	1.01	1.00 – 1.02	0.123	1.01	1.00 – 1.02	0.129	1.01	1.00 – 1.02	0.171
N	11 $_{fField}$			11 $_{fField}$			11 $_{fField}$		
Observations	110			110			110		

GEE - Bird Richness

- Model diagnostics are difficult for GEE!

```
fitted_values <- predict(M.gee1, type = "response")
residuals <- residuals(M.gee1, type = "pearson")
plot(fitted_values, residuals,
     main = "Pearson Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Pearson Residuals",
     pch = 16,
     col = "blue")
qqnorm(residuals)
```





Advantages

1. Computational simple
2. NO distributional assumptions
3. Have an inherent over-dispersion term
4. Robust to model misspecification
5. Provides population-averaged estimates

Disadvantages

1. Sensitive to data with only a few clusters (best with many clusters and few observations per cluster)
2. Limited to single-level clustering
3. Sensitive to Missing data (if not random)
4. Limited Diagnostic tools
5. No AIC based model selection

GEEs key-points

Interpretation: provides population-averaged estimates of the effects. Coefficients can be interpreted similarly to GLMs, adjusted for correlation.

Choosing Correlation Structures: accounts for repeated measures of clustered data.

Robust Standard Errors: GEEs use “sandwich” estimators for robust standard errors, which can be especially useful when the working correlation structure is misspecified.

Generalized linear mixed models (GLMM)

What are GLMMs?

- GLMMs extend the GLM model to allow for correlation between the observations, and nested data structures

- Good news: similar steps as to linear mixed model
- Bad news: parameter estimation and model selection is not straightforward – keep it simple

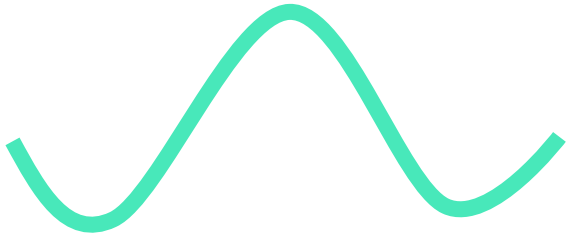
GLMMs are more complicated than LMM

LMM:

- everything is **normal + linear**
- we can compute the overall ("marginal") mean *exactly* by integrating over random effects

GLMM:

- response is **non-normal**
- link function is **nonlinear**
- random effects act on the **link scale**, not the response scale



The key issue:

- To get population-level results, we must **integrate over all possible random-effect values**.
- In LMMs this averaging is easy.
- In GLMMs it **has no exact solution**.

GLMM: parameter estimation

- Because the likelihood cannot be solved exactly,we use approximation methods:
  - Laplace approximation provides true likelihoods
  - Penalized quasi-likelihood estimation is commonly used but does not provide true likelihoods, so AIC-type measures are not applicable
  - Simulation-based techniques
  - Add priors and use Bayesian techniques

GLMM vs. LMM

Factor	Linear Mixed Model (LMM)	Generalized Linear Mixed Model (GLMM)
Response Variable	Continuous and normally distributed	Non-normal distributions (e.g., binary, count, proportions)
Distribution Assumptions	Assumes normality for both the response and residuals	Allows various distributions for response variable (e.g., binomial, Poisson)
Link Function	Uses an identity link (linear relationship)	Requires a link function (e.g., logit, log) to relate predictors to response variable
* **Random Effects**	Assumed to be normally distributed, relatively straightforward to model	Also normally distributed, but interact with link functions and non-normal response, adding complexity
* **Likelihood Calculation**	Direct, closed-form likelihood	Involves integration over random effects, often with no closed form, requiring approximate or numerical integration
* **Parameter Estimation**	Can be done using traditional methods like REML	Requires specialized estimation techniques (e.g., penalized quasi-likelihood, laplace, adaptive quadrature, MCMC in Bayesian GLMM)
Computational Complexity	Lower computational demand, faster to fit	High computational demand, may require advanced software or resources
Interpretation of Effects	Relatively straightforward to interpret	More complex, particularly with non-linear link functions
Diagnostics	Well-established diagnostic tools available	Fewer diagnostic tools, harder to assess model fit

Which R packages (functions) fit GLMMs?

- MASS::glmmPQL (penalized quasi-likelihood)
- lme4::glmer (Laplace approximation and adaptive Gauss-Hermite quadrature [AGHQ])
- MCMCglmm (Markov chain Monte Carlo)
- glmmML (AGHQ)
- glmmAK (AGHQ?)
- glmmADMB (Laplace)
- glmm (from Jim Lindsey's repeated package: AGHQ)
- glmmTMB (Laplace, AGHQ)

Which R packages (functions) fit GLMMs?

