Model checking and comparison

Bert van der Veen

Department of Mathematical Sciences, NTNU

Questions so far?

Outline

- Defining a residual
- Residual plots for checking VGLM(M) assumptions
- Checking (prior) random effect assumptions
- Model-comparison

Modeling checking

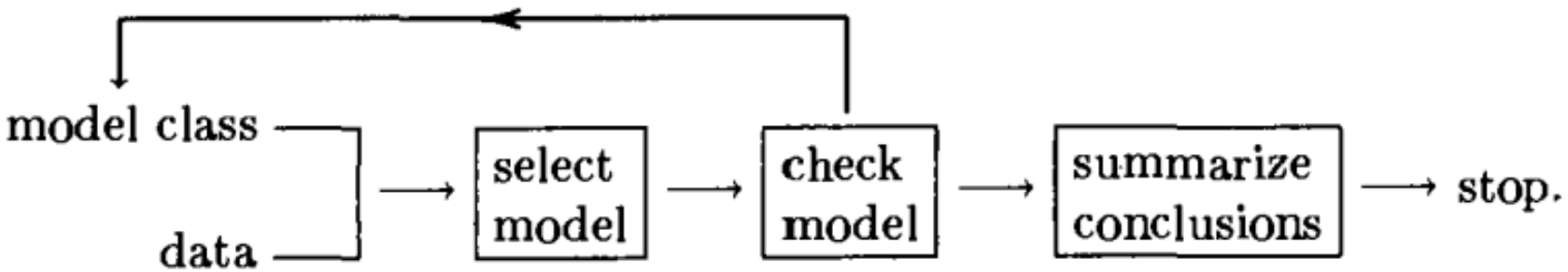


Figure 1: McCullagh and Nelder (1989) workflow

Assumptions

We have made various assumptions in the models we fitted so far:

1) Poisson/negative-binomial/binomial distributed responses (we could change to negative-binomial)
2) Correct link function, correct variance function
3) Same dispersion for species (move to vector GLM)
4) We have correctly specified the model structure (species-specific vs. common effects, autocorrelation)
5) Species responses are independent (move to JSDM, tomorrow)
6) No outliers

Mean-variance

Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8**, 1408–1414 doi: 10.1111/2041-210X.12843

FORUM

The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017)

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

The mean-variance relationship has been one of the main drivers of model-based multivariate methods.

Checking assumptions

Every method/model comes with assumptions. We need to check those to make sure our results are valid.

How do we check those?

The linear model

Writing the (V)linear model:

$$y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

This error is where we place our assumptions

The linear model

Writing the (V)linear model:

$$y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

This error is where we place our assumptions

We can use the the estimate of the error (residual) for assumption checking. This holds all information that is not included in our model.

The linear model

Writing the (V)linear model:

$$y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

This error is where we place our assumptions

We can use the the estimate of the error (residual) for assumption checking. This holds all information that is not included in our model.

Violated residual assumptions mean that some or all of your model's results are untrustworthy.

(V)Generalised linear model

$$
\begin{aligned}
g\{\mathbb{E}(y_i|x_i)\} &= \eta_i = \alpha + x_i\beta \\
\mathbb{E}(y_i|x_i) &= g^{-1}(\eta_i) = g^{-1}(\alpha + x_i\beta)
\end{aligned}
\tag{2}
$$

VGLMs do not have an error term, checking is harder

But we still want to check in the same way

Response residuals

We could use the same residual as in linear regression:

$$\epsilon_i = y_i - \hat{\mu}_i$$

But we do not expect these to look nice in (V)GLMs.

- Because the variance depends on the mean
- We want nice looking residuals when a model is good
- We want bad looking residuals otherwise

(V)GLM residual

Even though we lack the ϵ_i term in (V)GLMs, we still calculate the residuals similarly

(V)GLM residual

Even though we lack the ϵ_i term in (V)GLMs, we still calculate the residuals similarly

$$\hat{\epsilon_{i,pearson}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathsf{var}(y_i; \hat{\mu}_i, \hat{\phi})}} \tag{3}$$

(V)GLM residual

Even though we lack the ϵ_i term in (V)GLMs, we still calculate the residuals similarly

$$\hat{\epsilon_{i,pearson}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(y_i; \hat{\mu}_i, \hat{\phi})}} \tag{3}$$

$$\hat{\epsilon_{i,deviance}} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \qquad \text{so that } \sum_{i=1}^{n} \hat{\epsilon_{deviance,i}}^2 \tag{4}$$

(V)GLM residual

Even though we lack the ϵ_i term in (V)GLMs, we still calculate the residuals similarly

$$\hat{\epsilon_{i,pearson}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(y_i; \hat{\mu}_i, \hat{\phi})}} \tag{3}$$

$$\hat{\epsilon_{i,deviance}} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \qquad \text{so that } \sum_{i=1}^{n} \hat{\epsilon_{deviance,i}}^2 \tag{4}$$

Approximately normally distributed in large samples

(V)GLM residual

Even though we lack the ϵ_i term in (V)GLMs, we still calculate the residuals similarly

$$\hat{\epsilon_{i,pearson}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(y_i; \hat{\mu}_i, \hat{\phi})}} \tag{3}$$

$$\hat{\epsilon_{i,deviance}} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \qquad \text{so that } \sum_{i=1}^{n} \hat{\epsilon_{deviance,i}}^2 \tag{4}$$

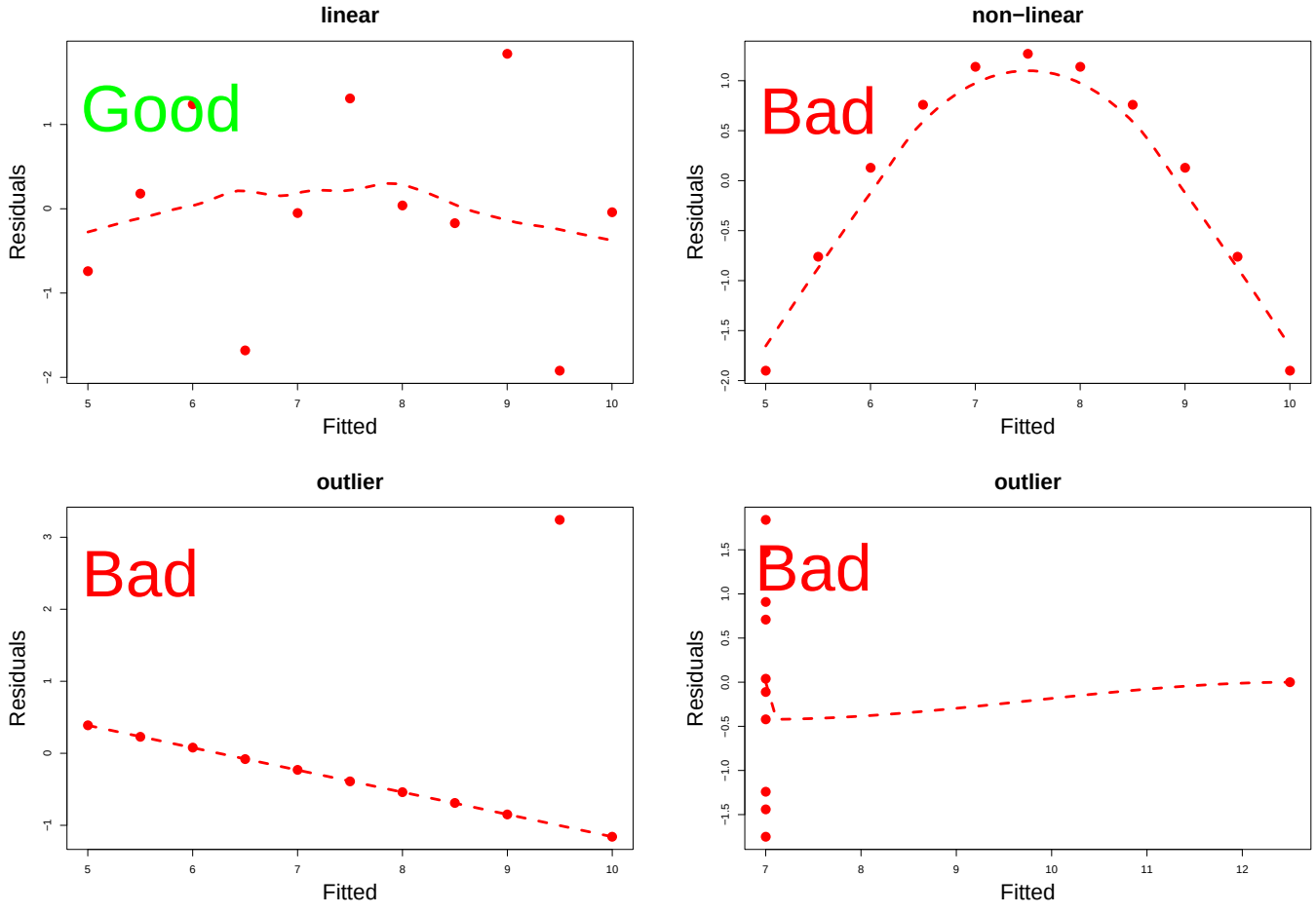
Approximately normally distributed in large samples

