Detecting dispersion problems in GLMMs

Melina de Souza Leite

melina.souza-leite@ur.de

Postodoctoral Researcher

University of Regensburg

Ecological data analysis

Increased complexity and flexibility in ecological data modeling:

- Generalized linear modes (GLMs)
- Mixture models (e.g. zero-inflated GLMs)
- Hiearchical/Multilevel models, GLMMs

- But still few tools for model diagnostics
- Problem: **failing to check model assumptions**

Can you trust your model?

Dispersion problems in count data

- Example count data:
  - Species richness
  - Abundance of indiviuals
  - Number of success (K) within a number of trials (N)
- Modeling count data, GL(M)M distributions:
  - Poisson
  - Binomial (K/N) proportion

Problem when data has **more or less variability than expected by the distribution** used for modeling:

UNDER or OVERDISPERSION

GOALS

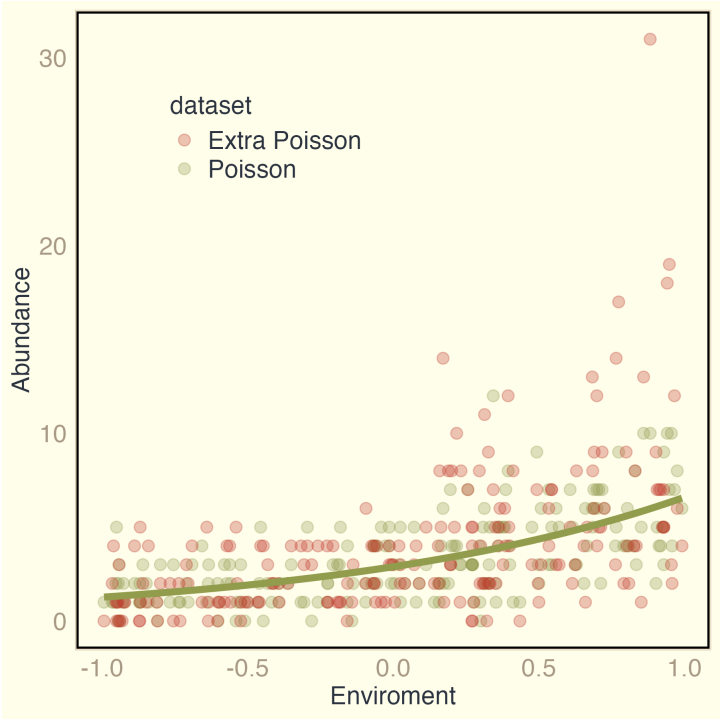
- Aware ecologists of dispersion problems with count data

- Identify and describe the 3 main causes by using model diagnostics tools with `DHARMa`

- Show modeling solutions for these causes

3 causes of dispersion problems

“Real” overdispersion:



More variance that expected by the model.

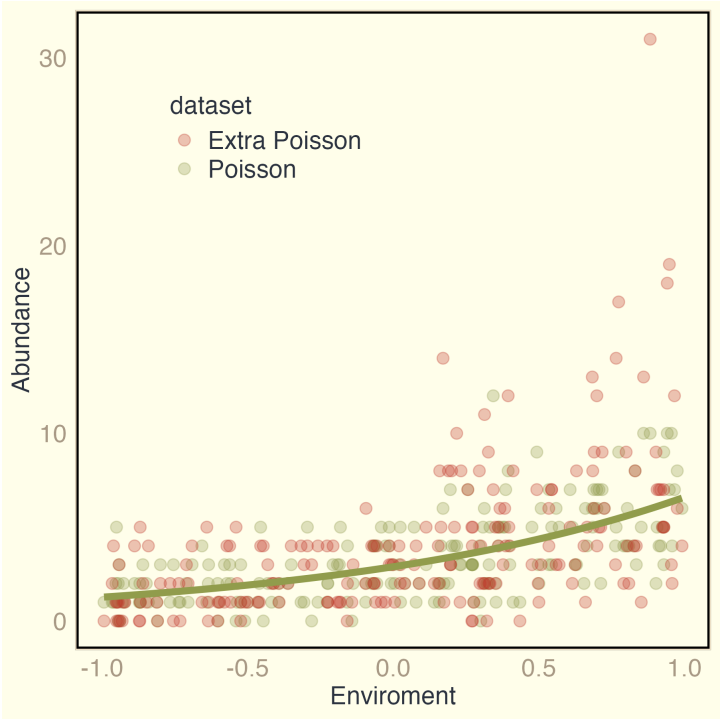
Heteroscedasticity:

Zero-inflation:

Examples for **OVERDISPERSION** for the Poisson distribution.

