1. Introduction

Generalized linear mixed models (GLMMs) are a powerful class of statistical models that combine the characteristics of generalized linear models (Chapter xxx) and *mixed models* (models that include both fixed and random predictor variables: see below). They handle a wide range of response distributions, and a wide range of scenarios where observations have been sampled in some kind of groups rather than completely independently. While they can't do everything – there are still situations where an expert might choose custom-built models for greater flexibility – GLMMs are fast, powerful, can be extended to handle additional complexities such as zero-inflated responses, and can often be fitted with off-the-shelf software. The only real downsides of GLMMs are due to their generality: (1) some of the standard recipes for model testing and inference that you have learned previously may not apply, and (2) it's easy to build plausible models that are too complex for your data to support. GLMMs are still part of the statistical frontier, and not all of the answers about how to use them are known (even by experts), but this chapter will try to provide practical solutions to allow you to use GLMMs with your data.

Generalized linear models, as discussed in Chapter xxx, allow modeling of many kinds of response variables, particularly those with binomial and Poisson distributions; you should definitely be comfortable with the material in Chapter xxx before attempting the methods described in this chapter. In contrast, the idea of mixed models, and the distinction between *fixed effects* (the typical way that you compare differences between treatments or the effects of continuous predictor variables) and *random effects* (roughly speaking, experimental or observational blocks within which you have several observations) may be new to you. Models with Normally distributed responses that incorporate some kind of random effects or blocking are called *linear mixed models* (LMMs); they are a special, slightly easier case of GLMMs. Hopefully you have a passing acquaintance with the idea of experimental blocks from a previous statistics course, or from a basic textbook such as Gotelli and Ellison (2004) or Quinn and Keough (2002), but this chapter will review the basic idea. If you are already well-versed in ANOVA approaches to blocked experimental designs, you may actually have to unlearn some things, as modern approaches to random effects are quite different from the classical approaches taught in most statistics courses.

As well as using a different conceptual definition of random effects, modern mixed models are more flexible than classical ANOVA approaches, allowing (e.g.) non-Normal responses, unbalanced experimental designs, and more complex grouping structures (crossed random effects: see below). Equally important is a new philosophy: modern approaches use a model-building approach rather than a hypothesis-testing approach, as discussed in Chapter XXX. You can (and should) still test hypotheses, but instead of a list of F statistics and p values the primary outputs of the analysis are quantitative parameter estimates describing (1) how the response variable changes as a function of the fixed predictor variables and (2) the variability among the levels of the random effects.

While many ecologists have embraced the idea of model-based approaches, random effects such as variation among experimental blocks are often neglected in model-based analyses because they are relatively difficult to incorporate in custom-built statistical models. While one can use tools such as WinBUGS, AD Model Builder, or SAS PROC NLMIXED to incorporate such components in a general model, generalized linear mixed models are general enough to encompass many of the problems that ecologists will encounter, yet can be fitted with off-the-shelf software, without building your own model completely from scratch.

2. Running examples

Before jumping into the technical and philosophical details of random and fixed effects, I will introduce several real case studies from the literature or from my own work which will serve as running examples.

- *Coral symbiont defense*: McKeon et al. (2012) ran a field experiment with coral (XXX spp.) inhabited by invertebrate symbionts (crabs and shrimp) and exposed to predation by sea stars (*Culcita* spp.) to understand the complementary or synergistic effects of symbionts: were symbionts from different species more, less, or as effective in defending corals from attack as one would expect based on their independent effects? The design is a randomized complete block design with a small amount of replication (2 replications per treatment per block; 4 treatments (no symbionts, crabs alone, shrimp alone, both crabs and shrimp), with each of these units of 8 repeated in 10 blocks). The response is binomial with a single trial per unit (also called Bernoulli or binary); treatment, a categorical variable, is the only fixed effect input variable; block is the only grouping variable, with intercepts (i.e. baseline predation probability) varying among blocks.

- *Gopher tortoise shells*: Ozgul et al. (2009) analyzed the numbers of fresh gopher tortoise shells in different areas to estimate whether shells were more common (implying a higher mortality rate) in areas with higher disease prevalence. The response is the count of fresh shells, for which we will consider Poisson and negative binomial distributions; seroprevalence of mycoplasma (i.e. the fraction of tortoises carrying antibodies against the disease) is the single, continuous, fixed predictor variable. We would like to consider both year and site as crossed grouping variables (see below) with variation in baseline numbers among them, but as discussed below we treat year as fixed because there are only 3 levels (2003-2005).

- *Red grouse ticks*: Elston et al. (2001) used data on numbers of ticks sampled from the heads of red grouse chicks in Scotland to explore patterns of aggregation. Ticks have potentially large fitness and demographic consequences on red grouse individuals and populations, but the goal in this particular paper was just to decompose patterns of variation into different

scales (within-brood, within-site, by altitude and year). The response is the tick count (again Poisson or negative binomial); altitude (treated as continuous) and year (treated as categorical and fixed because there are only 3 years; it could be treated as continuous, but it costs only one additional parameter to relax the assumption of linearity in this case) are fixed input variables. Individual within brood and brood within location are nested random-effect grouping variables, with the baseline expected number of ticks (intercept) varying among groups. (See *overdispersion* for an explanation of treating individual as a random effect.)

In each of these case studies, the data are non-Normal (counts in the tick and gopher tortoise examples and binary (attacked/not attacked) in coral symbiont example), and the structure of the data includes some kind of grouping (experimental blocks for the sea star example; areas and years for the gopher tortoise example; and individuals within broods within sites, for the tick example). These are the basic characteristics that require the use of GLMMs.

3. Concepts

3.1 Model definition

3.1.1 Random effects

The traditional way to look at random effects is as a way to do the correct statistical tests when some observations are correlated. When samples are collected in groups (within species in the example above, or within experimental blocks of any kind), there will be some variation within groups (σ^2_{within}) and some among groups (σ^2_{among}); the total variance is $\sigma^2_{\text{total}}=\sigma^2_{\text{within}}+\sigma^2_{\text{among}}$; and the correlation between any two observations in the same group is $\rho = \sqrt{\sigma^2_{within} / \sigma^2_{total}}$ (observations that come from *different* groups are uncorrelated). Such grouping which violates the assumption of independent observations that is part of most statistical models. Sometimes one can solve this problem by analyzing the data at the level of independent groups, rather than at the level of partially correlated individual observations. For example, in a balanced, nested design where fixed effects are constant within groups – for example, if were testing for the differences between deciduous and evergreen plants, where every member of a species has the same leaf habit – we could simply calculate species averages, throwing away the variation within species, and do a *t*-test between the deciduous and evergreen species means. This procedure is exactly equivalent to testing the fixed effect in a classical mixed model ANOVA with a fixed effect of leaf habit and a random effect of species. This classical approach correctly incorporates the facts that (1) repeated sampling within species reduces the uncertainty associated with within-group variance, but (2) we have fewer *independent* data points than observations – in this case, as many as we have groups (species) in our study.

These basic ideas underlie all classical mixed model ANOVA analyses, although the formulas get more complex when treatments vary within grouping variables, when different fixed effects can vary at the levels of different grouping variables (e.g., randomized block and split-plot designs). Murtaugh (2007) points out that mixed model ANOVA is unnecessarily complicated for simple nested designs, recommending simpler approaches like the averaging procedure described above. However, mixed model ANOVA is still extremely useful for a wide range of more complicated designs, and as discussed below, traditional mixed model ANOVA itself falls short for cases such as unbalanced designs or non-Normal data.