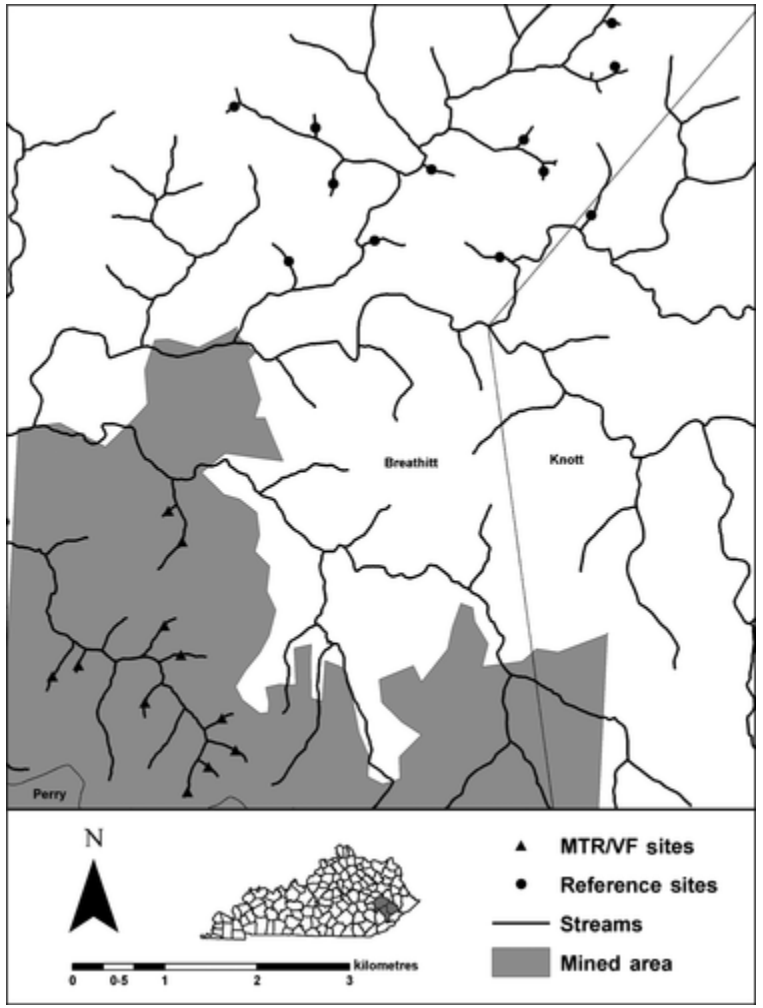
Package	Function	Zero inflation	Over-disperison	Temporal/ Spatial	Multiple random effects	Crossed random effects
lme4	glmer	No	Limited	Limited	Yes	Limited
glmmTMB	glmmTMB	Yes	Yes	Yes	Yes	Yes
glmmADMB	glmmADMB	No	Yes	Yes	Yes	Yes
MASS	glmPQL	No	Limited	No	No	No

GLMM: Salamanders Study

- Mountaintop removal mining and valley filling (MTR/VF) is one form of land use which can be a stressor to stream ecosystems.
- Salamanders are sensitive to land use changes that impact streams and their catchment
- In this study, the effects of MTR/VF were evaluated on stream salamanders in eastern Kentucky (Price et al. 2016).