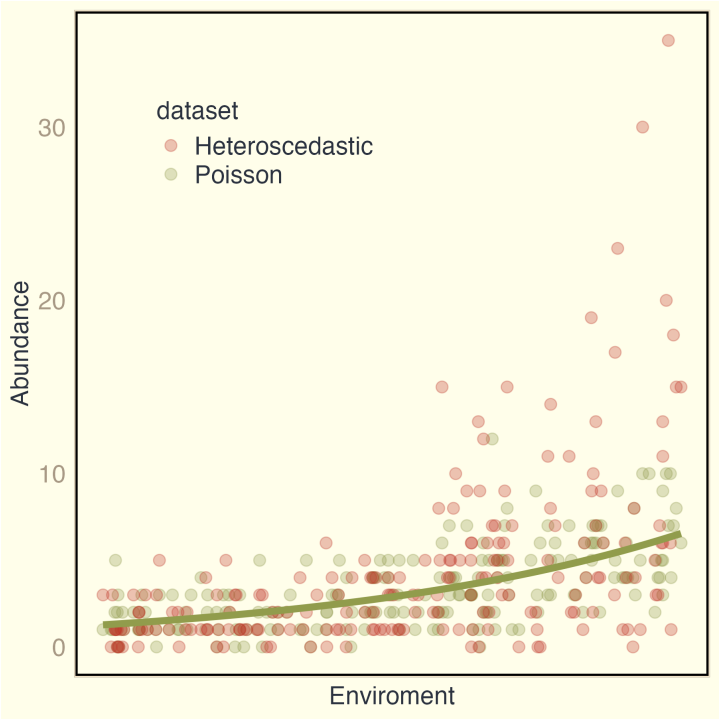
3 causes of dispersion problems

“Real” overdispersion:



More variance that expected by the model.

Heteroscedasticity:



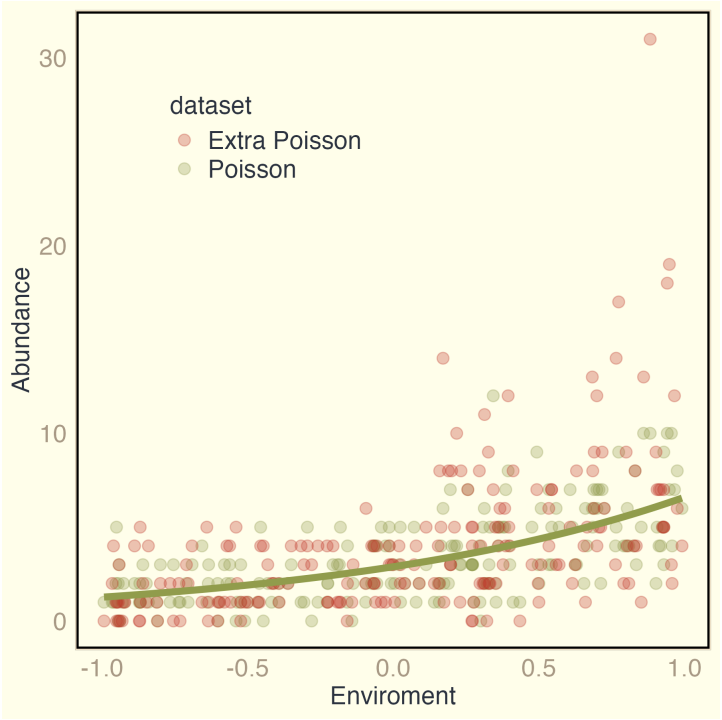
Variance increases/ decreases with a predictor.

Zero-inflation:

Examples for **OVERDISPERSION** for the Poisson distribution.