We can also think of random effects as a way to combine information from different levels within a grouping variable. Suppose that you had estimated photosynthetic rate from multiple individuals from each of many species. If you had only a few samples from a few species, you might be forced to *pool* the data, ignoring the differences among species. Pooling assumes that σ^2_{among} is effectively zero, so that the individual observations are uncorrelated ($\rho=0$). On the other hand, if you had many individuals from each species, and especially if you had a small number of species, you might choose to estimate the photosynthetic rate for each species individually, or in other words to estimate a fixed effect parameter for each species. Treating the grouping factor as a fixed effect assumes that information about one species gives us no information about any other species; this is equivalent, for the purposes of parameter estimation, to treating σ^2_{among} as infinite. Treating species as a random effect compromises between the extremes of pooling and estimating separate (fixed) estimates; we acknowledge, and try to quantify, the variability among species. Because the species are assumed to come from a population with a well-defined mean, the predicted photosynthetic rates for each species are a weighted average between the mean for that species and the overall mean of the population; the smaller and noisier the sample for a particular species, the more its prediction is "shrunk" toward the population mean – the random effects predictions are sometimes called *shrinkage estimates* (Figure 1). (For technical reasons, the value we retrieve from the model for each species is called a "prediction" or more generally a *conditional mode*, rather than an "estimate"; they are often loosely called "random effects", but this can get confusing …) For example, if we had estimated the maximum photosynthetic rate for species 1 as 5 (in some sensible units), with a variance of 1 (in the same units), while the mean rate of all the species in the group was 8 with a variance of 3, then our predicted value would be $(\mu_{species}/\sigma^2_{species} + \mu_{group}/\sigma^2_{among})/(1/\sigma^2_{species} + 1/\sigma^2_{among})$ = $(5/1+8/3)/(1/1+1/3)=5.75$. Because $\sigma^2_{species} < \sigma^2_{among}$, the prediction is closer to the species-specific value than to the group mean. (Stop and convince yourself that this formula agrees with verbal description above of how variance-weighted averaging works when σ^2_{among} is either very small or very large.)

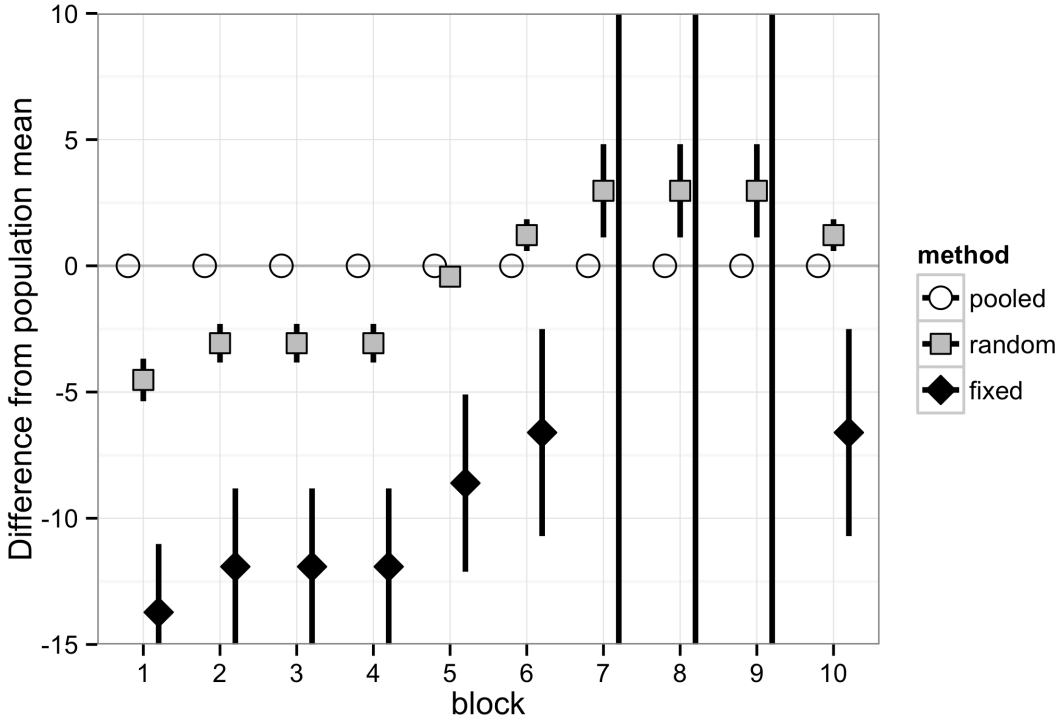

Figure 1: Estimated block effects from the *Culcita* analysis (deviations from population mean). The mixed estimates are (mostly) intermediate between the fixed estimates and the pooled estimate of zero. (In this case there is actually a problem with the fixed effects, which can't be estimated properly for blocks 7-9 because *all* corals in those blocks were attacked, which makes the estimate of the attack probability infinite on the logit scale, and messes up the estimates of the other block deviations as well, so that the estimated random and fixed effects in blocks 6 and 8 have different signs … this is another reason to prefer the shrinkage estimator, which handles this case just fine!) Notice also that the confidence intervals on the estimates are much smaller for the mixed-effect than the fixed-effect estimates; this doesn't matter very much in this example because we're not particularly interested in the block effects, but could be very important if the random effects were (e.g.) conservation areas that we wanted to prioritize.

Random effects are especially useful when we have (1) lots of levels (i.e. many species), (2) relatively little data on each species (although we need multiple samples from at least some of the species), and (3) uneven sampling across species.

This idea of a random effect as an effect where we combine information from different levels differs from the standard frequentist definition, which is a categorical variable whose levels were chosen *at random from a larger population*, e.g. species chosen at random from a list of endemic species. This definition is philosophically coherent, and you will encounter researchers (including reviewers and supervisors) who insist on it, but it is practically problematic. At first glance it rules out using species as random effect when you have observed *all* of the endemic species at your field site (since your sample is no longer a sample from a larger population), or using year as a random effect (since researchers very rarely choose the years in which to run an experiment

randomly – usually they either use a series of consecutive years, or the haphazard set of years when they were able to run replicates of the experiment). This problem applies to both the gopher tortoise and tick examples, each of which (coincidentally) have samples collected in three successive years. It would be nice to be able to formally generalize across years (i.e., calculate the among-year variance, rather than the usual procedure of just hoping that the sample is reasonably representative of other, unsampled years), but it simply isn't practical with such a short sample.

- TABLE 1.

You may want to treat an effect as random if:

- you have sampled at least 5 levels of the grouping factor;
- you don't want to test hypotheses about differences between responses at specific levels of the grouping factor;
- you do want to quantify the variability among levels of the grouping factor;
- you want to making predictions about unobserved levels of the grouping factor;
- you want to use shrinkage estimates to combine information across levels;
- there is variation in information per level (samples or noisiness);
- your levels are(randomly?) chosen from a larger population.
- the effect is a nuisance variable (i.e. it is not of direct interest, but should be controlled for)

cf. Crawley (2002), Gelman (2005)

A more useful definition of a random effect is a predictor variable where you are interested in making inferences about the distribution of values (i.e., the variance among the values of the response at different levels) rather than in testing the differences of values between particular levels. Choosing a random effect is essentially trading the ability to test hypotheses about particular levels (low vs. high nitrogen, 2001 vs. 2002 vs. 2003) for the ability to (1) quantify the variance among levels (variability among sites, among species, etc.) and (2) to generalize to levels that were not measured in your experiment. (If you treat "species" as a fixed effect, you can't say anything about an unmeasured species; if you use it as a random effect, then you can make a guess that an unmeasured species will have a value equal to the population mean estimated from the species you did measure.) Of course, as with all statistical generalization, your levels (e.g. years) must be chosen in some way that, if not random, is at least *representative* of the population you want to generalize to.

You will also hear people say that "random effects are used to represent factors that you aren't interested in". This is not always true. While it is often the case in ecological experiments (ecologists usually don't care much about the variance among sites in experiments – it is just a nuisance), it is sometimes of great interest, for example in evolutionary studies where the variation among genotypes is the raw material for natural selection, or in demographic studies

where among-year variation can have significant impacts on long-term growth rates.

You will also hear that "you can't say anything about the value of a level of a random effect". This is not true either – it may be impossible to formally test the null hypothesis that the value is equal to zero, or that the values of two different levels are equal, but it is still perfectly sensible to look at the predicted value, and even to compute a confidence interval around the predicted value. Particularly in management contexts, researchers may care very much about *which* sites are particularly good or bad relative to the population average, and how good or bad they are.

Bayesians are much more relaxed about these philosophical and inferential issues, because the Bayesian framework makes these particular problems much easier. To a Bayesian, a fixed effect is one where we estimate each parameter (e.g. the mean for each species within a genus) independently (with independently specified priors, while for a random effect the individual parameters are modeled as being drawn from a distribution (usually Normal); in standard statistical notation, species_mean ~ Normal(genus_mean, $\sigma^2_{\text{genus/species}}$).

I said above that random effects are most useful when the grouping variable has many measured levels. Conversely, random effects are generally ineffective when the grouping variable has too few levels. Random effects will be usually be infeasible when the grouping variable has fewer than 5 levels and unstable with fewer than 8; you are essentially trying to estimate a variance from a very small sample. In the classic ANOVA approach, where all of the variance estimates are derived from simple sums-of-squares calculations, random effects calculations can work as long as you have at least two samples (although their power will be very low, and in some circumstances the variance estimates are negative); in the modern mixed modeling approach, you tend to get warnings and errors from the software instead, or estimates of zero variance, but in any case the results will be unreliable (see SECTION for more details, and solutions to this problem).