Still inappropriate for discrete data and small samples

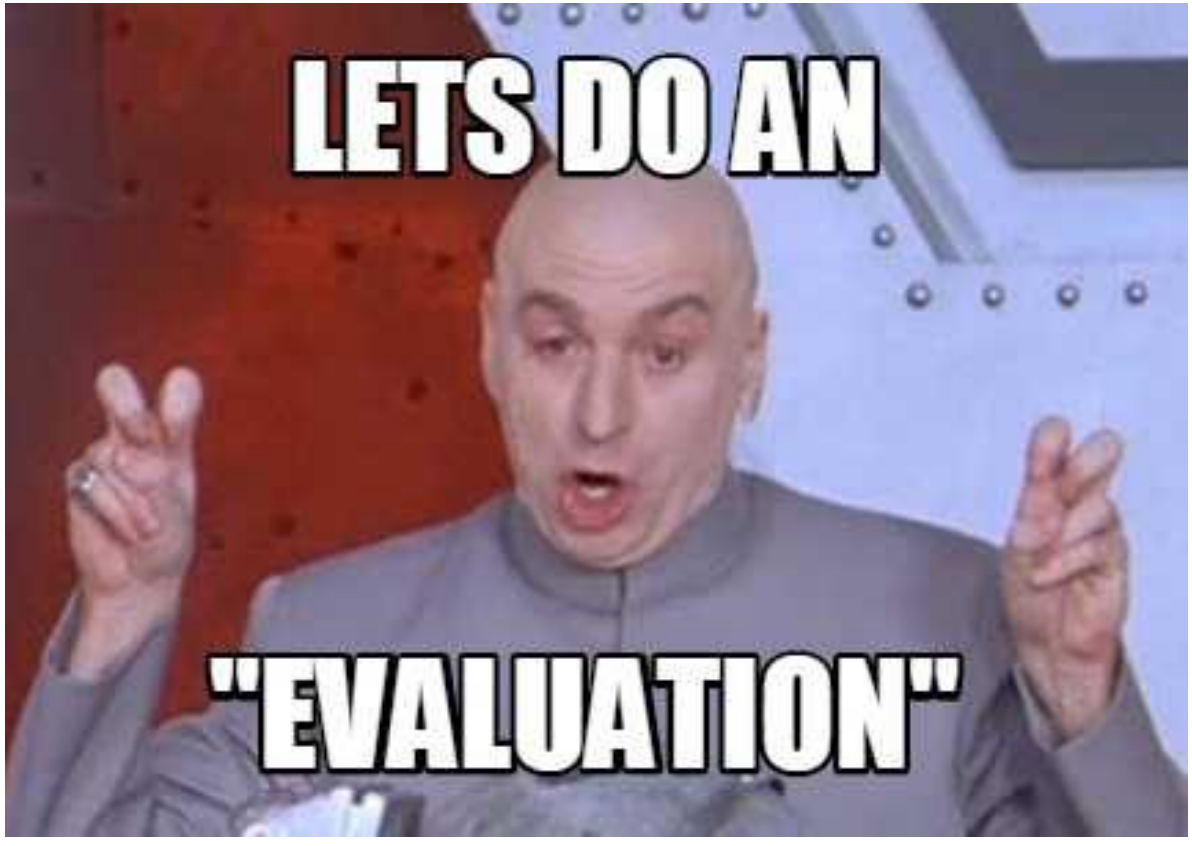
In practice

In practice, both Pearson's and Deviance residuals are often non-normally distributed.

Residual plots: residuals vs. fitted



Example 1

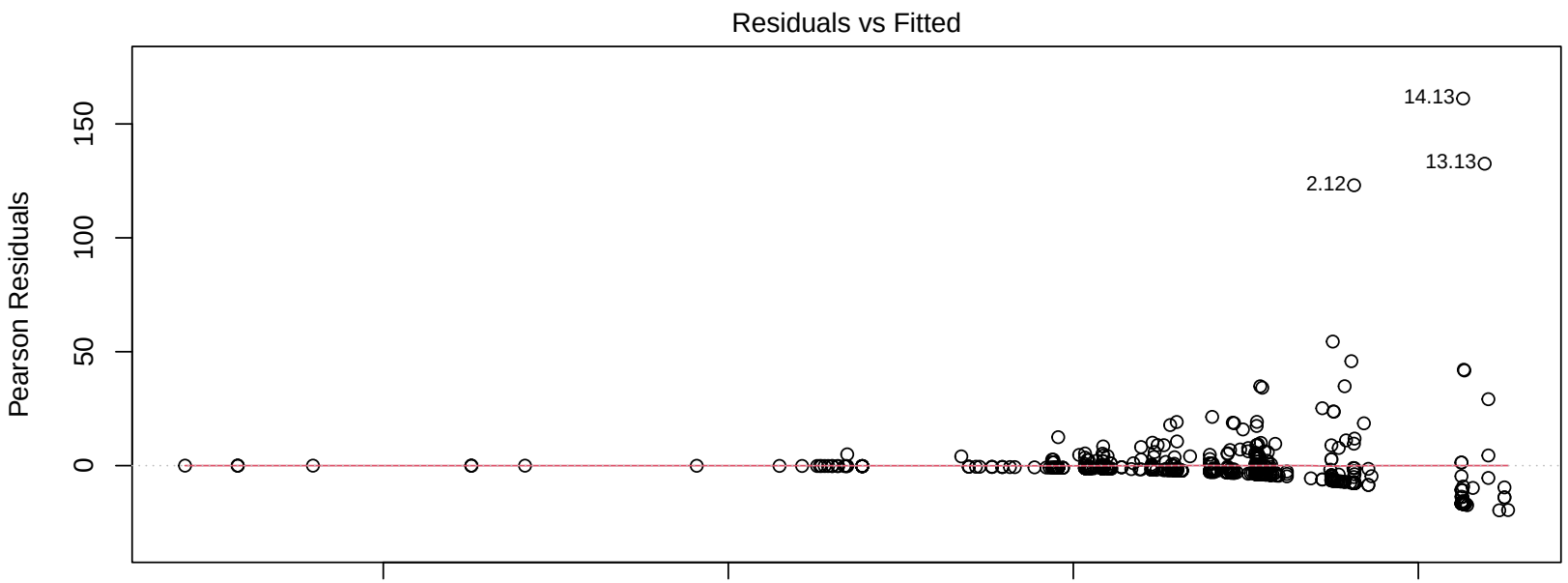


Wetlands: residual diagnostics

We had this model:

```
model1 <- glm(Count~0+Species+NO3+NO3:Species, data = long, family = "pois
```

With the following residual plot:

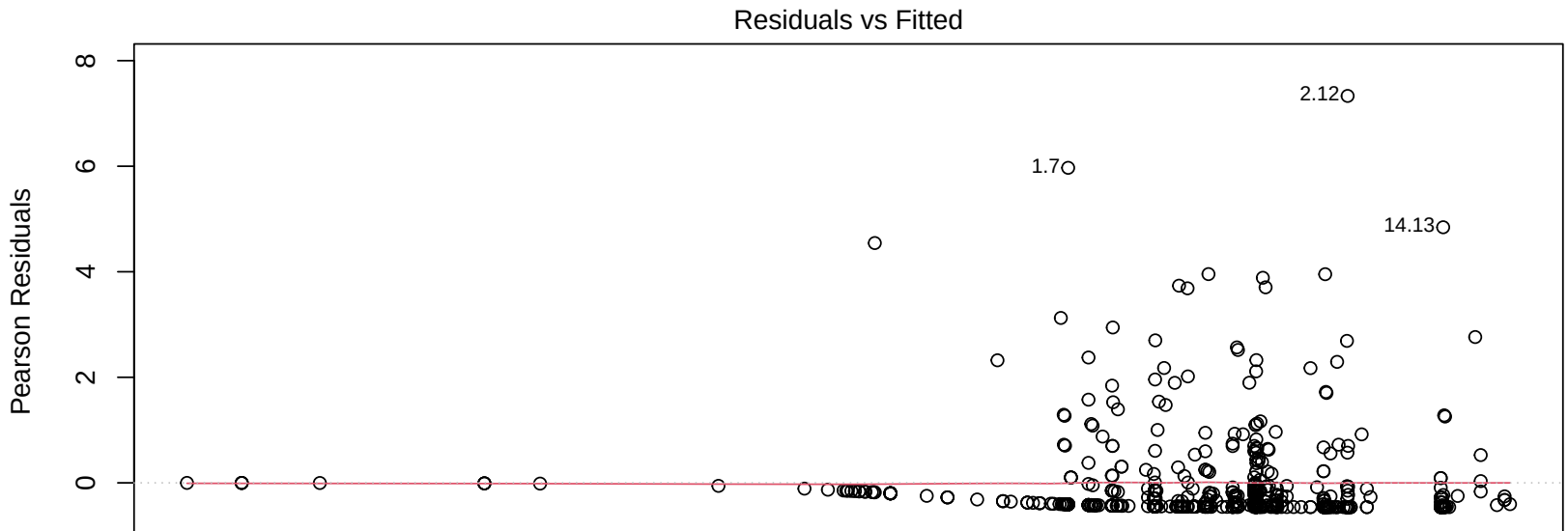


Wetlands: residual diagnostics

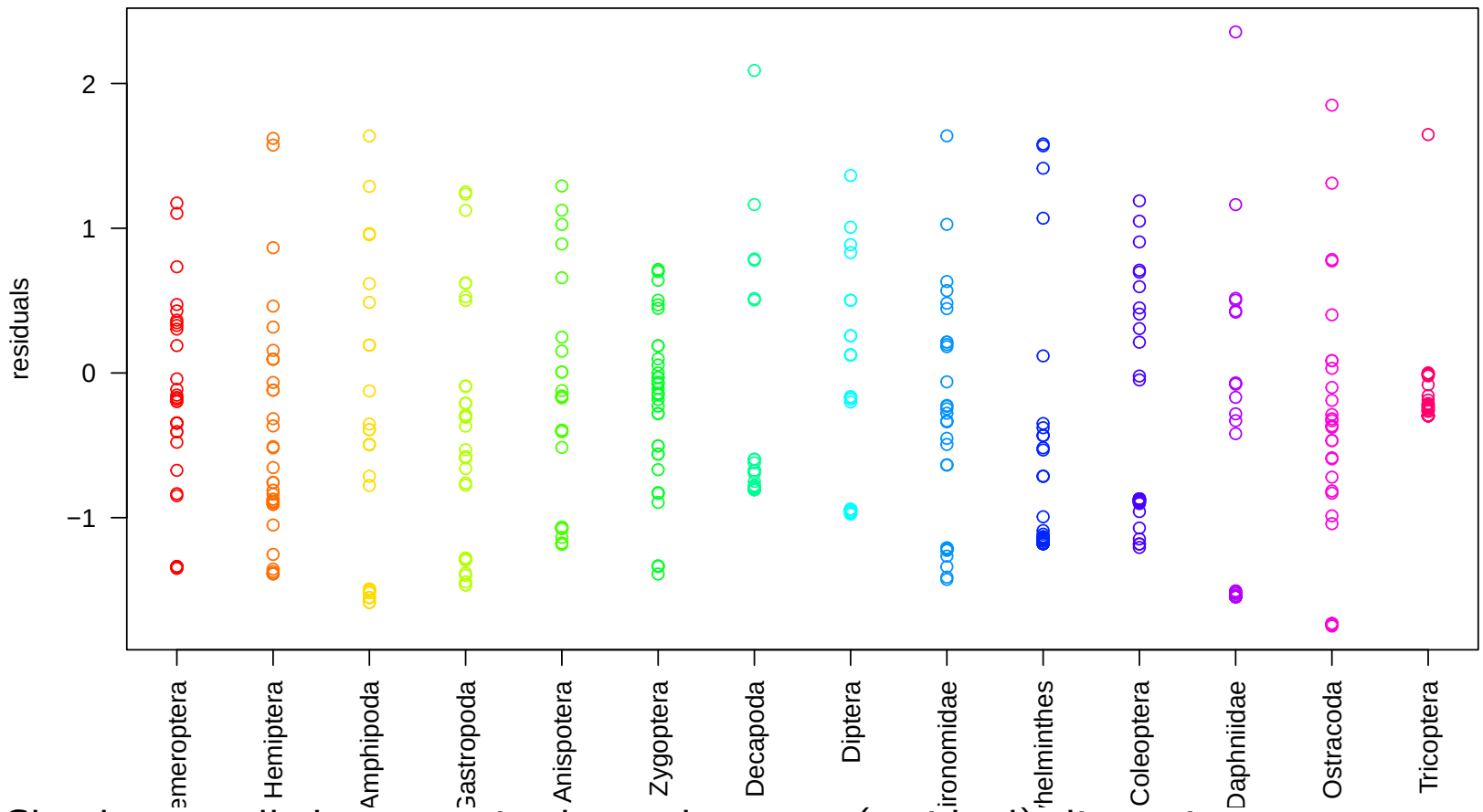
We also had this model:

```
model2 <- MASS::glm.nb(Count~0+Species + NO3:Species,
                       data = long)
```

With the following residual plot:



Wetlands: residual diagnostics (species)



Clearly, not all these species have the same (residual) dispersion.

(V)GLMM checking

Random effects

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e} \qquad (5)$$

Random effect

Random effects

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e} \qquad (5)$$

Random effect

Random effect estimates $\hat{\mathbf{u}}$ have a very relation to the random effect $\mathbf{u}$ as the residual $\hat{\epsilon}$ does for the error ϵ.

(V)GLMM residuals

Conditional

$$\mathrm{g}\{\mathbb{E}(y_{ij}|\mathbf{x}_i)\} = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_j + \mathbf{z}_i^\top \hat{\mathbf{u}}_j \tag{6}$$

Unconditional

$$\mathrm{g}\{\mathbb{E}(y_{ij}|\mathbf{x}_i)\} = \mathbf{x}_i \hat{\boldsymbol{\beta}}_j \tag{7}$$

How do we calculate the residual?

- should we condition on the estimate of the random effect?
- simulate from conditional distribution?

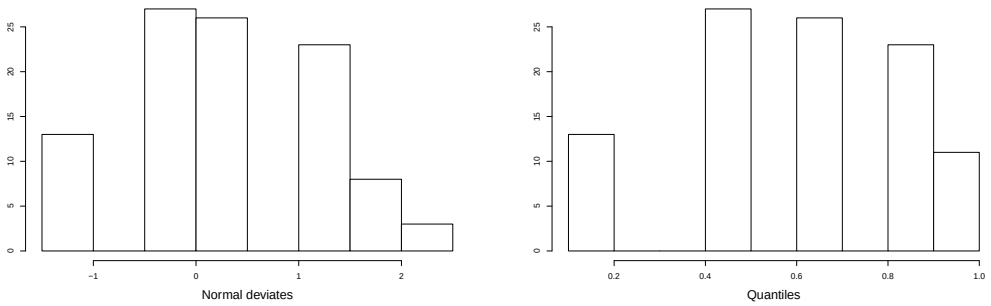
i.e. a range of options

Randomized Quantile residual (Dunn and Smyth 1996)

- Gold standard residual
- Better suited for small samples and discrete data types
- Exactly normally distributed
- Suitable for all kinds of models

Continuous

$$r_Q = \Phi^{-1}\left\{\mathcal{F}\left(y_i; \hat{\mu}_i, \hat{\phi}\right)\right\} \tag{8}$$