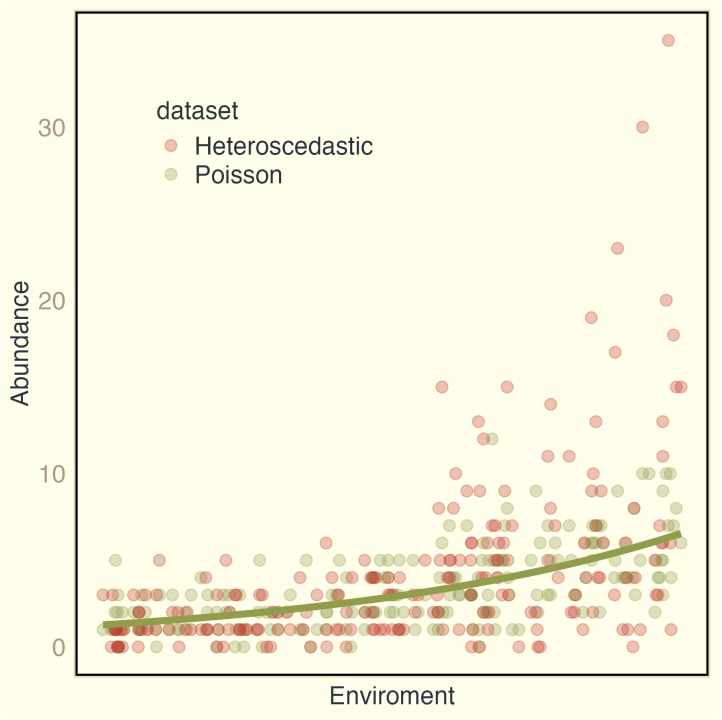
3 causes of dispersion problems

“Real” overdispersion:



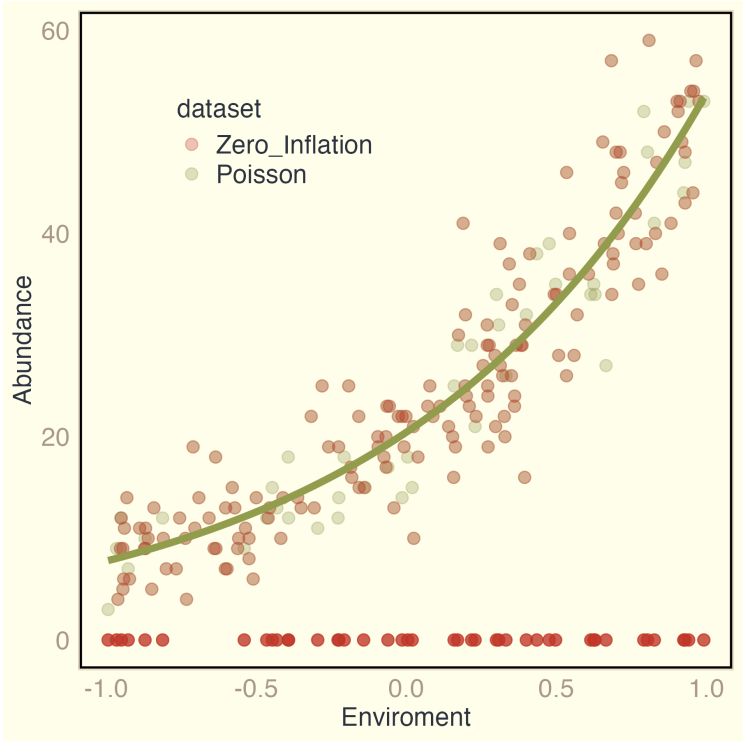
More variance that expected by the model.

Heteroscedasticity:



Variance increases/ decreases with a predictor.

Zero-inflation:



More zeros than expected by the model.

Examples for **OVERDISPERSION** for the Poisson distribution.

Consequences of dispersion problems

- Too small standard error of estimates -> narrower confidence intervals
- Larger chance of type I error: find an effect when it doesn’t exist

- Wrong estimates by ignoring other processes (e.g. zero-inflation causes) in your data-generating process.

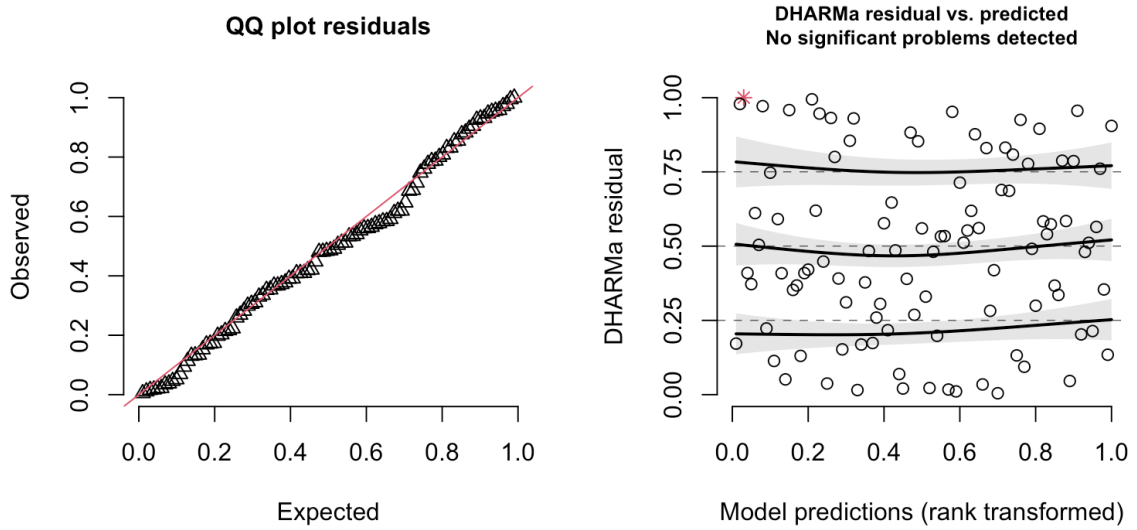
- Missing the opportunity to learn and get more info from your data/system. Ecological meanings for modeling/understanding unexpected variability?