3.1.2 Scalar vs. non-scalar and intercept vs. non-intercept random effects

Up to now I have described random effects in terms of the differences in the baseline values of the response variable among levels of a categorical grouping variable (e.g. baseline numbers of ticks in different sites). Although technically sites is the *grouping variable* in this case, and the thing that varies among levels is the intercept term of a statistical model, we often call this simply a "random effect of site". This is a *random intercept* model, which is also a *scalar* effect (there is only one value per level of the grouping variable). In R it would typically be specified within a modeling formula as `~group` or `~(1):group` (`MCMCglmm` package), `~1|group` (`nlme` or `glmmADMB` packages), or `(1|group)` (`lme4` or `glmmADMB` packages) depending on the modeling

function (the `1` specifies an intercept effect; it is implicit in the first example). More generally, we might have observed the effects of a treatment within each level, and want to know how the effect of the treatment (described by either a categorical or a continuous predictor) varies across levels. Since both the intercept and all of the parameters describing the treatment would vary across levels, this would be a *non-scalar* or *vector* random effect. This could be specified in R as `~1+x|group` (nlme/glmmADMB), `(1+x|group)` `(lme4/glmmADMB)`, or `~us(1+x):group` (MCMCglmm). (In many cases the `1` is optional – `(x|group)` would also work – but I include it here for concreteness. The `us` in the third specification refers to an "unstructured" variance-covariance matrix, i.e. allowing the different effects to be correlated with each other.) For example, the coral symbiont data follow a randomized block design, with replicates of all treatments within each block, so we could in principle use `(1+ttt|block)` to ask how the effects of symbionts varied among different blocks, with four random parameters (intercept and three treatment parameters) per block describing the difference between the effects of symbionts in that block and the overall population average effects. (Although such a model is theoretically OK because all treatments are performed in each block, in practice it's not feasible because we have too little information – only two binary samples per treatment per block.)

Such effects can be classified as *interactions* between the random effect of block and the fixed effect (symbionts), and are themselves random – we assume, for example, that the difference in predation rate between corals with and without symbionts is drawn from a distribution of predation rates. The interaction between a random effect and a continuous predictor would also be random, and describes the variation in slopes among levels; this type of interaction is the only case in which it makes sense to consider a random effect of a continuous variable. One should in general consider the random × fixed effect interactions whenever it is feasible, i.e. for *all* treatments that are applied within levels of a random effect; doing otherwise assumes *a priori* that there is no variation among groups in the treatment effect, which is rarely warranted biologically (Schielzeth and Forstmeier 2009, Barr et al. 2013). It is often impossible or logistically infeasible to apply treatments within groups: in the gopher tortoise example the prevalence of disease is fundamentally a site-level variable, and can't vary within sites. Or, as in the coral symbiont example, we may have so little statistical power to quantify the among-group variation that our models don't work, or that we estimate the variation as exactly zero. In these cases we have to accept that there probably is a real interaction that we are ignoring, and temper our conclusions accordingly.

3.1.3 Nesting and crossing

What about the interaction between two random effects? Here we have to specify whether the two effects are *nested* or *crossed*. If at least one of the levels

of each effect is represented in multiple levels of the other effect, then the random effects are crossed; otherwise, one is nested in the other. For example, in the gopher tortoise example, each site is measured in multiple years, and multiple sites are measured in each year, so site and year are crossed (although as pointed out above we don't actually have data for enough years to treat them as random): this would be specified for example as `(1|site)+(1|year)`. On the other hand, in the tick example each chick occurs in exactly one brood, and each brood occurs in exactly one site (`(1|site/brood/chick)`, read as "chick nested within brood nested within site", or equivalently `(1|site) + (1|site:brood)+(1|site:brood:chick)`; if the broods and chicks are uniquely labeled, so that the nesting can be detected `(1|site) + (1|brood)+(1|chick)` will also work). Another way of thinking about the problem is that, in the gopher tortoise example, there is variation among sites that is similar across years, variation among years that applies across all sites, and variation among site-by-year combinations. In the tick example, there is variation among broods and variation among chicks within broods, but there is no sensible way to define variation among chicks *across* broods. In this sense a nested model is a special case of crossed random effects that sets one of the variance terms to zero.

Crossed random effects are more challenging computationally than nested effects (they are largely outside the scope of classical ANOVAs), and so this distinction is often ignored in older textbooks. Most of the software that can handle both crossed and nested random effects can automatically detect when a nested model is appropriate, provided that the levels of the nested factor are uniquely labeled. That is, if you have individuals numbered 1, 2, … 10 in species A, B, and C, the software can't tell that individual #1 of species A is not in some way similar to individual #1 of species B. Although you can specify nesting explicitly, it is safer to label the nested individuals uniquely as A1, A2, …, A10, B1, B2, … B10, ... etc..

Interactions between two or more fixed effects are usually best treated as crossed, because in general the levels of fixed effects are generalizable across levels of other fixed effects ("high nitrogen" means the same thing whether we are in a low- or high-phosphorus treatment). Random effects can be nested in fixed effects, but fixed effects would only be nested in random effects if we really wanted (e.g.) to estimate different effects of nitrogen in each plot.

3.1.4 Overdispersion and observation-level random effects

Linear mixed models assume the observations to be Normally distributed conditional on the fixed-effect parameters and the conditional modes and thus need to estimate the residual variance at the level of observations. Most GLMMs, in contrast, assume binomial or Poisson distributions where the variance ("dispersion") parameter is fixed to 1 – that is, if we know the mean then we

assume we know the variance (equal to the mean for Poisson distributions, or to $Np(1-p)$ for binomial distributions). However, as discussed in the GLM chapter, we frequently observe *overdispersion* – variances higher than would be predicted from the model, due to missing covariates, or among-individual heterogeneity. (However, note that overdispersion is not identifiable with binary responses, as long as each observation has a unique set of predictor values.) You can allow for overdispersion in GLMMs in some of the same ways as in regular GLMs – use quasi-likelihood estimation to inflate the size of the confidence intervals appropriately, or use an overdispersed distribution such as a negative binomial – but these options are not always available in pre-packaged GLMM software.

A GLMM-specific solution to overdispersion is to add *observation-level* random effects, i.e. to add a new grouping variable with a separate level for every observation in the data set. While this may seem like magic – how can we estimate a separate parameter for every observation in the data set? – it is essentially just a way to add more variance to the data distribution. For Poisson distributions, the resulting *lognormal-Poisson* distribution is quite similar to a negative binomial distribution (also called a *Gamma-Poisson* distribution because it represents a Poisson-distributed variable with underlying Gamma-distributed heterogeneity). Most GLMM packages allow observation-level random effects (for technical reasons, MCMCglmm *always* adds an observation-level random effect to the model). Another advantage of using observation-level random effects is that this variability is directly comparable to the among-group variation in the model; Elston *et al.* (2001), the source of the tick data example, exploit this principle (see also Agresti 2002, section 13.5).

3.1.5 Correlation within groups (R-side effects)

As described above, grouping structure induces a correlation $\rho = \sqrt{\sigma^2_{within} / \sigma^2_{total}}$ between every pair of observations within a group. Observations can also be differentially correlated within groups; that is, an observation can be strongly correlated with some of the observations in its group, but more weakly correlated with other observations in its group. These effects are sometimes called *R-side effects* because they enter the model in terms of correlations of residuals (in contrast to correlations that occur because of group membership, which are called *G-side effects*). The key feature of R-side effects is that the correlation between pairs of observations within a group typically decreases with increasing distance between observations. As well as physical distance in space or time, one can also consider genetic relatedness (distance along the branches of a pedigree or phylogeny) as a distance. To include R-side effects in a model, one typically needs to specify both the distance between any two observations (or some sort of coordinates – observation time, spatial location, or position on a phylogeny – from which distance can be computed), as well as a

model for the rate at which correlation decreases with distance. While incorporating R-side effects in *linear* mixed models is relatively straightforward, putting them into GLMMs is, alas, rather challenging at present (see the "Challenges" section at the end of the chapter).