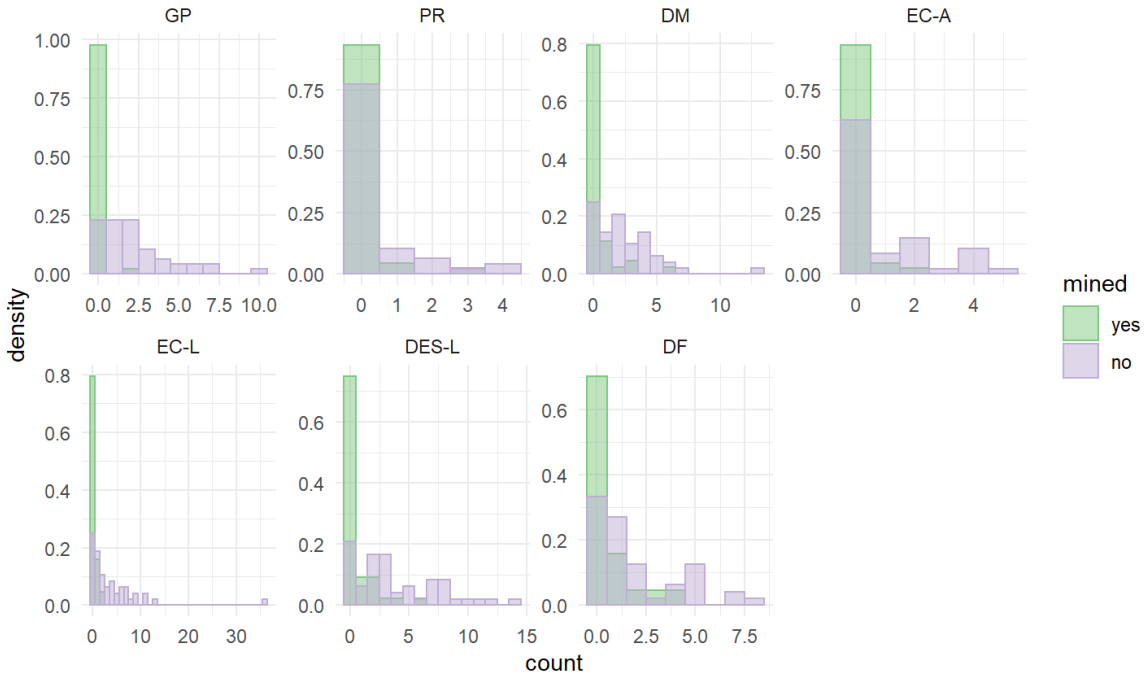
Study Design

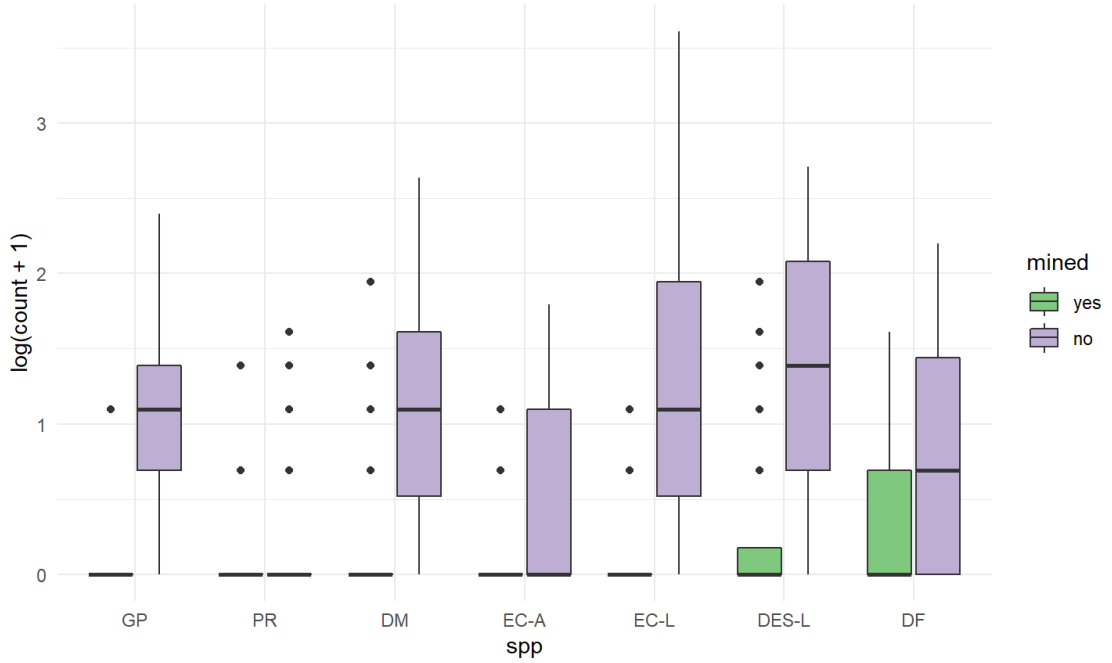
- Counts of salamanders were recorded at 23 headwater streams (sites); 11 streams located on mined land, and 12 reference streams
- In each stream, a single 10-m sampling transect was set up and was sampled 4 times (usually monthly) from March through June 2013
- Salamanders were identified to species and in some instances life stage (i.e. adult vs. larva).
- As multiple samples were taken from each site, at different times, we have pseudo-replication. We need a random effect for site to account for this.



Plot the data

How does mining impact the abundance of Salamanders?
Are these responses consistent across species/life stage?





What model should we use?
Fixed? Random? Distribution?