Detecting dispersion problems with DHARMa

Residual diagnostics with DHARMa

- Scaled quantile residuals -> Simulating from the model
- Residuals between 0 and 1 for ANY model complexity or distribution
- Interpreted the SAME way:

> If your model is correctly specified, i.e. your have the “data-generating process”, scaled quantile residuals will present a uniform “flat” distribution between 0 and 1.



Detecting dispersion problems: DHARMa

Create DHARMa residuals

```
DHARMaResiduals <- simulateResiduals(model)
```

Test dispersion problems

```
testDispersion(DHARMaResiduals)
```

Test heteroscedasticity

```
plotResiduals(DHARMaResiduals,
              form = data$Environment1, # the predictor
              absoluteDeviation = T)
```

Test zero inflation

```
testZeroInflation(DHARMaRresiduals)
```

Solving dispersion problems: glmmTMB

Overdispersion

```
glmmTMB(observedResponse ~ Enviroment1,
        family = nbinom2(), ...) # from poisson()
```

Heteroscedasticity

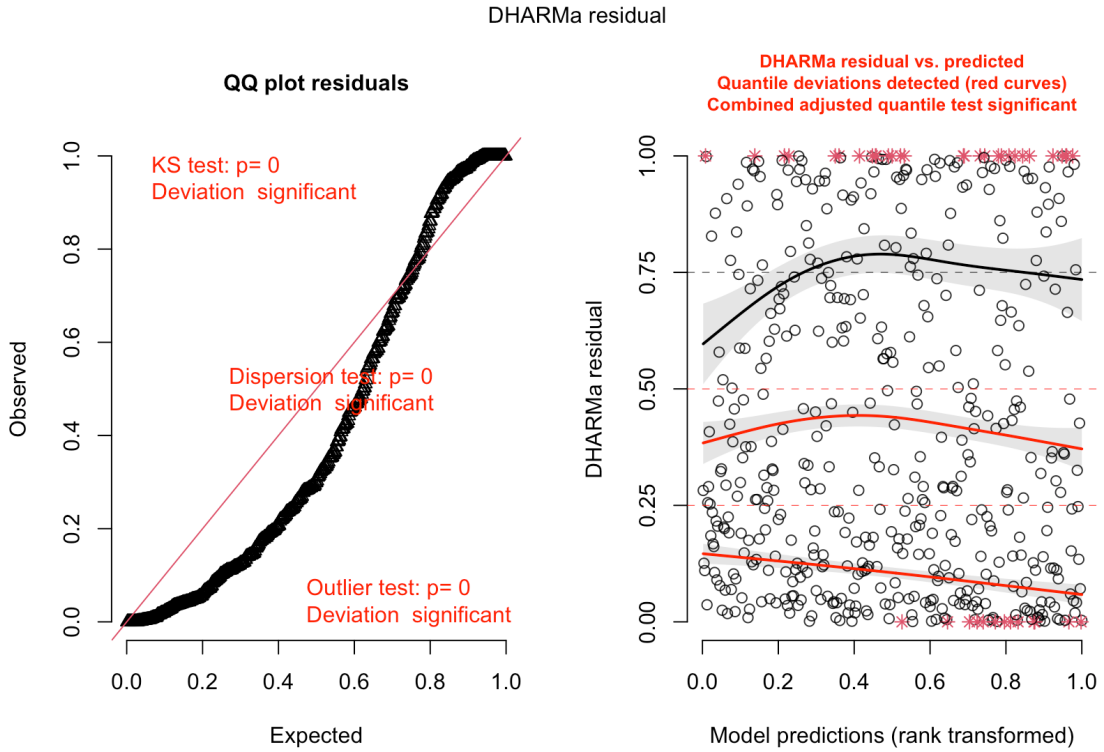
```
glmmTMB(observedResponse ~ Enviroment1,
        dispformula = ~ Enviroment1, # dispersion formula
        family = nbinom2(), ...) # needs to be negative binomial
```

Zero-inflation

```
glmmTMB(observedResponse ~ Enviroment1,
        ziformula = ~ 1, # zero-inflation formula / can add predictor
        family = poisson(), ...) # can also be negative binomial
```

What is the problem?