3.1.5 Fixed effects and families

Of course, for a complete model you need to specify the *fixed effects* part of your model, and the family (distribution and link function) as well as the random effects. These are both specified in the usual way as for standard (non-mixed, fixed-effect-only) GLMs.

Depending on the package you are using, the fixed effects may be specified separately or in the same formula as the random effects; typically the fixed-effect formula is also where you specify the response variable as well (the model has only one response variable, which is predicted by both the fixed and the random effects. In the coral symbiont example, the fixed effect is the categorical treatment variable (control/shrimp/crabs/both). In the gopher tortoise example we have the effects of both disease prevalence and, because we didn't have enough parameters to treat it as random, of year (treated as a categorical variable); we also have an offset term that specifies that the number of shells is proportional to the site area (i.e., we add a log(area) term to the predicted number). Finally, in the grouse tick example we have fixed effects of year and height.

TABLE 2. model specifications for the examples.

	`nlme/glmmADMB`	`lme4/glmmADMB`	`MCMCglmm`
coral symbiont	`fixed=pred~ttt, random=~1|block, family="binomial"`	`formula=pred~ttt+(1|block), family="binomial"`	`fixed=pred~ttt, random=~block, family="categoric`
gopher tortoise	`fixed=shells~factor(year) prev+offset(log(area)), random=~1|Site, family="poisson"`	`formula=shells~factor(year)+ prev+offset(log(Area))+ (1|Site), family="poisson"`	`fixed=shells~facto prev+offset(log(A random=~Site, family="poisson"`
grouse tick	`fixed= ticks~1+factor(year)+height, random=~(1|location/brood/index), family="poisson"`	`formula= ticks~1+factor(year)+height+ (1|location/brood/index), family="poisson"`	`fixed= ticks~1+factor(yea random=~location+ family="poisson"`

3.2 Conditional, marginal, and restricted likelihood

Once you have defined your GLMM, specifying (1) the conditional distribution of the data (family) and link function (see Chapter GLM); (2) the categorical and continuous predictors and their interactions (see Chapter GLM); and (3) the random effects and their pattern of crossing and nesting, you are ready to try to fit the model. Chapter XXX describes the process of maximum likelihood estimation, which we need to extend here to allow for random effects.

3.2.1 Conditional likelihood

If we (magically) knew the values of the conditional modes of the random effects for each level (e.g. the baseline predation rates for each block), we could use standard numerical procedures to find the maximum likelihood estimates for the fixed effect parameters, and all of the associated things we'd like to know (confidence intervals, AIC values, *p*-values for hypothesis tests against null hypotheses that parameters or combinations of parameters were equal to zero …). The likelihood we obtain this way is called a *conditional likelihood*, because it depends (is conditioned on) a particular set of values of the conditional modes. If x is an observation, β is a vector of one or more fixed effects, and u is a conditional mode of a random effect, then the conditional likelihood for x would be expressed as $L(x|\beta,u)$. If u were a regular fixed effect parameter, then we could go ahead and find the values of β and u that jointly gave the maximum likelihood, but that would ignore the fact that the conditional modes are random variables that are drawn from a distribution.

3.2.2 Marginal likelihood

The *marginal* likelihood is the modified form of the likelihood that allows for the randomness of the conditional modes. It essentially compromises between the goodness of fit of the conditional modes to their overall distribution and the goodness of fit of the data within grouping variable levels. For example, an observation of attack on a coral which was well-defended and would be typically expected to have a low attack probability could be explained either by saying that the coral was an unlucky individual within its (perfectly typical) block or that the coral was no unluckier than average but the block was unusual, i.e. subject to higher-than-average attack rates. Because the block effect is treated as a random variable, in order to get the likelihood we have to average the likelihood over *all possible values* of the block effect, weighted by their probabilities of being drawn from the Normal distribution of blocks. The result is called the marginal likelihood, and we can treat it in most respects the same way we would handle an ordinary likelihood. In mathematical terms, this average is expressed as an integral. If we take the definitions of x (observation), u (conditional mode), and β (fixed effect parameter) given above and additionally define σ^2 as the among-group variance (i.e. the variance of the distribution of the u values, which are defined to have zero mean), then the likelihood of a given value of u is $L(u|\sigma^2)$ and the marginal likelihood of x is the integral of the conditional likelihood weighted by the likelihood of u:
$L(x \mid \beta,\sigma^2) = \int L(x \mid u,\beta) \cdot L(u \mid \sigma^2)\, du$. The marginal likelihood is a function of β and σ^2, which are the parameters we want to estimate. (In a more complex model, σ^2 would be replaced by a vector of parameters, representing the variances of all of the random effects and the covariances among them.)

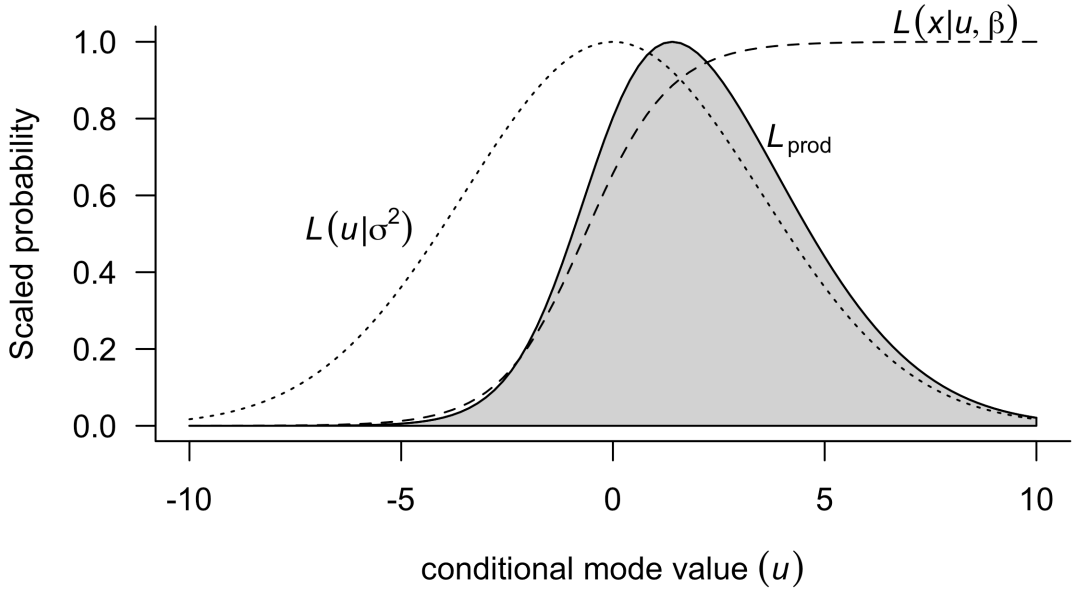

Figure 2. Conditional and marginal likelihoods. For block 5, "shrimp" treatment, replicate 2. The Normal curve (dotted line) shows the likelihood of the conditional mode *u*; the logistic curve (dashed line) shows the conditional likelihood of the observation *x* given *u*; the solid line shows their product, and the gray area under the curve represents the marginal likelihood. (All likelihoods scaled to a maximum of 1.0 for ease of presentation.) If the focal observation were the only one in the block, the conditional mode would be estimated at the peak of L_{prod}, $\hat{u}_5$=1.4. However, the contribution of the other 7 observations in the block makes the overall estimate of the conditional mode $\hat{u}_5$=-0.43.

3.3.3 Restricted likelihood

One of the often-ignored properties of maximum likelihood estimates is that many of their useful properties like lack of bias, maximal power, and so forth, only hold asymptotically – that is, when the data set is large. Maximum likelihood estimates of variances are biased downward: you may remember that the formula for estimating sample variance is $\sum(x-\bar{x})^2/(n-1)$, rather than $\sum(x-\bar{x})^2/n$ (which is the maximum likelihood estimate), for exactly this reason. *Restricted maximum likelihood* (REML) is a generalization of this rule that generally gives better (less biased) estimates of the variances in mixed models. Technically, it is based on finding some way to combine the observations so as to factor out the fixed effects. For example, in a pairwise *t* test the difference between the two observations in a pair is equal to the difference between treatments, which is the fixed effect. We are usually interested in the difference between the treatments, so we compute the difference between treatments in each pair. If we instead took the mean of each pair, we would

cancel out the fixed effect – we could then compute an unbiased estimate of the variance among the pairs. A broader way of thinking about REML is that it describes any statistical method where we integrate over the fixed effects when estimating the variances. One issue with using REML is that you cannot sensibly compare the restricted likelihoods of two models with different sets of fixed effects, because they are likelihoods of completely different models for the variance. Another is that while REML in principle applies to GLMMs as well as LMMs, they are more easily defined and more accessible in software for LMMs than for GLMMs (Bellio and Brazzale 2011).