Fit a model

```
poisfit = glmmTMB(count~  spp + mined + spp:mined + (1|site),
                  salamanders, family="poisson")
```

- In this model we have:
  - count as the response
  - fixed effects: species (spp), mined and their interaction
  - random effect: random intercept for each site
  - poisson distribution

Examine the output

```
summary(poisfit)
tab_model(poisfit)
```

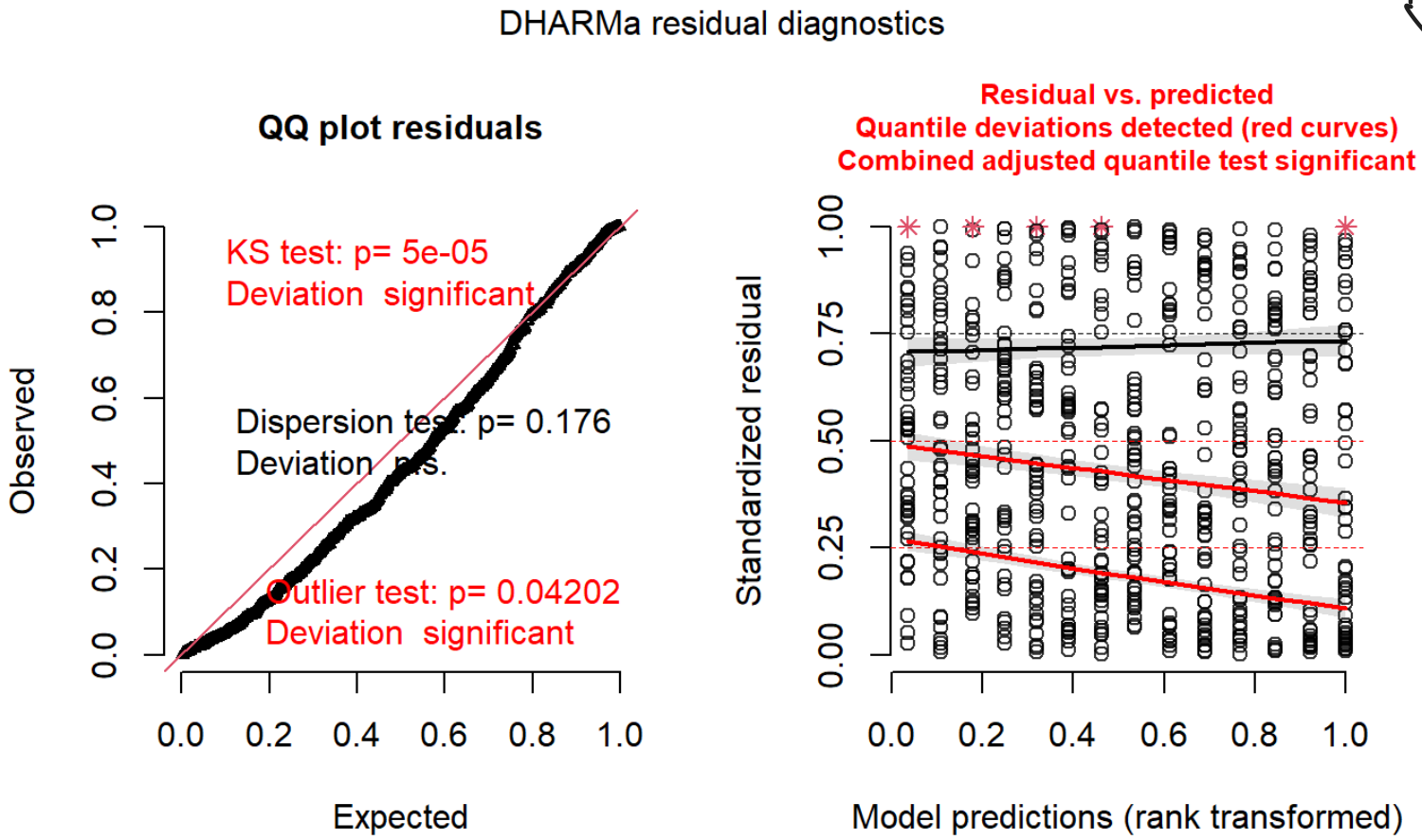
```
 Family: poisson  ( log )
Formula:          count ~ spp + mined + spp:mined + (1 | site)
Data: Salamanders

     AIC      BIC   logLik deviance df.resid
  1940.2   2007.3   -955.1   1910.2      629

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 site   (Intercept) 0.3316   0.5759
Number of obs: 644, groups:  site, 23

Conditional model:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.3771     0.7340  -4.601 4.20e-06 ***
sppPR          0.9163     0.8367   1.095 0.273435
sppDM          2.2513     0.7434   3.028 0.002458 **
sppEC-A        0.6931     0.8660   0.800 0.423494
sppEC-L        1.7047     0.7687   2.218 0.026576 *
sppDES-L       2.5257     0.7348   3.437 0.000588 ***
sppDF          2.5257     0.7348   3.437 0.000588 ***
minedno        4.1109     0.7587   5.418 6.02e-08 ***
sppPR:minedno -2.4887     0.8688  -2.864 0.004178 **
sppDM:minedno -2.1526     0.7554  -2.850 0.004377 **
sppEC-A:minedno -1.5279   0.8838  -1.729 0.083853 .
sppEC-L:minedno -1.1212   0.7782  -1.441 0.149670
sppDES-L:minedno -1.9527  0.7448  -2.622 0.008748 **
sppDF:minedno  -2.6674    0.7485  -3.563 0.000366 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

	count		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	0.03	0.01 – 0.14	**<0.001**
spp [PR]	2.50	0.49 – 12.89	0.273
spp [DM]	9.50	2.21 – 40.78	**0.002**
spp [EC-A]	2.00	0.37 – 10.92	0.423
spp [EC-L]	5.50	1.22 – 24.81	**0.027**
spp [DES-L]	12.50	2.96 – 52.77	**0.001**
spp [DF]	12.50	2.96 – 52.77	**0.001**
mined [no]	61.00	13.79 – 269.89	**<0.001**
spp [PR] × mined [no]	0.08	0.02 – 0.46	**0.004**
spp [DM] × mined [no]	0.12	0.03 – 0.51	**0.004**
spp [EC-A] × mined [no]	0.22	0.04 – 1.23	0.084
spp [EC-L] × mined [no]	0.33	0.07 – 1.50	0.150
spp [DES-L] × mined [no]	0.14	0.03 – 0.61	**0.009**
spp [DF] × mined [no]	0.07	0.02 – 0.30	**<0.001**
Random Effects			
σ^2	0.90		
$\tau_{00\ site}$	0.33		
ICC	0.27		
N_{site}	23		
Observations	644		
Marginal R^2 / Conditional R^2	0.633 / 0.732		

Model Fit

If Overdispersion exists in the model, try negative Binomial.