Problem 1



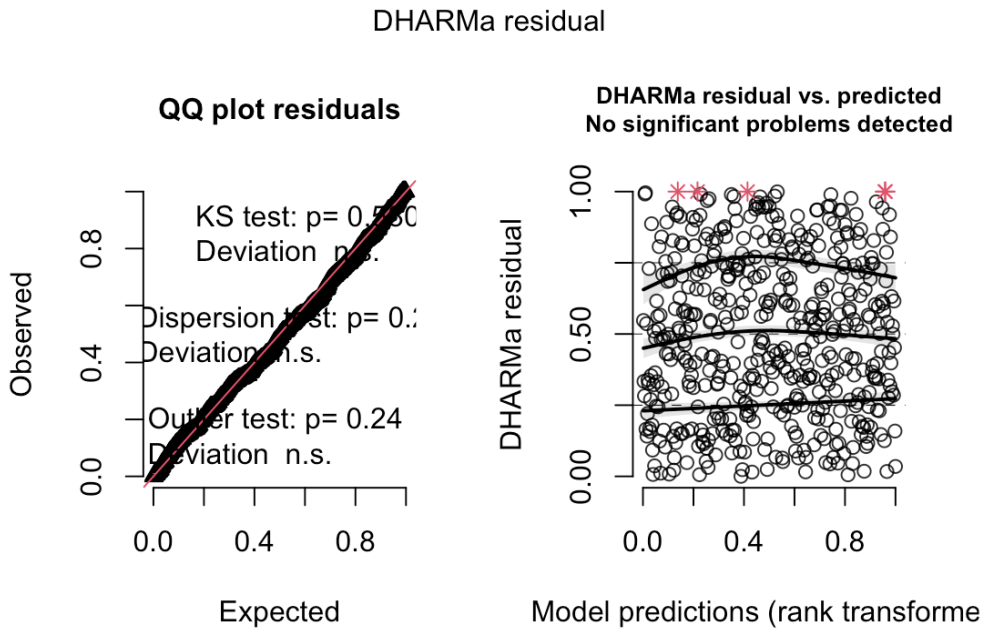

```
1  testDispersion(overRes,plot=F)
```

```
    DHARMa nonparametric dispersion test via sd of residuals
fitted vs.
    simulated

data:  simulationOutput
dispersion = 5.2662, p-value < 2.2e-16
alternative hypothesis: two.sided
```

Try a negative binomial model

Modeling “real” overdispersion

```
glmmTMB(observedResponse ~ Environment1 + (1|group),
        family = nbinom2(), data = overData)
```



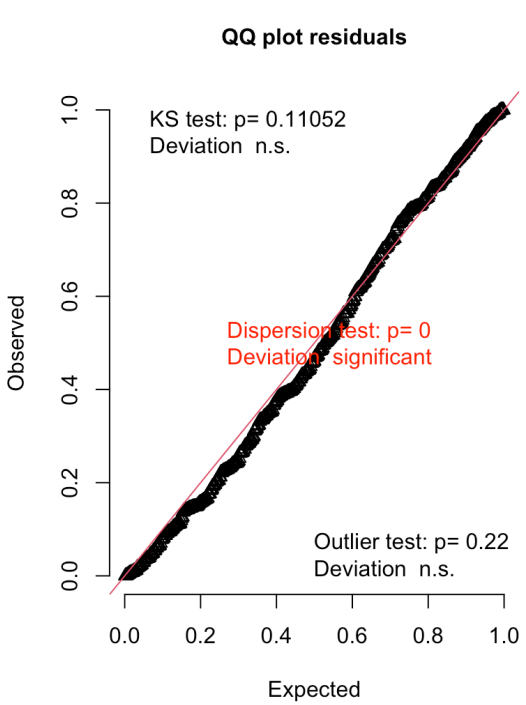

```
    DHARMa nonparametric dispersion test via sd of
residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 1.1935, p-value = 0.224
alternative hypothesis: two.sided
```