4. Setting up a GLMM

Now that most of the concepts and terminology we'll need are defined, we can consider the basic components of a GLMM. This section discusses more of the practical considerations you need to think about when deciding on the structure of a GLMM.

4.1 Response distribution

The conditional distribution of the response variable, which we often abbreviate to "the response distribution" or "the distribution of the data", is the expected distribution of each observed response around its predicted mean, given the values of all of the fixed and random effects for that observation. That is, when we collect a data set of (e.g.) counts, we don't expect the overall (marginal) shape of the data to be Poisson distributed; we expect each point to be drawn from a Poisson distribution with its own mean that depends on the predictors for that point. In the gopher tortoise example, the distribution of number of shells in a given site s (with prevalence $P(s)$) and year y is $x_{sy} \sim \text{Poisson}(\beta_0+\beta_y+\beta_P P(s)+u_s)$.

If the conditional distribution is Gaussian, or can sensibly be transformed to be Gaussian (e.g. by log transformation) then we have a *linear* mixed model, and several aspects of the modeling process are simpler (we can more easily define R-side effects and restricted maximum likelihood; statistical tests are easier, as discussed below). As with generalized linear models (chapter XXX), binomial (including binary or Bernoulli, i.e. 0/1 responses) and Poisson responses comprise the vast majority of GLMMs. The Gamma distribution is the other common distribution handled by GL(M)Ms; it is useful for continuous, skewed distributions, but treating such data as log-normal (i.e. log-transforming and then using a linear mixed model) usually gives very similar results and is often simpler.

In addition to these standard distributions, there are other useful distributions that do not technically fall within the scope of GLMMs, but can sometimes be handled using simple extensions. These include the negative binomial distribution for overdispersed count data; zero-inflated distributions for count data with excess

zeros; the Beta distribution for proportional data where the denominator is unknown; and the Tweedie distribution for continuous data with a spike at zero. Ordinal responses (i.e. categorical responses that have more than two ordered categories) and multinomial responses (categorical responses with more than two categories, but without ordering) can be handled by extensions of binomial GLMMs. These extensions are often useful, but using them will generally make it harder to analyze your model (i.e. you are more likely to run into computational difficulties, which will manifest themselves as warnings and errors from software), and restrict your choice of software more than if you stick to the simpler (Normal, binomial, Poisson) choices of distributions.

As is typical in ecological applications, the examples for this chapter all use either binary (coral symbiont) or Poisson (gopher tortoise, grouse tick) conditional distributions (Table 2). The family is specified almost exactly as in standard GLMs, with a few quirks. `glmmADMB` requires the family argument and link functions to be given as a quoted strings (e.g. `family="binomial`), in contrast to base R, nlme, and lme4, which allow more flexibility (e.g. `family="binomial"` for the default link, or `binomial()`). MCMCglmm has its own names for binary/logit (`family="categorical"`) and binomial (`family="multinomial2"`) models.

4.2 Link function

As with GLMs, we also have to choose a link function to describe the shape of the response curve as a function of continuous predictor variables. The rules for picking a link function are the same as for GLMs: when in doubt, use the default link for the response distribution you have chosen. We will follow this rule in the examples, using the default logit link for the coral symbiont (binary) example and a log link for the gopher tortoise and grouse tick (Poisson) examples (Table 2), although we did also consider a log link for the coral symbiont example (see below). In nlme and lme4 links are specified along with the family as for standard GLMs in R, e.g. `family=binomial(link="logit")` or `binomial(link="log")`; in glmmADMB they are specified as a separate string (`link="logit"`); and MCMCglmm offers uses alternative family names where alternate links are available (e.g. `family="ordinal"` for a binary/probit link model).

4.3 Number and type of random effects

As discussed above, it is not always easy to decide which variables to treat as random effects, and to figure out their structure (nested vs crossed). Section 3.1.4 already discussed the issues of random vs. fixed and nested vs. crossed for the examples. More generally, the more random effects you include in the model, and the more they are crossed rather than nested, the harder it will be to fit the model – as with unusual response distributions (Section 4.1), the model is more likely to have computational problems, or run very slowly, or you may have trouble finding

software to fit the model. Beyond computational difficulties, the model may be statistically poorly posed – when you include several random effects, it is likely that some of their variances will be estimated as exactly zero, or that pairs of random effects will be estimated as perfectly correlated. While this does not necessarily invalidate a particular model, it may break model-fitting software in either an obvious way (errors) or a non-obvious way (the model is more likely to get stuck and give an incorrect result, without warning you). Model complexities also interact: for example, some of the software available to fit models with non-standard distributions can only handle models with a single random effect. In general you should avoid: (1) fitting random effects to categorical variables with fewer than 5 levels; (2) fitting more than two or three random effects in a single model, unless you have huge amounts of data and a very fast computer.

5. Estimation

Once the model is set up, you need to estimate the parameters – the fixed effect parameters that describe overall changes in the response, the conditional modes of the random effects that describe the predicted differences of each level of the grouping factor from the population average, and the variances of, and covariances among, the random effects. In a perfect world this would be easy, but it isn't always; there are a variety of possible methods, with tradeoffs in speed and availability.

5.1 Method of moments

The traditional way to fit a mixed ANOVA model is to compute appropriate sums of squares (e.g. the sum of squares of the deviations of the group means from the grand mean, or the deviations of observations from their individual group means) and dividing them by the appropriate degrees of freedom to obtain mean squares, which are estimates of the variances. This approach is called the "method of moments" because it relies on the correspondence between the sample moments (mean squares) and the theoretical parameters of the model (random effects variances). This approach is simple, fast, always gives an answer – and is extremely limited, applying only to Normal responses (i.e., linear mixed models), in balanced or nearly balanced designs, with nested random effects only. I mention it here for two reasons: (1) you may want to use it for simple problems that fall within its scope; (2) it is the traditional method, and it may be useful to know what more classically trained ecologists (e.g., supervisors or manuscript reviewers) have in mind if you have to discuss methodology with them.

5.2 Deterministic/frequentist algorithms

Instead of computing sums of squares, modern estimation approaches try to find efficient and accurate ways to compute the marginal likelihood (section 3.2.2). Because the marginal likelihood involves an integral that is typically at least as high-dimensional as the number of random effects in the model, computing it can be challenging. The first class of approaches for estimating mixed models involve

numerically tractable approximations of the integral. Because they try to find the exact value of the integral, I call them *deterministic* approaches; they are typically used in a frequentist statistical framework to find the maximum likelihood estimates and confidence intervals.

- *Penalized quasi-likelihood* (PQL, Breslow 2004) is the fastest, most flexible, and least accurate technique for approximating the marginal likelihood. It can handle any number of random effects quickly, and can fairly easily be extended to handle R-side effects, but it has two important limitations. (1) It is the least accurate approximation technique; in particular, it generally gives biased estimates of variance parameters, especially when the amount of information per sample is small, as with binary or low-count data. More accurate variants of PQL exist, but are not widely implemented in open source software. There is still considerable debate about the importance of these biases; for some applications, especially those focusing on the fixed effect parameters, the biases may be unimportant – but it's hard to know for sure. (2) Because of the way that PQL is derived, it computes a quantity called the "quasi-likelihood" rather than the likelihood, which may not be appropriate for model comparison by the likelihood ratio test. Depending on the software you are using, therefore, you may be limited to inference based on Wald tests (see XXX below). In a similar vein, it is not always clear what statistical model is being estimated. PQL is closely related to *generalized estimating equations* (GEE), another general statistical approach that is even more flexible but shares at least disadvantage #2 and possibly disadvantage #1.
- *Laplace approximation* is a somewhat slower and less flexible, but still very general, procedure for approximating the marginal likelihood. It approximates the integral based on the curvature of the likelihood around the conditional mode (i.e., using the Taylor expansion of L_{prod} around u=0).
- *Gauss-Hermite quadrature* (GHQ) is an extension of Laplace approximation that uses multiple points to integrate the marginal likelihood. One generally has to specify how many quadrature points to use – that is, how much computation you're willing to do for a more accurate answer. The default is usually around 8-12 quadrature points (1 quadrature point corresponds to Laplace approximation). GHQ is feasible for models with 2-3 random effects (e.g. 2-3 grouping factors with scalar random effects or a single grouping factor with 2-3 effects), but most software restricts GHQ to models with a single random effect.