```
library(DHARMa)
simulateResiduals(fittedModel = poisfit, plot = T)
```



Dispersion Models

- Let's try a dispersion model. We will fit a negative binomial model (nbinom2), which assumes the variance increases quadratically with the mean.

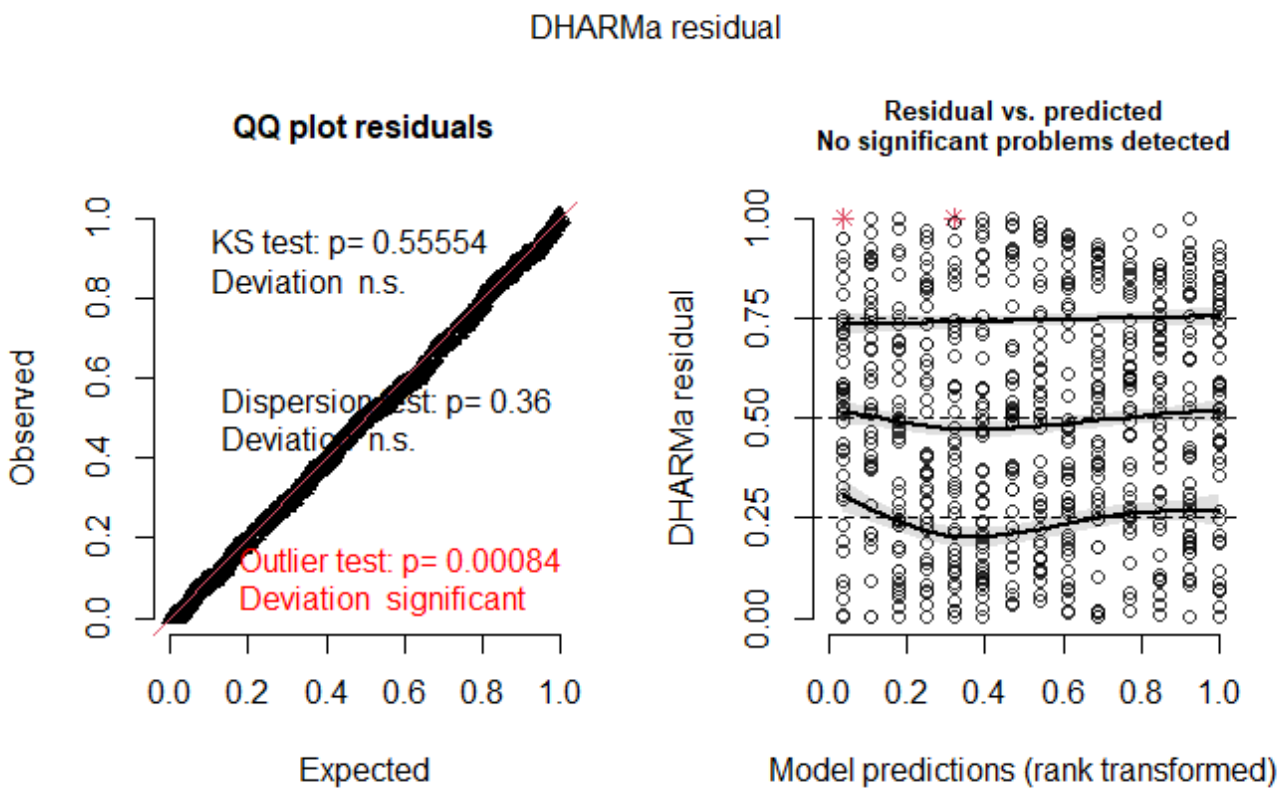
```
nbfit1 = glmmTMB(count~  spp + mined + spp:mined + (1|site),
                 salamanders, family="nbinom2")
```

- If counts were expected to become more dispersed (more variable relative to the mean) over the four months, then this can be included to model how the dispersion changes with the day of the year (DOY):