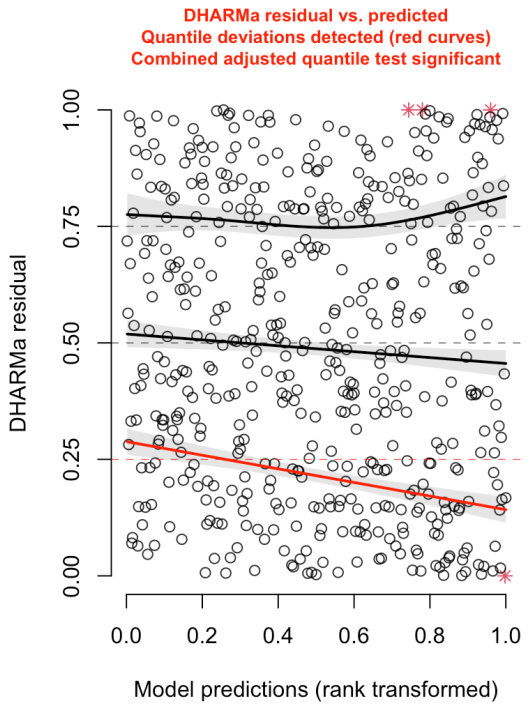
```
    DHARMa zero-inflation test via comparison to
expected zeros with
    simulation under H0 = fitted model

data:  simulationOutput
ratioObsSim = 0.97326, p-value = 0.848
alternative hypothesis: two.sided
```

Problem solved!

Problem 2

DHARMa residual



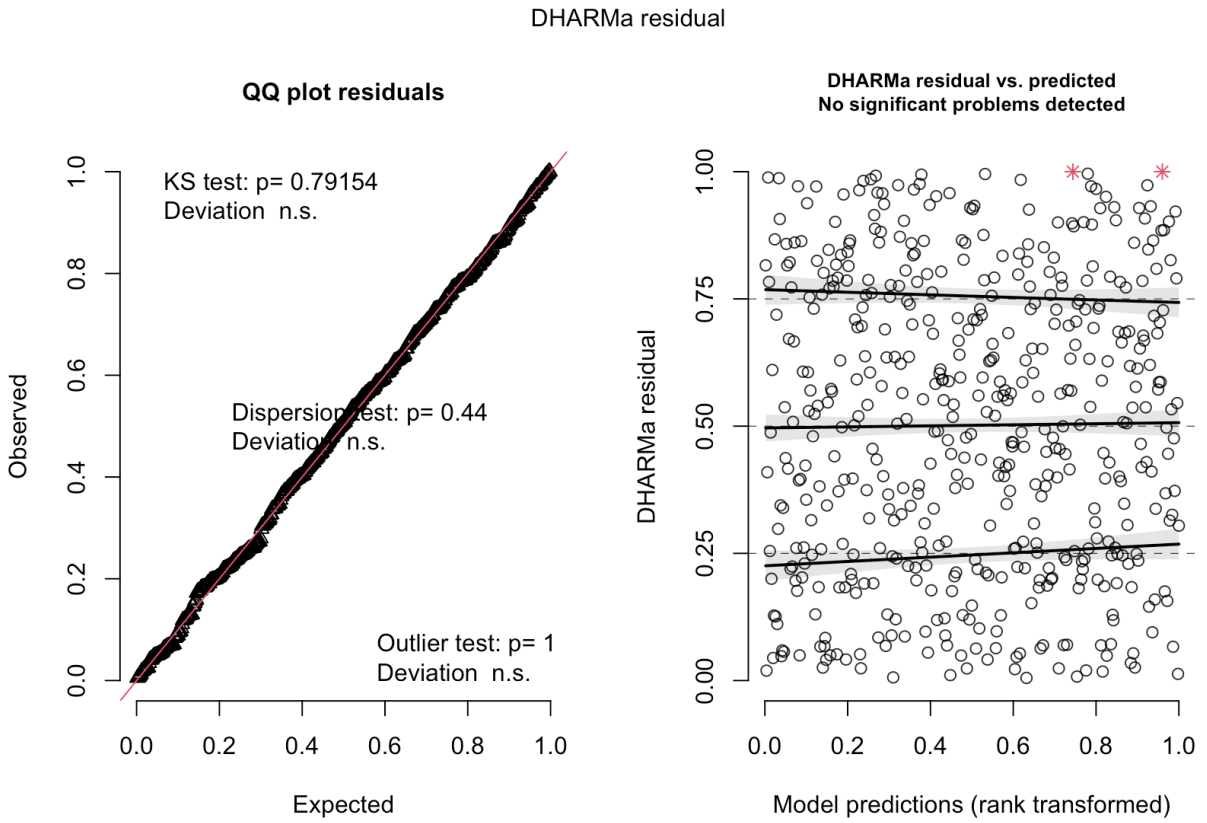



```
    DHARMa nonparametric dispersion test via sd of residuals
fitted vs.
    simulated

data:  simulationOutput
dispersion = 1.9199, p-value < 2.2e-16
alternative hypothesis: two.sided
```