5.3 Stochastic/Bayesian algorithms

Another approach to GLMM parameter estimation uses the Markov chain Monte Carlo algorithm, typically embedded in a Bayesian statistical framework that attempts to estimate the posterior distributions of the parameters rather than the maximum likelihood estimates and likelihood profiles. These algorithms are typically much slower than deterministic algorithms, and may require more tuning

of optimization parameters, although a single run of the algorithm generally gives enough information to obtain both point estimates (posterior means or medians) and confidence intervals, in contrast to deterministic algorithms where computing reliable confidence intervals may take several times longer than just finding the point (maximum likelihood) estimates.

Although there is at least one "black box" R package (`MCMCglmm`) that allows the user to define the fixed and random effects symbolically (i.e., as usual for software packages), many researchers who opt for stochastic GLMM parameter estimation instead choose to use the BUGS language (i.e. the WinBUGS package or one of its variants such as OpenBUGS or JAGS) to fit their models. BUGS is a very flexible, powerful framework for fitting ecological models to data in a Bayesian context (McCarthy 2007, Kéry 2010), not just GLMMs, but it comes with its own steep learning curve.

The Bayesian model-fitting framework also has some big advantages when it comes to computing confidence intervals that incorporate all the relevant sources of uncertainty and account properly for the size of the data set (see 6.2.5 below).

For researchers who are interested in stochastic parameter estimation but are still reluctant to use Bayesian methods, there are several stochastic parameter estimation methods that work within a frequentist framework. The older method, Monte Carlo expectation-maximization, is potentially very powerful but has not been widely implemented in general-purpose software (Booth and Hobert 1999, Sung and Geyer 2007). The newer method, *data cloning*, adapts the framework of MCMC, and particularly the BUGS package, to compute parameter estimates in a way that is consistent with frequentist theory (Ponciano et al 2009, Sólymos 2010).

5.4 Model diagnostics and troubleshooting

Much of the process of model checking for GLMMs falls back to the same procedures as for GLMs. You should plot appropriately scaled residuals (i.e., deviance or Pearson residuals) as a function of the fitted values and as a function of the input variables, looking for unexplained patterns in the mean and variance; look for outliers and/or points with large influence (leverage); and check that the distribution of the residuals is reasonably close to what was assumed. For Poisson or binomial GLMMs with $N>1$, you should compare the sum of the squared Pearson residuals to the residual degrees of freedom (number of observations minus number of fitted parameters) to check for overdispersion (unless your data are binary, or the model already contains an observation-level random effect).

The first GLMM-specific check is to see whether non-zero variances (and non-perfect correlations among random effects, i.e. $|\rho|<1$) could be estimated for all the random effects in the model. If some of the variances are zero or some correlations

are ±1, that indicates that the model is probably overfitted – not only was the among-group variation not significantly different from zero, the best estimate was zero. Although in principle the model coefficients estimated in this case will be identical to those that would have been estimated if you had just left the zero-estimate terms out of the model in the first place, it would probably be worth refitting the model without them to make sure that you haven't hit any numerical glitches. Although the most principled approach to model selection for hypothesis-testing purposes is to simply pick the largest reasonable model and stick with it, you can also to use information-theoretic approaches (AIC or BIC) to choose among possible candidate random-effects models (see section 6.2.4 below), especially if you are interested in prediction rather than hypothesis testing.

Also specific to GLMMs is examining the estimates of the conditional modes. In theory these should be Normally distributed, but you should only worry about fairly extreme deviations from Normality: it's unknown how badly a non-Normal distribution of conditional modes will bias the results of a mixed model analysis, and furthermore relaxing the assumption of Normality is difficult. You should look particularly for extreme conditional modes, and treat these as you would typically handle outliers, e.g. figure out whether there is something wrong with the data for those groups, or try fitting the model with these groups excluded and see whether the results change significantly.

For Bayesian MCMC analyses (e.g. via `MCMCglmm`), you should perform the usual diagnostics to ensure convergence and mixing – check quantitative convergence statistics such as the Gelman-Rubin $\hat{R}$ (if available) and effective sample size, and examine graphical diagnostics (trace and density plots) for both the fixed and random effects parameters. For analyses of small data sets, it is quite common for the variance-covariance parameters to mix badly, sticking close to zero much of the time and occasionally spiking near zero; the corresponding density plots typically show a spike at zero with a long tail of larger values. There are no really simple fixes for this problem, but some reasonable strategies include (1) running much longer chains; (2) adding a weakly informative prior to push the variance away from zero; (3) taking the results with a grain of salt.

As you try to troubleshoot the random effects component of your analysis, you should keep an eye on the fixed-effect estimates and confidence intervals associated with models with different random effects structures; you will often find that the fixed-effect estimates don't vary greatly among models with different random effects. This can be comforting if your main interest is in the fixed effects, although you should be careful since fitting multiple models also allows some scope for cherry-picking the results you like …

5.5 Examples

I tried fitting all three examples with lme4, glmmADMB, and MCMCglmm (using Laplace approximations for the first two). I discussed above some of the issues that arise, such as the impracticality of fitting a treatment × block interaction in the coral symbiont example, or the need to fit year as a fixed rather than a random effect in the gopher tortoise example. Other points that arose as a result of fitting and diagnosing the models:

- *Coral symbionts*: the main issue with this fit, as discussed above, is that because of the small number of points and binary data, some of the estimates can be extreme. Random-effects fitting of the blocks takes care of the extremes we saw when we tried to fit block as a fixed effect, but when we look at the Pearson residuals we see one very extreme value, an observation that we would expect to have a very high probability of predation (it's in the no-symbiont treatment in an otherwise frequently attacked block) that nevertheless escaped predation. Re-fitting the model without this observation makes all of the treatment effects much more extreme, essentially suggesting near-complete separation of the control and treatment groups once block effects are accounted for.

  This near-complete separation also leads to some issues with the MCMCglmm fit; it also estimates complete separation between the control and non-control treatments, resulting in extremely large positive estimates of predation probability for the control treatment and extremely large negative estimates for the non-control treatments. The qualitative conclusions are similar.

  Other aspects of the model look OK – for example, (1) the distribution of conditional modes is sensible, and (2) refitting with Gauss-Hermite quadrature makes very little difference to any of the estimates.

- *Gopher tortoise shells*: even after we reluctantly switched year to a fixed effect, we still find that Poisson sampling variation can account for nearly all the variation in the data – the maximum likelihood estimates of both the variance of an observation-level random effect (if included) and the among-site variance are very near zero. Thus, the conditional modes (which are scaled by the among-site standard deviation) are also all near zero. The Pearson residuals look reasonable, and are approximately equivalently distributed across sites.

  The `MCMCglmm` fit (which includes both among-site and among-observation variation) shows the pathology described above in section 5.5. We can't really simplify the model, because `MCMCglmm` requires at least one explicit random effect (so we can't remove the site effect), and always includes an observation-level random effect. Adding a weakly informative prior on the variances, on the other hand, cleans up the model nicely. It doesn't change

the point estimate of the prevalence effect much, but it does increase its uncertainty slightly.

- *Grouse ticks*:
  The Pearson residuals and estimated conditional modes all look reasonable. We don't bother to test for overdispersion since we already have observation-level random effects in the model. The deterministic/frequentist approaches (lme4 and glmmADMB) give positive estimates for all of the variance components, suggesting that the model is not overfitted, but MCMCglmm disagrees; unless we add a prior, it estimates the location variance as being near zero, suggesting that the brood vs. location variance decomposition is unstable.

6. Inference

6.1 Approximations for confidence intervals

Point estimates of parameters are useless without confidence intervals, or hypothesis tests, that inform us how much we really know about the system. Confidence intervals and hypothesis tests for GLMMs require a series of assumptions which are inherited from either GLMs or linear mixed models, and which (as with the estimation methods described above) require a series of tradeoffs between accuracy, computation time, and convenience or availability in software.

Quadratic approximations: the easiest and least accurate versions of confidence intervals and hypothesis tests assume that the log-likelihood surface has a quadratic shape, i.e. that the goodness of fit of the model as measured by the log-likelihood decreases as the square of the distance in parameter space from the best fit model. This is exactly true for linear models, but only approximately true for GLMs and GLMMs, and the approximation gets worse the smaller the effective size of the data set.

