5.2.3. Over-dispersion in GLMM

Over-dispersion is "the polite statistician's version of Murphy's law: if something can go wrong, it will" (Crawley, 2013). It is particularly relevant when working with count or proportion data where variation of a response variable does not strictly conform to the Poisson or binomial distribution, respectively. Fundamentally, over-dispersion causes poor model fitting where the difference between observed and predicted values from the tested model are larger than what would be predicted by the error structure. To identify possible over-dispersion in the data for a given model, divide the deviance (−2 times the log-likelihood ratio of the reduced model, e.g. a model with only a term for the intercept, compared to the full model; see McCullagh and Nelder, 1989) by its degrees of freedom: this is called the dispersion parameter. If the deviance is reasonably close to the degrees of freedom (i.e. the dispersion or scale parameter = 1) then evidence of over-dispersion is lacking.

Causes of over-dispersion can be apparent or real. Apparent over-dispersion is due to model misspecification, i.e. missing covariates or interactions, outliers in the response variable, non-linear effects of covariates entered as linear effects, the wrong link function, etc. Real over-dispersion occurs when model misspecifications can be ruled out, and variation in the data is real due to too many zeros, clustering of observations, or correlation between observations (Zuur *et al.*, 2009). Solutions to over-dispersion can include: i) adding covariates or interactions, ii) including individual-level random effects, e.g. using bee as a random effect, where multiple bees are observed per cage, iii) using alternative distributions: if there is no random effect included in the model consider quasi-binomial and quasi-Poisson; if there are, consider replacing Poisson with negative-binomial, and iv) using a zero-inflated GLMM (a model that allows for numerous zeros in your dataset, the frequency of the number zero is inflated) if appropriate. Over-dispersion cannot occur for normally distributed response variables because the variance is estimated independently from the mean. However, residuals often have "heavy tails", i.e. more outlying observations than expected for a normal distribution, which nevertheless can be addressed by some software packages.