Add a dispersion formula

Modeling heteroscedasticity

```
glmmTMB(observedResponse ~ Environment1 + (1|group),
        dispformula = ~ Environment1,
        family = nbinom2(), data = heteroData)
```

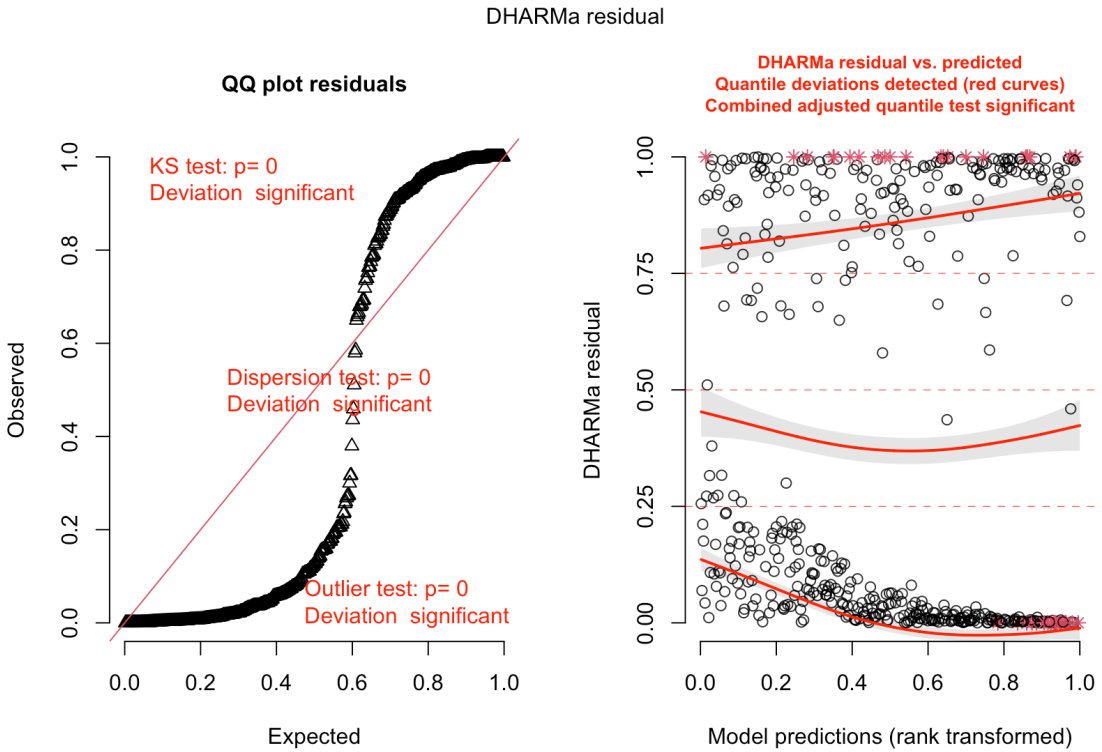



```
    DHARMa nonparametric dispersion test via sd of
residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 1.106, p-value = 0.44
alternative hypothesis: two.sided
```

```
    DHARMa zero-inflation test via comparison to
expected zeros with
    simulation under H0 = fitted model

data:  simulationOutput
ratioObsSim = 0.98758, p-value = 0.92
alternative hypothesis: two.sided
```

Problem Solved!