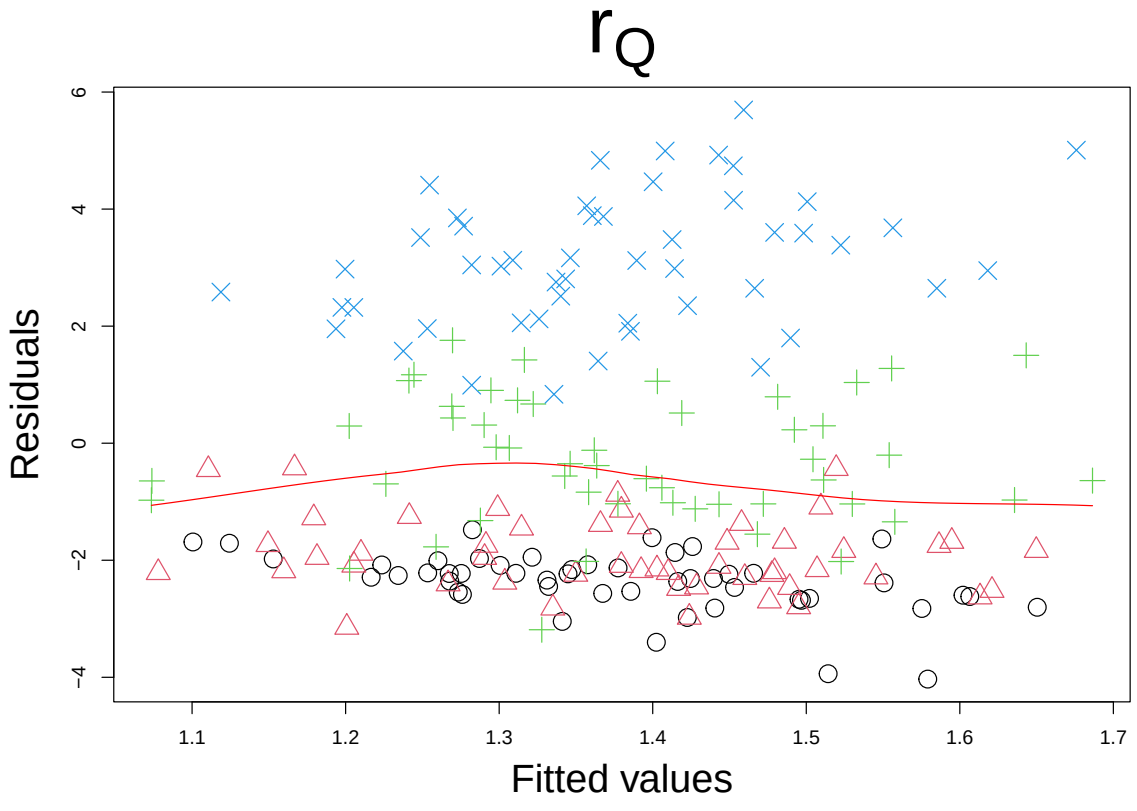
Packages that implement Quantile residuals

- You might have heard of `DHARMa` (Hartig 2024)
- `mvabund` (Wang et al. 2024)
- `gllvm` also implements (conditional) Quantile residuals

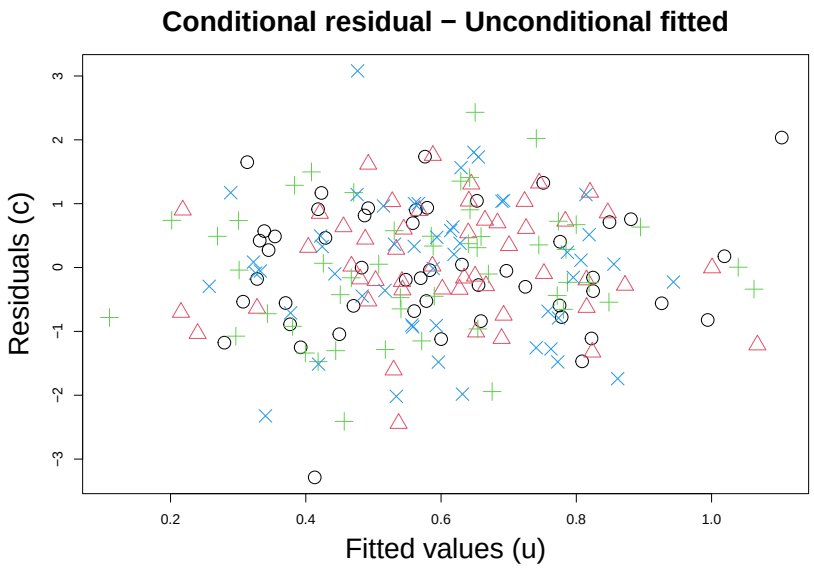
Simulation: grouping of errors

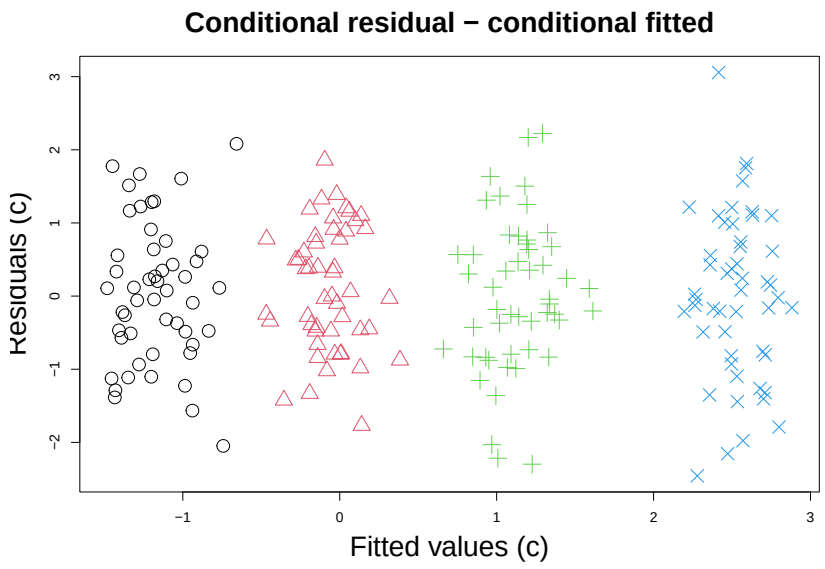
```
n <- 200
ngroups <- 4
alpha <- 0.5
beta <- -1
x <- rnorm(n, sd = 0.2)

fac<-rep(1:ngroups,each=n/ngroups)
e <- seq(from=-2,to=2,length.out=ngroups)
mu <- exp(alpha + beta*x + e[fac])
y <- rpois(n = n, lambda = mu)
```

Example: Poisson residuals (grouping)



Residual diagnostics: Poisson residuals (grouping)





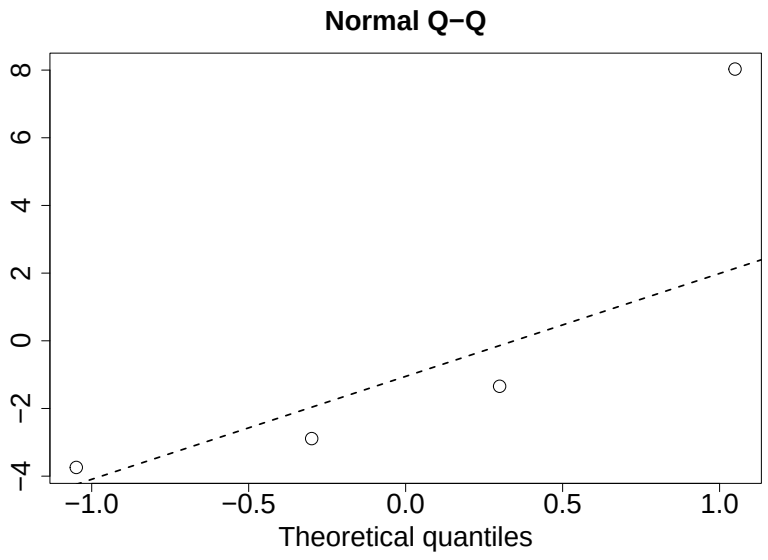
GLMM: checking random effect assumptions

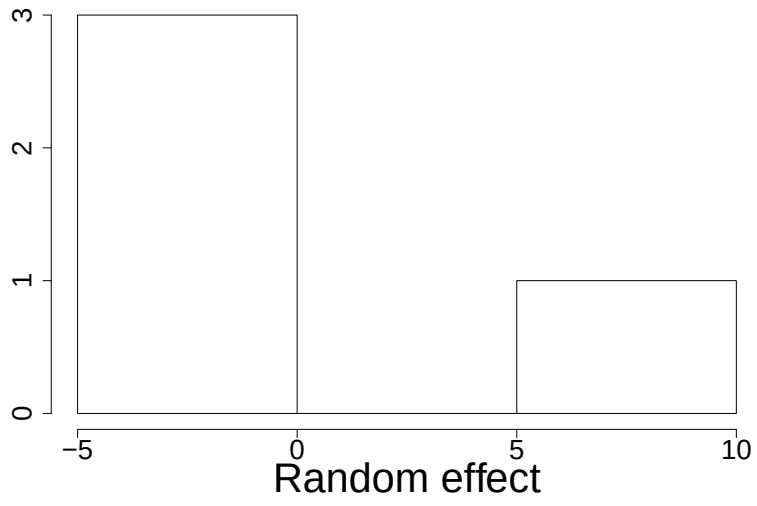
- random effect is a type of residual
- $\hat{u}_j$ is an estimate of the mean or mode of $p(u_j|y_i)$
- we treat $\hat{u}_j$ as a sample of the random effect distribution
- so we check assumptions (marginal normality, constant variance, independence, no outliers)!
- difficult with small number of groups
- needs to be done for every random effect