Finite-size corrections: Even when using a procedure such as likelihood profiling (see below) that doesn't make assumptions about the shape of the likelihood surface, we still need to some assumptions about the distribution of the maximum log-likelihood under the null hypothesis in order to determine appropriate *p*-values or critical values for the confidence intervals.

- For response distributions such as the Normal or Gamma with a freely varying scale parameter (in contrast to the Poisson and binomial, which assume the variance is a fixed, known function of the mean), there is uncertainty in the estimate of the scale parameter. In the case of a Normal response (i.e. LMMs) this uncertainty causes the sampling distribution of individual parameters to be *t* rather than Normal, and the sampling difference in log-likelihood between nested models when the simpler one is

correct to be proportional to an F rather than to a χ^2 distribution; these are the standard distributions used to construct confidence intervals and test hypotheses in standard linear models as well. The degrees of freedom parameter for the t, or the denominator degrees of freedom for the F, is a measure of the effective size of the data set: the number of observations, counted at an appropriate level of replication, minus the number of parameters estimated. The additional complexity that comes with LMMs is that these distributions are only *approximate* as soon as we depart from simple classical (balanced, nested, no R-side effects) experimental designs, and that the approximate degrees of freedom for the t distribution, or the denominator degrees of freedom for the F, can be difficult to compute. If your experimental/observational design is nested and balanced, you can either use a software package that computes the denominator degrees of freedom for you or look the experimental design up in a standard textbook (e.g. Gotelli and Ellison 2004 or Quinn and Keough 2002). If it is not, then you will need to rely on a computational approximation such as the Kenward-Roger correction (Kenward and Roger 1997, Højsgaard 2013), or use a resampling-based approach (see below).

- For GLMMs (i.e. non-Normal response distributions), whether or not the scale parameter is fixed (e.g. for binomial or Poisson distributions as well as for the Gamma discussed in the previous point), there is an additional component of approximation and uncertainty that arises because the null sampling distribution of parameter estimates or log-likelihood differences is only *approximately* Z- or χ^2-distributed. This component of uncertainty can usually be neglected if the effective sample size is greater than 40-50, and is almost always neglected in standard GLM analyses. In GLMMs, however, it is (1) more likely that the effective sample size (e.g. the number of blocks in a nested design), will be small and (2) likely that a reader will be coming from the world of LMMs, where researchers spend a lot of time worrying about effective sample size, rather than from the world of GLMs, where they routinely ignore them. *Bartlett corrections* (McCullagh and Nelder 1989) are one approach to finding adjustments to the null statistics that account for finite size, but they are not widely implemented; for reliable finite-size corrections you may need to rely on resampling (see below).

Boundary effects: another problematic feature of (G)LMM models is that the null-hypothesis values of variance parameters lie on the boundary of their allowable space, which causes technical difficulties with the statistical theory used to derive null distributions (Pinheiro and Bates 2000). That is, the null hypothesis in tests of random effects is that the variances are zero, but if they're not zero they must be positive rather than negative. In the simplest case of testing whether a single random-effect variance is zero, the p-value derived from standard theory is twice as large as it should be, leading to a conservative test (you're more likely to conclude that you can't reject the null hypothesis). In the simplest cases you can fix the problem by simply dividing the p-value by 2, but for more complex cases the

simplest approaches (other than ignoring the problems) involve simulating the null hypothesis.

6.2 Solutions

6.2.1 Wald tests

As discussed in chapter XXX , Wald tests, and the corresponding Wald confidence intervals, assume that the log-likelihood surface is quadratic, and so are subject to artifacts when assessing GLM parameters – for example when a binomial model has extreme parameter estimates ($|\beta|>10$) because some combination of treatments in the data gives rise to observations that are all zeros or all ones (*complete separation*). However, they are quick to compute and can be useful for a rapid assessment of parameter uncertainty. If you can guess the appropriate residual degrees of freedom, then you may try to use appropriate *t* statistics rather than *Z* statistics for the *p*-values and confidence interval widths in order to account for finite sample sizes, but be aware that this is a very crude approximation in the case of GLMMs.

6.2.2 Likelihood ratio tests

Using the actual shape of the log-likelihood surface rather than assuming that it is quadratic improves the accuracy of confidence intervals and *p*-values considerably. When comparing nested models to get *p*-values, this is fairly straightforward; you just fit the full model and the reduced model and compare the difference in the difference in log-likelihood. According to likelihood theory, in order to reject the null hypothesis that the simpler model is a sufficient description of the system (i.e., that the parameters you added to the model aren't making the fit of the data any better than would be expected by chance even if their true values were zero), you need to show that the difference in deviance (-2 times the log-likelihood) is larger than a critical value based on a χ^2 distribution.

To use the likelihood ratio test to find confidence intervals for parameters, or regions for combinations of parameters, we have to find the *profile likelihood* – that is, the best likelihood that can be achieved for each value of a focal parameter by optimizing over all of the other (non-focal) parameters, and then finding the values of the focal parameter for which the profile likelihood is greater than the χ^2-based cutoff described above. Computing profile likelihoods is straightforward in principle, but computationally much more challenging – depending on the number of parameters in the original model it can take tens or hundreds of times as long to compute the profile confidence intervals as to find the maximum likelihood estimates in the first place. Of the off-the-shelf GLMM approaches, only lme4 has built-in profiling. Furthermore, because profile likelihood calculations intentionally try to evaluate the likelihood for extreme parameter values, they are much more subject to computational warnings and errors than the original model fit.

Finally, although likelihood-based comparisons are more reliable than curvature-based (Wald) comparisons, they still fail to account for the non-χ^2 sampling distribution of the likelihood; that is, they assume infinite residual or "denominator" degrees of freedom. If your effective sample size is large enough (e.g., the number of levels in the smallest grouping factor is >40), then you don't need to worry about this: otherwise, if you want accurate confidence intervals and *p*-values you may need to use a stochastic resampling method such as parametric bootstrapping or Markov chain Monte Carlo.

6.2.3 Bootstrapping

Bootstrapping usually refers to resampling data with replacement to get a new pseudo-data set. *Parametric* bootstrapping (PB) refers to simulating from the fitted model. If you want to test the significance of certain parameters in a model (equivalent to doing the likelihood ratio test between full and reduced models, but allowing for finite sample size), you fit the reduced model to the data; simulate pseudo-data from it many (say 1000) times; fit the both the reduced and full model to each set of pseudo-data, and calculate the difference in the log-likelihood in each case. This is the null distribution of the log-likelihood difference between the reduced and full model. The proportion of time that these null values are greater than or equal to the observed difference in log-likelihood between the full and reduced models (i.e., for the real data) is the *p*-value.

PB can also be used to compute the confidence intervals of the parameters for a single model, by simulating data from the *same* (full) model 1000 times and computing the quantiles of the distributions of each of the parameters.

PB is generally quite slow (it will take almost 1000 times as long as fitting the original model), and it is not perfect – the second approach in particular makes the assumption that the estimated parameters are close to the true parameters – but it is essentially the best way we know to compute *p*-values and confidence intervals for GLMMs.

Some specialized methods of parametric bootstrapping exist: for example, the `RLRsim` package in R (Scheipl et al. 2008) does a form of null-hypothesis simulation/parametric bootstrapping to compute *p* values for random effects terms in LMMs, in a way that is orders of magnitude faster than standard PB.

You can also try *nonparametric* bootstrapping, but you must do it in a way that respects the grouping structure of the data. For example, for a model with a single grouping variable you might do multi-stage bootstrapping where you first sample with replacement from the levels of the grouping variable, then sample with replacement from the observations within each sampled group. For more complex

models (with crossed random effects, or R-side effects), appropriate sampling may be difficult.