```
nbdisp = glmmTMB(count~  spp + mined + spp:mined + (1|site),
                 dispformula = ~ DOY,
                 salamanders, family="nbinom2")
```

Negative Binomial

```
tab_model(nbfit1)
simulateResiduals(nbfit1, plot=T)
```



	count		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	0.03	0.01 – 0.15	**<0.001**
spp [PR]	2.54	0.45 – 14.16	0.289
spp [DM]	9.47	2.02 – 44.37	**0.004**
spp [EC-A]	2.04	0.35 – 12.04	0.430
spp [EC-L]	6.13	1.25 – 30.15	**0.026**
spp [DES-L]	12.32	2.67 – 56.76	**0.001**
spp [DF]	13.15	2.85 – 60.66	**0.001**
mined [no]	64.20	13.56 – 303.87	**<0.001**
spp [PR] × mined [no]	0.08	0.01 – 0.47	**0.006**
spp [DM] × mined [no]	0.12	0.02 – 0.59	**0.009**
spp [EC-A] × mined [no]	0.21	0.03 – 1.31	0.094
spp [EC-L] × mined [no]	0.26	0.05 – 1.39	0.115
spp [DES-L] × mined [no]	0.14	0.03 – 0.71	**0.018**
spp [DF] × mined [no]	0.06	0.01 – 0.32	**0.001**
Random Effects			
σ^2	1.23		
$\tau_{00\ site}$	0.28		
ICC	0.19		
$N_{\ site}$	23		
Observations	644		
Marginal R^2 / Conditional R^2	0.584 / 0.662		

Model selection

```
nbfit1 = glmmTMB(count~  spp + mined + spp:mined + (1|site),
                 Salamanders, family="nbinom2")
```

```
drop1(nbfit1)
```

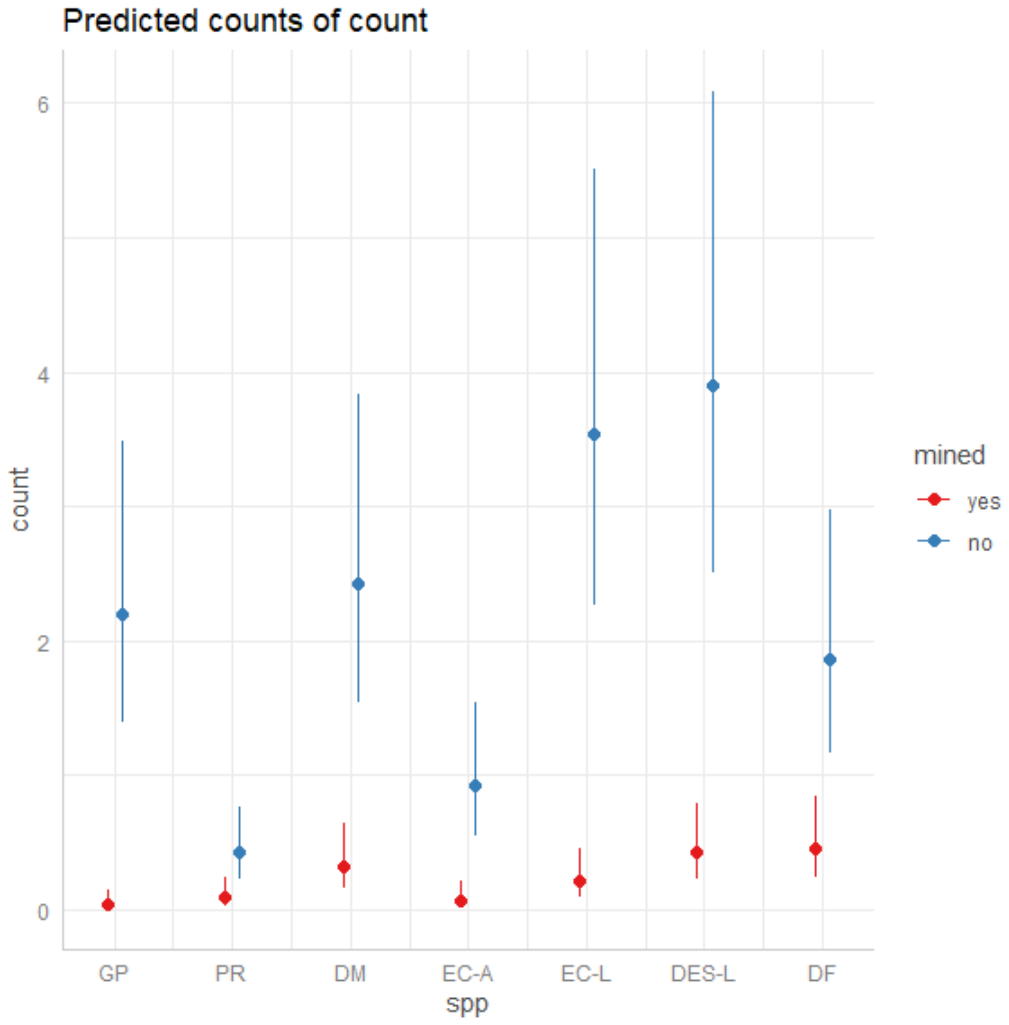
```
> drop1(nbfit1)
Single term deletions

