# 5.2. Generalised Linear Mixed Models (GLMM)

A central dogma in statistical analyses is always to apply the simplest statistical test to your data, but ensure it is applied correctly (Zuur *et al*., 2009). Yes, you could apply an ANOVA or linear regression to your data, but in the **vast majority of cases**, the series of assumptions upon which these techniques are based are violated by 'real world' data and experimental designs, which often include blocking of some kind or repeated measures. The assumptions typically violated are: i) normality; ii) homogeneity; and iii) independence of data.

1. Normality

Although some statistical tests are robust to minor violations of normality (Sokal and Rohlf, 1995; Sokal and Rohlf, 2012), where your dependent variable/data (i.e. the residuals, see section 5.1.) are clearly not normal (positively/negatively skewed, binary data, etc.), a better approach would be to account for this distribution within your model, rather than ignore it and settle for models that poorly fit your data. As an obvious example, a process that produces counts will not generate data values less than zero, but the normal distribution ranges from -∞ to +∞.

2. Homogeneity of variances

As stated above, minor violations of normality can be tolerated in some cases, and the same could be said for heterogeneous dependent variable/data (non-homogenous variance across levels of a predictor in a model, also called heteroscedasticity). However, marked heterogeneity fundamentally violates underlying assumptions for linear regression models, thereby falsely applying the results and conclusions of a parametric model, making results of statistical tests invalid.

3. Independence of data

See section 3.1.2. Simply, if your experimental design is hierarchical (e.g. bees are in cages, cages from colonies, colonies from apiaries) or involves repeated measures of experimental units, your data strongly violate the assumption of independence and invalidate important tests such as the *F*-test and *t*-test; these tests will be too liberal (i.e. true null hypotheses will be rejected too often).

GLMMs are a superset of linear models, they allow for the dependent variable to be samples from non-normal distributions (allowed distributions have to be members of the one and two parameter exponential distribution family; this includes the normal distribution, but also many others). For distributions other than the normal, the statistical model produces heterogeneous variances, which is a desired result if they match the heterogeneous variances seen in the dependent variable. The 'generalised' part of GLMM means that, unlike in linear regression, the experimenter can choose the kind of distribution they believe underlies the process generating their data. The 'mixed' part of GLMM allows for random effects and some degree of non-independence among observations. Ultimately, this level of flexibility within GLMM approaches allows a researcher to apply more rigorous, but biologically more realistic, statistical models to their data.

One pays a price for this advantage. The basic one is that the state of statistical knowledge in this area, especially computational issues, lags behind that for models based on the normal distribution. This translates into software that is buggy, which can result in many kinds of model estimation problems. Also, there are now far more choices to be made, such as which estimation algorithm to use (e.g. the Laplace and quadrature methods do not allow for correlation among observations), and which link function to use. The link function "links" the data scale to the model scale. For example, if dependent variable is assumed to be generated by a Poisson process, the typical link function is the log, i.e. $\log(E(\mu)) = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZU}$; in words, the natural log of the expected value of the mean is modelled as a sum of fixed and random effects). Tests are based on asymptotic behaviours of various quantities, which can give quite biased results for small samples. One is simultaneously working on two scales: the data scale and the model scale; the two are linked, but model estimates and hypothesis tests are done on the model scale, and so are less easily interpretable (i.e. a change in unit value of a predictor variable has different effects on the data scale depending on whether one is looking at low values or high values). One parameter and two parameter members of the exponential family have to be handled quite differently. Over-dispersion cannot be handled using a quasi-likelihood approach (e.g. using a quasi-binomial distribution); instead, appropriate random effects need to be added (e.g. one for every observation), which can lead to models with many parameters (Note: Over-dispersion means that one has a greater variability than expected based on the theoretical statistical distribution; for example the expected variance of a Poisson distribution is its mean - if the observed variance is larger than the estimated mean, then there is over-dispersion). For some one-parameter members of the exponential distribution (e.g. Poisson, binomial), one can try the analogous two-parameter member (e.g. for a Poisson distribution, it is the negative binomial distribution; for the binomial it is the beta-binomial). Model diagnosis is in its infancy. While we encourage researchers to explore the use of these models, we also caution that considerable training is necessary for both the understanding of the theoretical underpinnings of these models and for using the software. A recent book using GLMM methodology is Stroup (2013), which developed from experience with researchers in agriculture and covering both analyses and design of experiments. He discusses in detail what we can only allude to superficially; a shortcoming is that the worked examples only use the SAS software.

## 5.2.1. General advice for using GLMMs

## 5.2.2. GLMM where the response variable is mortality

## 5.2.3. Over-dispersion in GLMM