Simulation: GLMM (outlier)

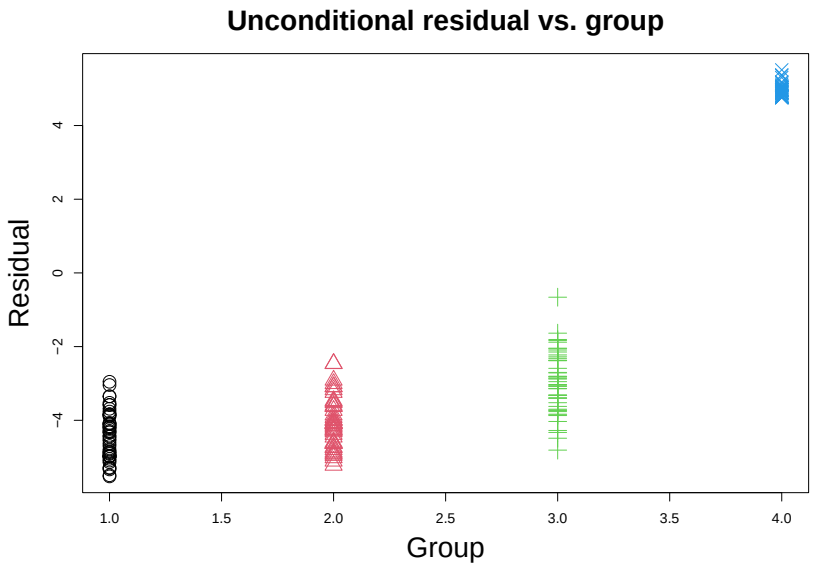
```
n <- 200
ngroups <- 4
alpha <- 0.5
beta <- -1
x <- rnorm(n, sd = 0.2)

fac<-rep(1:ngroups,each=n/ngroups)
e <- seq(from=-2,to=2,length.out=ngroups)
e[4] <- 10
mu <- exp(alpha + beta*x + e[fac])
y <- rpois(n = n, lambda = mu)
```

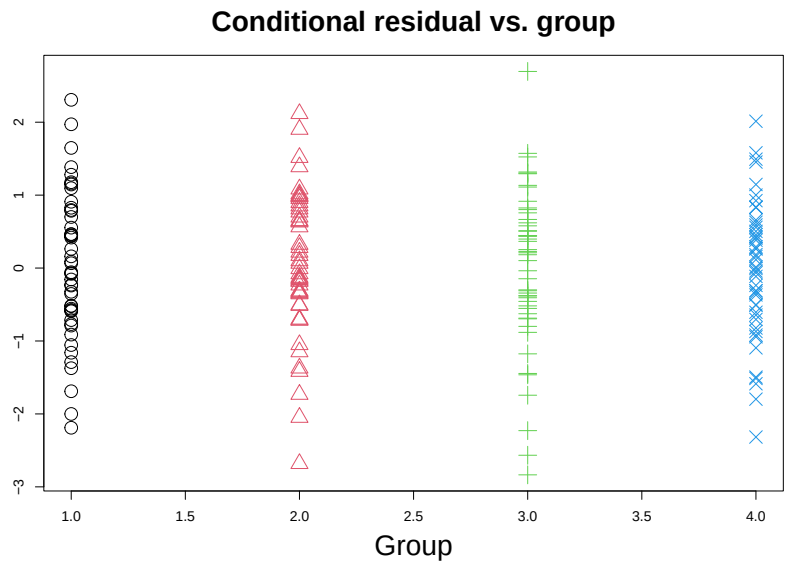
GLMM diagnostics





GLMM diagnostics

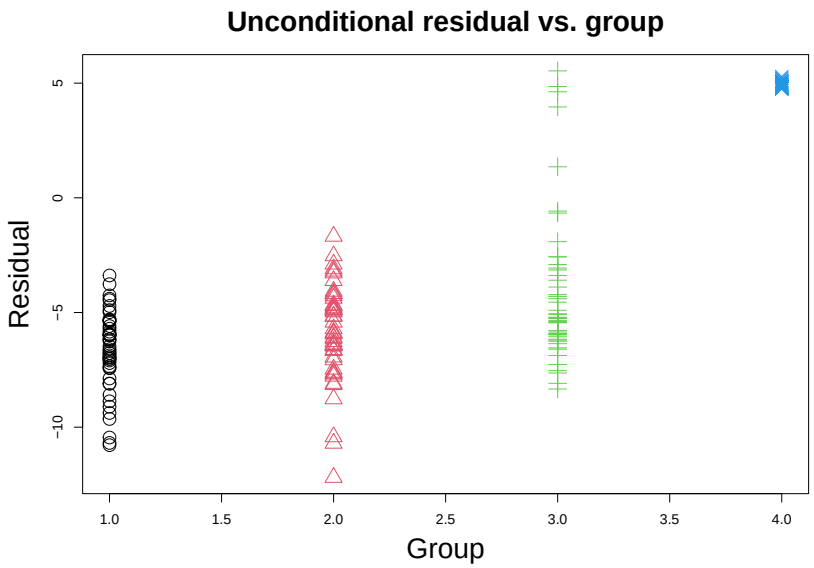


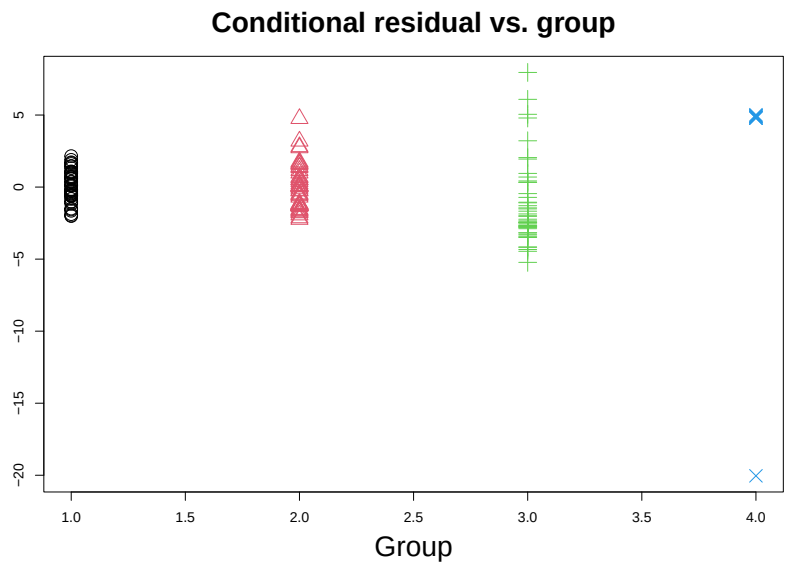


What if constant variance is violated

```
n <- 200
ngroups <- 4
alpha <- 0.5
beta <- -1
x <- rnorm(n, sd = 0.2)

fac<-rep(1:ngroups,each=n/ngroups)
e <- seq(from=-2,to=2,length.out=ngroups)
e[4] <- 10
e2 <- MASS::mvrnorm(1,rep(0,n), diag(rep(c(1,2,3,4),
                                         each=n/ngroups)))
mu <- exp(alpha + beta*x + e[fac] + e2)
y <- rpois(n = n, lambda = mu)
```

GLMM diagnostics





Residual checking for mixed effects models

- Check assumptions
  - Use both conditional and marginal residuals
  - Have a look at the `DHARMa` vignette
- Correct violations

Violation of some assumptions might be OK

Example 2

We had this model:

```
model3 <- gllvm(y, X = X, formula = ~(slp+asp|1),
                family = "binomial", num.lv = 0, beta0com = TRUE)
```

Let's compare it to:

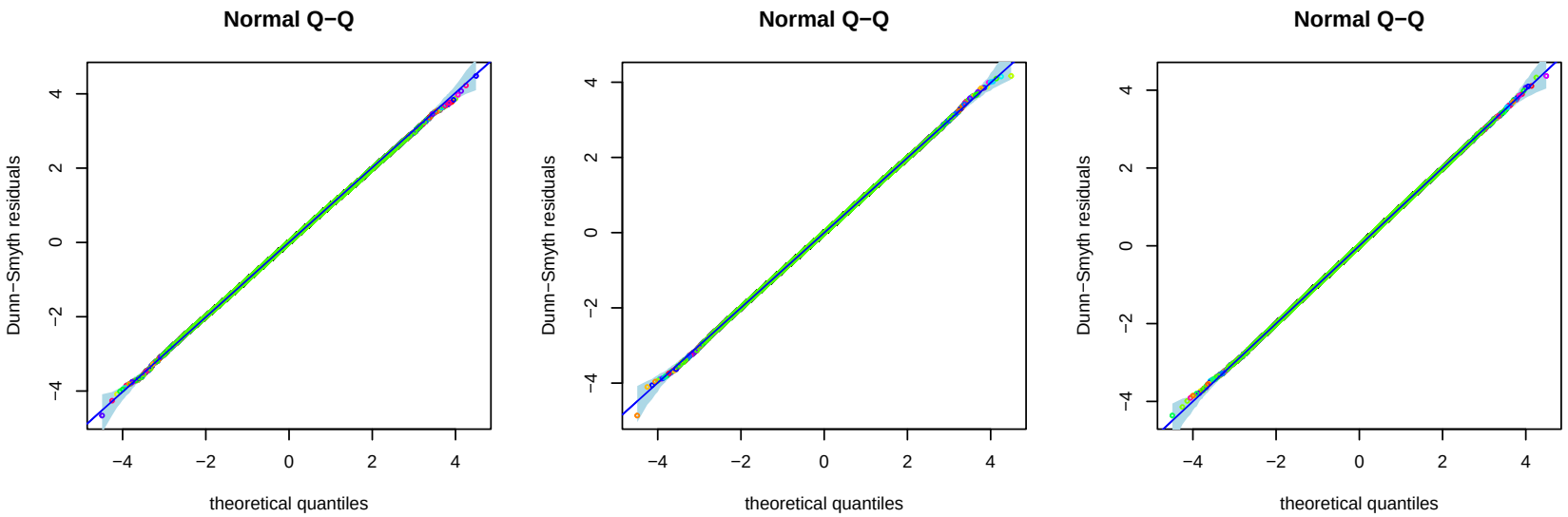
```
model4 <- gllvm(y, X = X, formula = ~(asp|1),
                row.eff = ~slp+asp, studyDesign = X,
                family = "binomial", num.lv = 0, beta0com = TRUE)
```

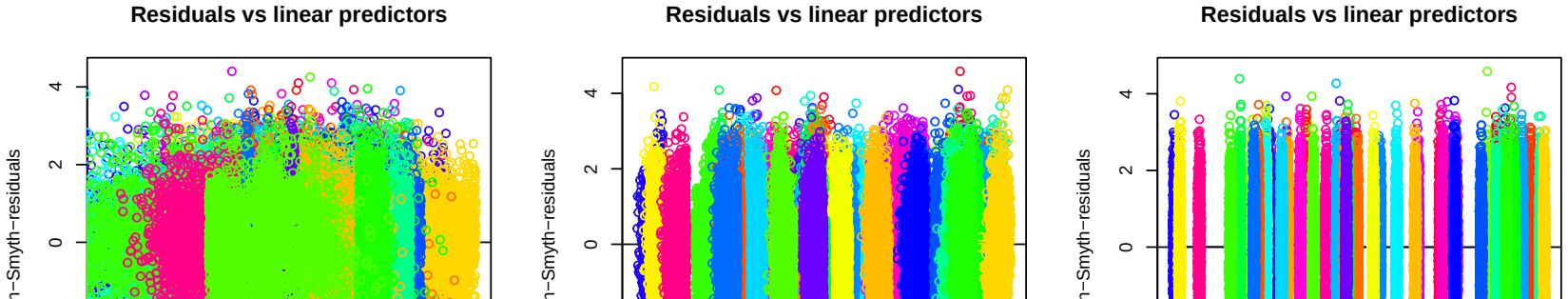
and perhaps

```
model5 <- gllvm(y, X = X, formula = ~(asp|1),
                family = "binomial", num.lv = 0, beta0com = TRUE)
```

Example 2: residuals

The residuals of all three look OK (but perhaps the last looks best):





Example 2: the models

The ecological assumptions of the models are a little different:

1) The first model assumes correlated species-specific responses to aspect and slope
2) The third model assumes species' responses to slope are the same
3) The third model assumes no responses to slope at all

So complexity is 1) $>$ 2) $>$ 3)

The models are nested, so we can use a hypothesis test, or information criteria, for comparison.

Example 2: comparison

```
anova(model4, model5) # model with mean slope effect is "better"
```

```
## Model  1 :  y ~ NULL
## Model  2 :  y ~ NULL
```

```
##   Resid.Df        D Df.diff P.value
## 1   141339   0.0000       0
## 2   141338 114.4127       1       0
```

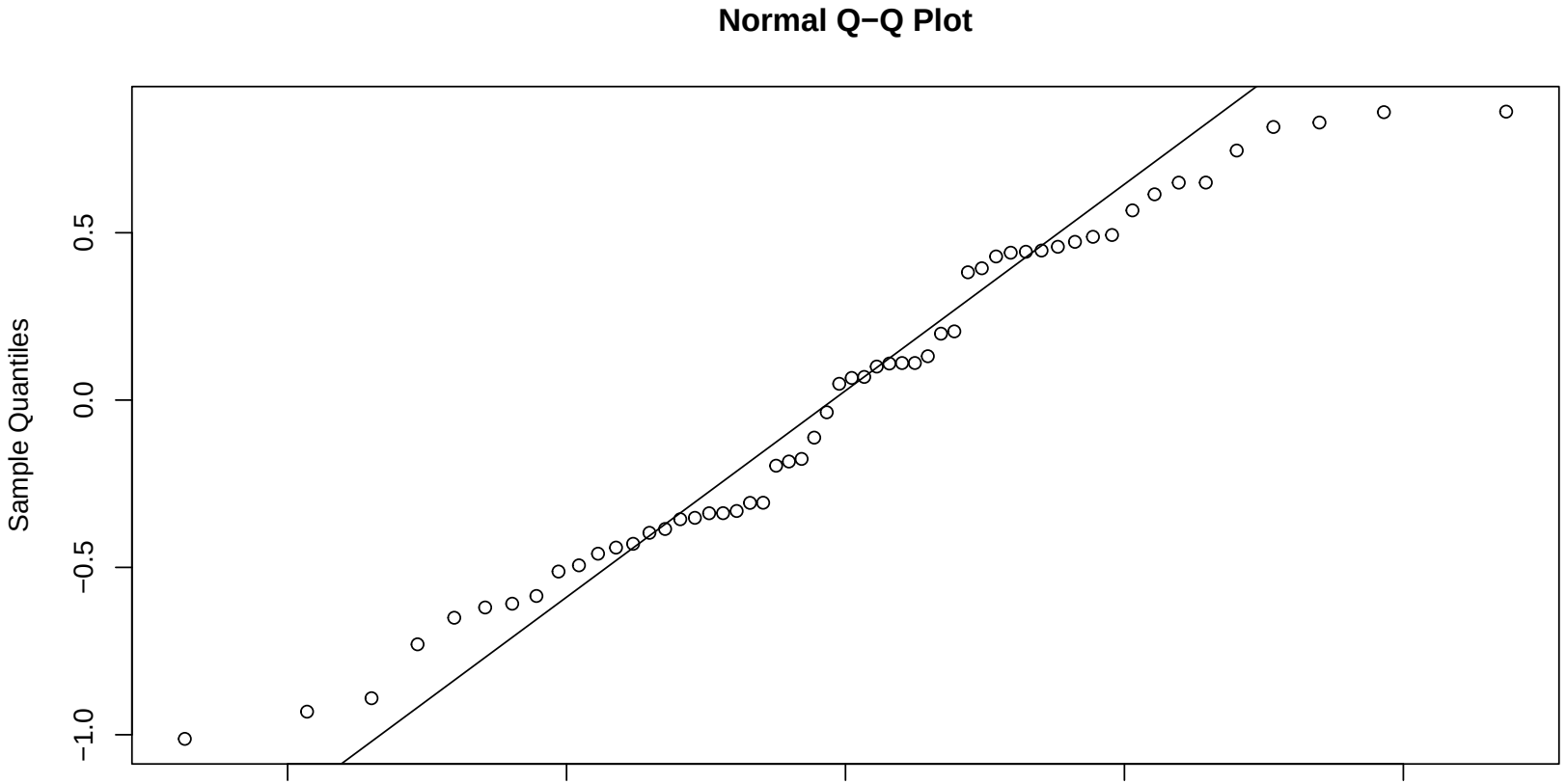
```
anova(model3, model4) # Model with slope REs is "better"; note null on the boundary
```

```
## Model  1 :  y ~ NULL
## Model  2 :  y ~ NULL
```

```
##   Resid.Df        D Df.diff P.value
## 1   141338     0.00       0
## 2   141335 12633.45       3       0
```