Model:
count ~ spp + mined + spp:mined + (1 | site)
           Df    AIC
<none>        1663.4
spp:mined   6 1672.4
```

```
nbfit.noint = glmmTMB(count~  spp + mined +  (1|site),
                      Salamanders, family="nbinom2")
lrtest(nbfit1, nbfit.noint)
```

```
> lrtest(nbfit1, nbfit.noint)
Likelihood ratio test

Model 1: count ~ spp + mined + spp:mined + (1 | site)
Model 2: count ~ spp + mined + (1 | site)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  16 -815.68
2  10 -826.20 -6 21.047   0.001799 **
---
```

Remember: Likelihood estimation and associated tests are approximate!

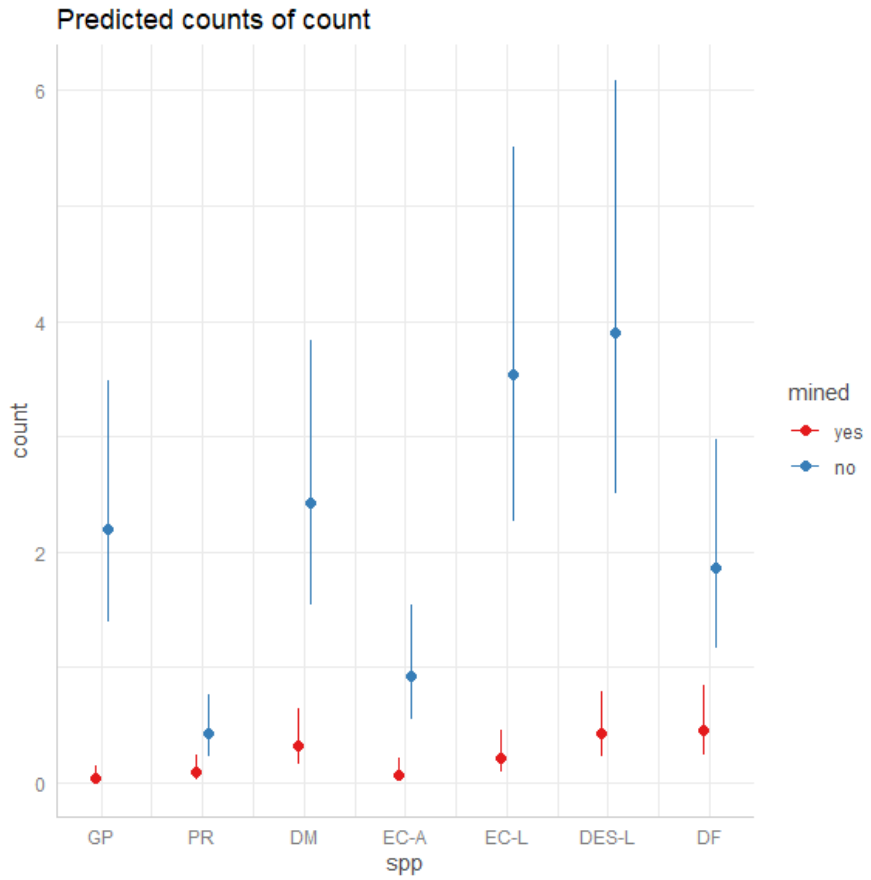
Model Interpretation



	count		
Predictors	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	0.03	0.01 – 0.15	**<0.001**
spp [PR]	2.54	0.45 – 14.16	0.289
spp [DM]	9.47	2.02 – 44.37	**0.004**
spp [EC-A]	2.04	0.35 – 12.04	0.430
spp [EC-L]	6.13	1.25 – 30.15	**0.026**
spp [DES-L]	12.32	2.67 – 56.76	**0.001**
spp [DF]	13.15	2.85 – 60.66	**0.001**
mined [no]	64.20	13.56 – 303.87	**<0.001**
spp [PR] × mined [no]	0.08	0.01 – 0.47	**0.006**
spp [DM] × mined [no]	0.12	0.02 – 0.59	**0.009**
spp [EC-A] × mined [no]	0.21	0.03 – 1.31	0.094
spp [EC-L] × mined [no]	0.26	0.05 – 1.39	0.115
spp [DES-L] × mined [no]	0.14	0.03 – 0.71	**0.018**
spp [DF] × mined [no]	0.06	0.01 – 0.32	**0.001**
Random Effects			
σ^2	1.23		
$\tau_{00\ site}$	0.28		
ICC	0.19		
$N_{\ site}$	23		
Observations	644		
Marginal R^2 / Conditional R^2	0.584 / 0.662		

Model Interpretation – posthoc comparison