Problem 3

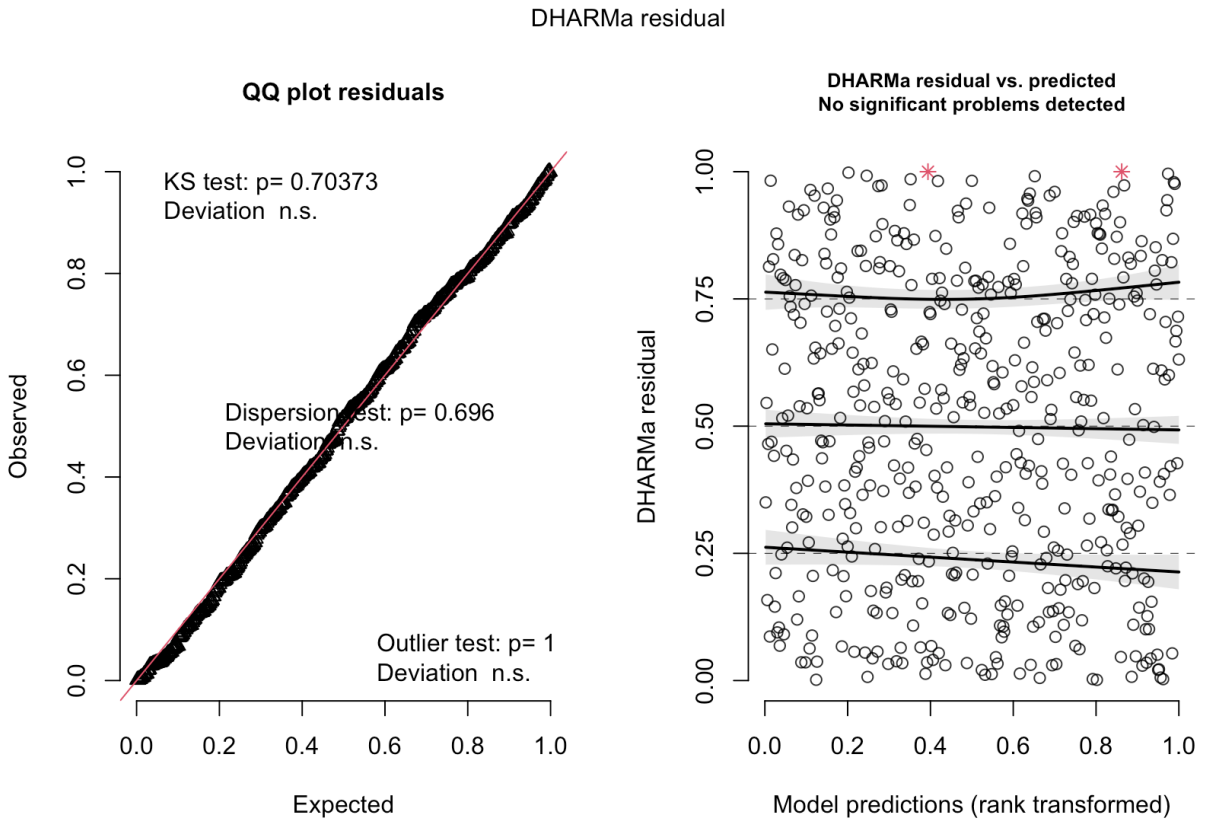



```
    DHARMa nonparametric dispersion test via sd of residuals
fitted vs.
    simulated

data:  simulationOutput
dispersion = 4.9124, p-value < 2.2e-16
alternative hypothesis: two.sided
```

Add a zero-inflation formula

Modeling zero-inflation

```
glmmTMB(observedResponse ~ Environment1 + (1|group),
        ziformula =  ~ 1,
        family = poisson(), data = zeroData)
```




```
    DHARMa nonparametric dispersion test via sd of
residuals fitted vs.
    simulated

data:  simulationOutput
dispersion = 1.0414, p-value = 0.696
alternative hypothesis: two.sided
```

```
    DHARMa zero-inflation test via comparison to
expected zeros with
    simulation under H0 = fitted model

data:  simulationOutput
ratioObsSim = 0.99592, p-value = 1
alternative hypothesis: two.sided
```

Problem Solved!

Solving dispersion problems

- Sometimes, residual patterns will not tell you which is the cause of overdispersion. E.g.:
  - ‘Real’ overdispersion will show significant test for zero-inflation, and vice-versa.
  - ‘Real’ overdispersion and zero-inflation may have significant heteroscedasticity/.

- Additional check: fit models addressing the potential problems and compare their fit (e.g. AIC, LRT) and residuals diagnostics.

Don’t always assume the most complex/complicated model is the correct one!

Conclusion

- There are many causes of dispersion problems in GLMMs
- Use `DHARMa` residuals tools to detect them
- Address the problem with adequate models, e.g, `glmmTMB`

Take home message

- Models should ALWAYS be checked: residual diagnostics!

- Avoid the oversimplistic view of dispersion problems

- Detecting and addressing the causes of dispersion problems may also be informative for your system/data.

Comming soon: Leite et al. *in prep.* Dispersion tests in GLMMs - a methods comparison and practical guide.

Thank you!

Vielen Dank!

Acknowledgements to Florian Hartig, Max Pichler, and the Theoretical Ecology Lab group