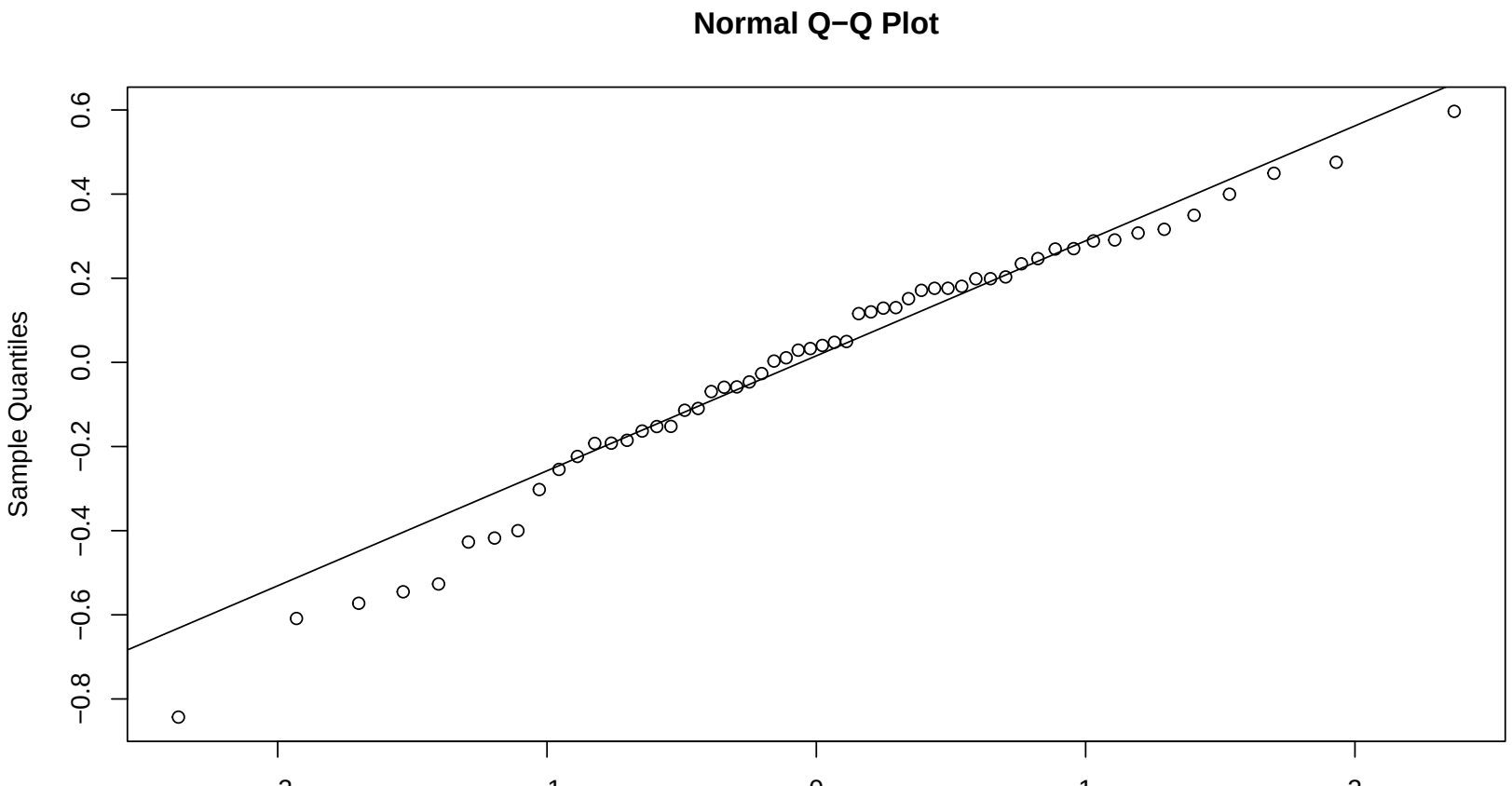
Example 2: examining the REs

```
qqnorm(coef(model3,"Br")["slp",])
qqline(coef(model3, "Br")["slp",])
```



Example 2: examining the REs

```
model6 <- gllvm(y, X = X, formula = ~(slp+bio_1+TE_canopy|1), family = "binomial", num.lv = 0, beta0com = TRUE
qqnorm(coef(model6,"Br")["slp",])
qqline(coef(model6, "Br")["slp",])
```



Example 2: information criteria

```
AIC(model3, model6) # lower is "better"
```

```
##        df      AIC
## model3  9 132304.1
## model6 14 120908.2
```

```
AICc(model3, model6)
```

```
## [1] 132304.1 120908.2
```

```
BIC(model3, model6)
```

Example 2: summary

```
##
## Call:
## gllvm(y = y, X = X, formula = ~(slp + bio_1 + TE_canopy | 1),
##     family = "binomial", num.lv = 0, beta0com = TRUE)
##
## Family:  binomial
##
## AIC:  120908.2 AICc:  120908.2 BIC:  121046.2 LL:  -60440 df:  14
##
## Informed LVs:  0
## Constrained LVs:  0
## Unconstrained LVs:  0
##
## Formula:  ~(slp + bio_1 + TE_canopy | 1)
## LV formula:  ~ 0
## Row effect:  ~ 1
##
## Random effects:
##  Name      Variance Std.Dev Corr
##  Intercept 0.8711   0.9333
##  slp       0.0893   0.2988   0.2118
##  bio_1     0.4446   0.6668  -0.0448 -0.2505
##  TE_canopy 0.0156   0.1250   0.2887  0.4636 -0.5255
##
## Coefficients predictors:
##            Estimate Std. Error z value Pr(>|z|)
## Intercept -0.769367   0.124935  -6.158 7.36e-10 ***
## slp        0.001113   0.040569   0.027  0.97811
## bio_1      0.236154   0.089521   2.638  0.00834 **
```

Model comparison

So, there are some more things we need to think about, besides formulating the model.

Likelihood ratio test

Is improved fit due to noise or is the alternative model actually better?

Procedure

- Fit two models: M_0 with k parameters and M_1 with r
- Calculate likelihood ratio $\Lambda = \log\left(\frac{\mathcal{L}(\mathbf{y};\Theta_0)_{M_0}}{\mathcal{L}(\mathbf{y};\Theta_1)_{M_1}}\right)$
- $\mathcal{L}(\mathbf{y};\Theta_0)_{M_0} \leq \mathcal{L}(\mathbf{y};\Theta_1)_{M_1}$
- $-2\Lambda \sim \chi^2(k_1 - k_0)$ under the null
- $p \geq 0.05$ difference in likelihood is due to sampling

LRT approximation assumptions

- $n \to \infty$
- Θ_0 contained in Θ_1: nested models
- The true parameter is in the interior of the parameter space
- Model is "identifiable"
- Hessian matrix is sufficiently close to the Fisher information
- y_i are independent