```
library(emmeans)
emmeans(nbfit1, pairwise ~ mined|spp)
```




```
Confidence level used: 0.95
Intervals are back-transformed from the log scale

$contrasts
spp = GP:
 contrast  ratio     SE  df null z.ratio p.value
 yes / no 0.0156 0.0124 Inf    1  -5.247  <.0001

spp = PR:
 contrast  ratio     SE  df null z.ratio p.value
 yes / no 0.2062 0.1239 Inf    1  -2.627  0.0086

spp = DM:
 contrast  ratio     SE  df null z.ratio p.value
 yes / no 0.1337 0.0556 Inf    1  -4.836  <.0001

spp = EC-A:
 contrast  ratio     SE  df null z.ratio p.value
 yes / no 0.0758 0.0472 Inf    1  -4.141  <.0001

spp = EC-L:
 contrast  ratio     SE  df null z.ratio p.value
 yes / no 0.0594 0.0269 Inf    1  -6.238  <.0001

spp = DES-L:
 contrast  ratio     SE  df null z.ratio p.value
 yes / no 0.1079 0.0427 Inf    1  -5.623  <.0001

spp = DF:
 contrast  ratio     SE  df null z.ratio p.value
 yes / no 0.2419 0.0966 Inf    1  -3.553  0.0004

Tests are performed on the log scale
```

GLMM key-points

Handles non-normally distributed data, such as counts, proportions, or binary outcomes

Provide a flexible framework for analyzing complex data structures by accommodating both fixed and random effects

Likelihood estimation is approximate and complex models often fail to converge

GEE is the average effect
GLMM is the effect found in the average group

Aspect	GEE	GLMM
Interpretation	Population-level estimates	Individual/group-level estimates
Computation	Simpler, often faster	More complex, often slower
Correlation Structure	Working correlation structure (not complex)	Random effects structure (multi-level); assumes normally distributed
Parameter estimation method	Quasi-likelihood- solves estimating equations derived from likelihood	Likelihood to estimate both fixed and random effects
Sample size	Requires larger sample size for robust estimates	Can be effective with smaller samples
Variance structure	Assumes that the mean and variance are related through a specified form, but the variance structure can be misspecified without affecting robustness.	Requires a specific variance structure linked to the distribution of the response variable; misspecification can lead to biased results.
Model selection/significance	Wald Z, QIC, robust standard errors	AIC, BIC, likelihood ratio tests
Model diagnostics	Not clear	DHARMa
Typical Uses	Longitudinal data, epidemiology	Ecology

GEE and GLMM summary

- **GEE** and **GLMM** provide valuable frameworks for analyzing non-normal datasets with correlated observations
- **GEEs** provide **population-averaged** estimates and are effective for **correlated data**, but they do not model hierarchical structures and are **sensitive to missing or unbalanced data**, especially with small sample size
- **GLMMs** allow for **individual- or site-specific** estimates through **random effects**, effectively handle **hierarchical and nested data** structures, and are **more robust to missing or unbalanced data**, including small sample sizes.