These assumptions may fail, especially in models more complex than VGLMs

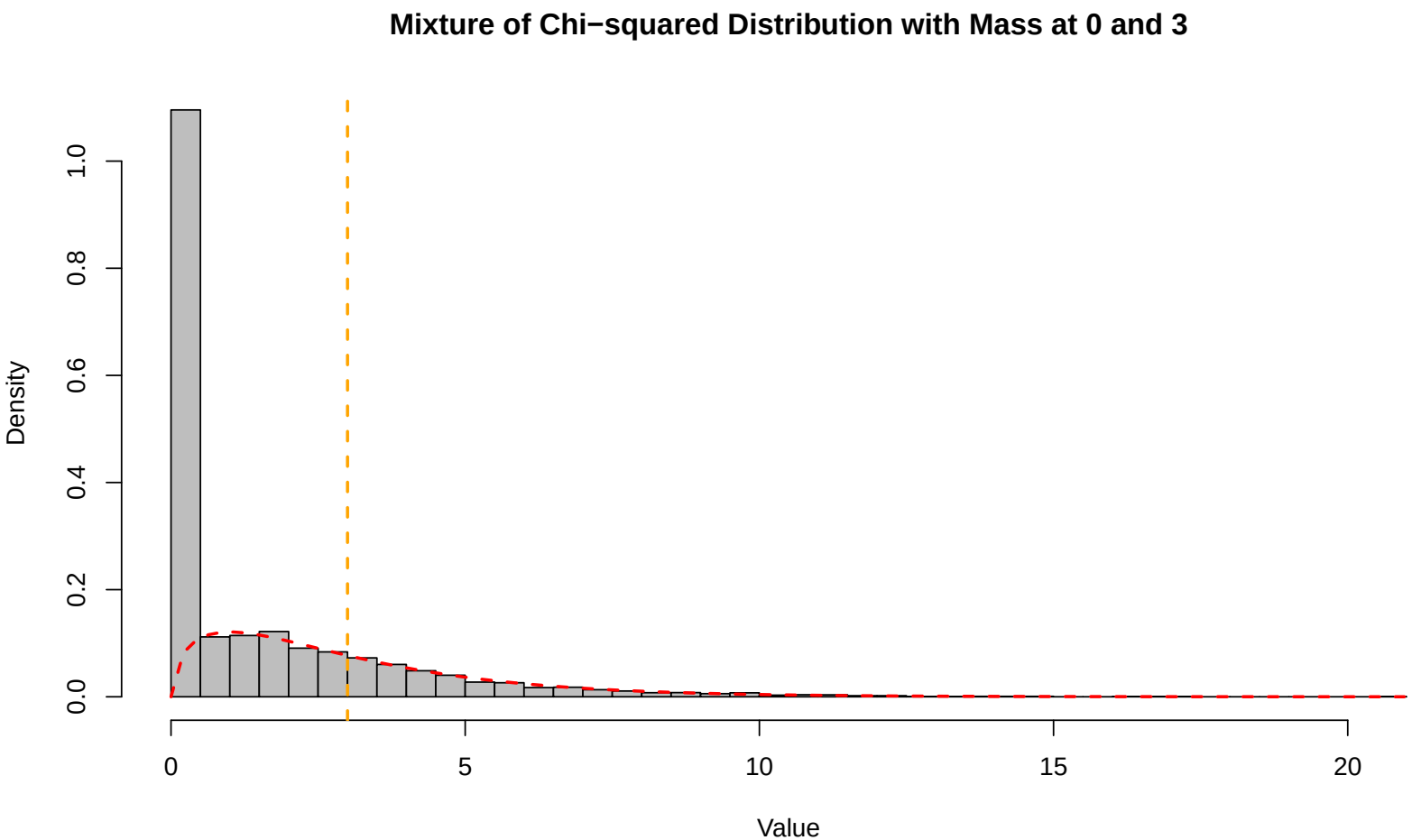
Alternatively: LRT by simulation.

Boundary issues

Variances σ^2 are positive only. If we compare to a model without random effect, it is like saying $\sigma^2 = 0$, so that it is "on the boundary".

And, the sampling distribution often has a spike at zero.

Boundary issues: example



So two things can happen: we can be too optimistic (test on the

Wald-statistic

The wald-statistic is reported in `summary`:

$$W = (\hat{\theta} - \theta)^2 / var(\hat{\theta}) \tag{9}$$

we test if the parameter estimate is different from zero. For this, we make the following assumptions:

- We have a good estimate of the standard error
- The estimator is normally distributed (large samples)
- The estimator is centered around the true parameter

Information criteria

A different paradigm:
Find the best model amongst a set of models.

Best:

- Penalise complexity (number of parameters)
- By fit (likelihood)

Most commonly:

1) AIC: Akaike's Information Criterion (Akaike 1974)
2) BIC: Bayesian Information Criterion (Schwarz 1978)

Lower = better

Akaike's Information Criterion

$$\text{AIC} = -2\mathcal{L}(\mathbf{y}; \Theta) + 2k \tag{10}$$

- Penalizes model complexity
- (approximately) Measures information loss to the true data generating process
- Asymptotically

AIC tends to select too complex models with little data. Finite sample correction (Sugiura 1978):

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \tag{11}$$

Find the model that predicts best

Bayesian Information Criterion

$$\text{BIC} = -2\mathcal{L}(\mathbf{y}; \Theta) + k \log(n) \tag{12}$$

So the penalty is different.

Find the model closest to the "true" model

Connection of AIC and LRT

Rule of thumb: difference of 2 points means a model is better

$$\begin{aligned}\Delta\text{AIC} =& \text{AIC}_{M_1} - \text{AIC}_{M_0} \\ &= 2\mathcal{L}(\mathbf{y};\Theta_0) - 2\mathcal{L}(\mathbf{y};\Theta_1) + 2k_1 - 2k_0 \\ &= -2\Lambda + 2(k_1 - k_0)\end{aligned} \tag{13}$$

So AIC with a rule of $= 2$ can be seen as a more liberal LRT (Sutherland et al. 2023)

The cult of (A)IC

Presentation by Mark Brewer

"Always use (A)IC for model comparison"

My perspectives

- Use common sense
- Do not blindly test all models ("dredging")
- Use model comparison techniques in moderation

Don't take the "best" model paradigm too seriously

Freedman's paradox

Just by chance, predictors with no relationship to the response will be selected.

Convergence

All of this assumes our model has converged. If the model has not converged, it may not be valid in any way.

Convergence

All of this assumes our model has converged. If the model has not converged, it may not be valid in any way.

see Ben Bolker's GLMM FAQ, `lme4` page on performance, and the `glmmTMB` troubleshooting vignette

Maximum Likelihood Estimation

At the maximum of the likelihood:

- The gradient is zero (tangent is straight)
- The hessian (of -LL) should
  - have positive diagonals
  - positive eigenvalues
  - be symmetric
  - and is thus invertible (we go up in both directions)
- Asymptotic covariance matrix is given by the inverse of the negative Hessian

These are important concepts to understand error messages and convergence in mixed-models.

Maximum Likelihood Estimation

At the maximum of the likelihood:

- The gradient is zero (tangent is straight)
- The hessian (of -LL) should
  - have positive diagonals
  - positive eigenvalues
  - be symmetric
  - and is thus invertible (we go up in both directions)
- Asymptotic covariance matrix is given by the inverse of the negative Hessian

These are important concepts to understand error messages and convergence in mixed-models.

Note: the method we use for estimating the model (optimisation) also makes assumptions.

Assessing arrival at the MLE

1. Stopping criteria
   - Maximum iterations
   - Gradient close to zero
   - Relative criterion: objective function value improvement
   - Absolute criterion: objective function becomes zero (say)
2. Gradient
3. Hessian

lme4 warnings: hessian

- `Warning: Problem with Hessian check (infinite or missing values?)`
- `Warning: Hessian is numerically singular: parameters are not uniquely determined`
- `Warning: Model failed to converge: degenerate Hessian with 2 negative eigenvalues`
- `Warning: Model is nearly unidentifiable: very large eigenvalue - Rescale variables?`
- `Warning: Model is nearly unidentifiable: very large eigenvalue ratio - Rescale variables?`

lme4 warnings: hessian

- `Warning: Problem with Hessian check (infinite or missing values?)`
- `Warning: Hessian is numerically singular: parameters are not uniquely determined`
- `Warning: Model failed to converge: degenerate Hessian with 2 negative eigenvalues`
- `Warning: Model is nearly unidentifiable: very large eigenvalue - Rescale variables?`
- `Warning: Model is nearly unidentifiable: very large eigenvalue ratio - Rescale variables?`

The most warning you will get in `gllvm` is a "sd.errors could not be calculated, due to singular fit".

Singular matrix

- determinant is zero
  - has zero eigenvalue(s)
- does not have inverse

$$\mathbf{HA} = \mathbf{I} \tag{14}$$

Numerical optimisation: best practices

1. Standardise (center and scale) explanatory variables
2. Try different optimisation routines
3. Different starting values
4. Carefully consider your model structure

Mixed-effects model troubleshooting

see Ben Bolker's GLMM FAQ

- Check data for mistakes
- Check model formulation
  - correct distribution and link-function
  - few random effects levels
  - few (non-zero) observations in a category
  - overly complex: drop terms with zero variances
- Double-check hessian calculation (finite differences)
- Use random effect as fixed effect instead (few levels)
- '?lme4::convergence' (and see the last line "convergence issues" for large datasets)

Difficult things

Everything gets more difficult when we use mixed-effects models:

- Residuals are harder to define
- We should test extra (random effect) assumptions
- Wald-statistic and p-values no longer (really) apply
- Hypothesis test does not always work well (on the boundary)
- Model selection does not usually work well (boundary issues, number of parameters is hard to define)

Difficult things

Everything gets more difficult when we use mixed-effects models:

- Residuals are harder to define
- We should test extra (random effect) assumptions
- Wald-statistic and p-values no longer (really) apply
- Hypothesis test does not always work well (on the boundary)
- Model selection does not usually work well (boundary issues, number of parameters is hard to define)

And even more so for VGLMMs

Crisis

- Inference in GLMMs can be difficult
- Use "ordinary" regression tools with caution

Take away tips

No free lunch in statistics

- There are some technical considerations
  - Predictor scaling
  - Convergence
  - How to safely use tools at your disposal
- Keep your model as simple as possible, but not simpler
- Take warnings seriously and check convergence
- Different packages have different ways of dealing with these issues
  - `gllvm` vs. `glmmTMB` vs. `lme4`
- Always check (residual) assumptions

End