6.2.5 MCMC

There's not nearly enough room in this chapter to give a proper explanation of Markov chain Monte Carlo; for now, you can just think of it as a computational recipe for sampling values from the posterior distribution of a model. MCMC is very general, and includes GLMMs as one special case. If you use a Bayesian software tool like WinBUGS or JAGS to set up your GLMM, then you get confidence intervals on the parameters "for free" by computing quantiles, or other kinds of Bayesian confidence interval, of the posterior sample. (While there are Bayesian definitions of *p*-values, most Bayesians don't use them to test null hypotheses.) The MCMC approach to computing confidence intervals on parameters, or on predicted values from the model, is very powerful – it automatically allows for finite size effects, and incorporates the uncertainty in all the components of the model, which is otherwise difficult. It's so powerful, in fact, that some frequentist/deterministic tools such as AD Model Builder allow the user to run a *post hoc* form of MCMC, assuming completely flat priors, to compute confidence intervals. (This sort of pseudo-Bayesian approach is often much more convenient than setting up a fully Bayesian analysis, but setting flat priors in this way is problematic when the information in the data is weak enough that moderately weak priors would have a strong effect on the results.) One challenge of MCMC, beyond the technical difficulty of setting up the model in the first place and the computational burden of the running the model (generating 1000 useful samples from the posterior distribution can take almost as long as the same number of parametric bootstrap replicates) is that for small, noisy data sets the posterior distribution of the variance parameters is often composed of a spike at zero along with a second component with a mode away from zero. In this case, most standard MCMC algorithms have a tendency to get stuck sampling either the spike or the non-zero component, and thus give poor results.

6.2.4 Information-theoretic approaches

In addition to the classical frequentist and Bayesian inferential frameworks, many ecological researchers use information-theoretic approaches to select models and generate parameter importance weights or weighted multimodel averages of parameters and predictions (Burnham and Anderson 2002, chap ?) In principle, AIC or other information criteria such as BIC should apply just as well to marginal log-likelihoods as they do to standard log-likelihoods, but several of the theoretical difficulties discussed above affect information criteria as well as classical frequentist tests (Greven and Kneib 2010, Müller et al. 2013).

- Parameters whose maximum likelihood values are on the boundary (e.g. variances that are estimated as zero) give similar problems to those encountered in frequentist hypothesis testing.

- Counting the number of parameters that should be associated with a random effect is tricky. If you are using AIC to compare models that differ only in their fixed effects, then it doesn't matter how many parameters you assign to a random effect, since only the difference in the number of parameters matters. However, if you are trying to decide whether to incorporate a random effect in the model, then you do have to address this issue. It turns out that the answer depends on whether you are trying to make predictions at the population level (i.e., predicting the average value of a response from individuals across all random effects levels, or predicting the response from an individual from a previously unmeasured random effects level) or at the individual level (i.e., for individuals within a specific level). In the former case (*marginal prediction*), you should count one parameter for each random effects variance-covariance parameter. In the latter case (*conditional prediction*), the correct answer is somewhere between 1 and *n*-1, where *n* is the number of random effects levels: there are recipes for computing the relevant value (Vaida and Blanchard 2005), although they are not as widely implemented as they might be. In my experience, academic ecologists are more generally interested in marginal prediction (they want to know what the effects are at the whole population), which allows them to use the easy one-parameter-per-variance rule; applied ecologists might be more likely to want conditional predictions for specific groups. If you are using Bayesian MCMC to fit your models there is an analogous metric called the *deviance information criterion* (DIC: Spiegelhalter et al. 2002), which has a similar issue in that the so-called "level of focus" must be defined explicitly.

- If you are trying to use an information-theoretic score that includes a finite-size correction term, such as AICc, you need to decide on what to include as the number of observations (e.g. for a nested design is it the number of individual observations, or the number of groups?) as well as the effective number of parameters; this is analogous to the "denominator degrees of freedom" issue discussed above, and you can probably use the same solutions (e.g. assign one degree of freedom for a scalar random effect if you are interested in population-level estimation), but be aware that the AICc has really not been tested in the context of GLMMs.

In general it is best practice to *pick one method of modeling and inference in advance*, or after brief exploration of the feasibility of different approaches for a specific problem, in order to avoid the ever-present temptation of cherry-picking the best results.

6.3 Examples

For completeness, I tried to compare a variety of different inference methods for each example (i.e. Wald, profile, parametric bootstrap, Bayesian credible interval). I

also give an example of how I might report the results in each case. You should in general report something about the among-group variation, whether it is of primary interest or not; if it is not, don't report *p*-values. Whether you report among-group variation as standard deviation or variance depends on the goals of your analysis. If you want to partition variance across levels, then report among-group variances; otherwise (and probably more generally), report among-group standard deviations, as these are expressed in the same units as the corresponding fixed effects.

In some places below I quote results (estimates, confidence intervals, *p*-values) from several different methods, for comparison purposes only. As recommended above, you should choose one method and stick to it in any given analysis.

FIGURE: TO DO: *add multi-panel figure comparing point estimates and CIs for fixed effects and random effects variances for different approaches for each example …*

- *Coral symbionts*: in the original paper (McKeon et al. 2012), we used a log link function to test the null hypothesis that the effect of multiple defenders on predation probability was independent and proportional (i.e., that if crab-protected and shrimp-protected corals were attacked with probability p_c and p_s respectively, that doubly protected corals would be attacked with probability $p_c \cdot p_s$), by quantifying a two-way interaction between crab and shrimp presence. We failed to reject that null hypothesis: "the best estimate of the [multiple defender effect] on frequency of predation was only a 6% reduction, but the confidence interval was wide (51% reduction to 90% increase)." (The paper also reported effects on volume removed when predation did occur, which was analyzed with a LMM and did show significant effects.) We did not report the size of the block effect, but we should have. For the analysis done here (logit link, one-way comparison of crab/shrimp/both to control) I would quote either the fixed-effect parameter estimates (clarifying to the reader that these are differences between treatments and the baseline control treatment, on the logit or log-odds scale), or the changes in predation probability from one group to another. For example: "Crab and shrimp treatments had similar effects (-3.8 log-odds decrease in predation probability for crab, -4.4 for shrimp); the dual-symbiont treatment had an even larger effect (-5.5 units), but although the presence of any symbiont caused a significant drop in predation probability relative to the control, none of the symbiont treatments differed significantly from each other (likelihood ratio test p=0.27, parametric bootstrap test (N=220) p=0.23). The among-block standard deviation in log-odds of predation was 3.4, nearly as large as the symbiont effect." Alternately, one could quote the predicted predation probabilities for each group, which might be more understandable for an ecological audience.

- *Gopher tortoise*: The main point of interest here is the effect of prevalence on the (per-area) density of fresh shells. This makes reporting easy, since we can focus on the estimated effect of prevalence. Because the model is fitted

on a log scale and the parameter estimate is small, it can be interpreted as a proportional effect. For example: "A 1% increase in seroprevalence was associated with an approximately 2.1% increase (log effect estimate=0.021) in the density of fresh shells (CI 0.013-0.29 [Wald]; 0.012-0.29 [likelihood profile]; 0.013-0.031 [parametric bootstrap=PB]). Both of the years subsequent to 2004 had lower shell densities (log-difference =-0.64 (2005), -0.43 (2006)), but the differences were not statistically significant (95% PB CI: 2005={-1.34,0.05}, 2006={-1.04,0.18}). There was no detectable overdispersion (Pearson squared residuals/residual df=0.85; estimated variance of an among-observation random effect was zero). The best estimate of among-site standard deviation was zero, indicating no discernable variation among sites, with a 95% PB CI of {0,0.38}."

- *Grouse ticks*: In this case the random effects variation is the primary focus, and we report the among-group variance rather than standard deviation because we are interested in variance partitioning. "Approximately equal amounts of variability occurred at the among-individual, among-brood, and among-location levels (glmer/Laplace: σ^2_{ind}=0.29, σ^2_{brood}= 0.56, σ^2_{loc}=0.28; glmmADMB/Laplace, 95% Wald CI: σ^2_{ind}=0.31 [0.06-0.42], σ^2_{brood}= 0.48 [0.18-0.84], σ^2_{loc}=0.13 [0.14-0.47]; MCMCglmm (default priors), 95% credible intervals: σ^2_{ind}=0.31 [0.02-0.44], σ^2_{brood}= 0.88 [0.52-1.30], σ^2_{loc}=0.26 [0-0.39]; MCMCglmm (stronger priors), 95% credible intervals: σ^2_{ind}=0.31 [0.2-0.43], σ^2_{brood}= 0.59 [0.36-0.93], σ^2_{loc}=0.57 [0.29-1.0]). The among-brood variance is estimated to be approximately twice the among-individual and among-location variances, but there is considerable uncertainty in the brood/individual variance ratio (`MCMCglmm`: $\sigma^2_{\text{brood}}/\sigma^2_{\text{ind}}$=2.01 [95% CI 1.007-3.37]), and the among-location variance is somewhat unstable. There are also strong effects of year and altitude (glmer/Laplace, 95% Wald CI). In 1996, tick density increased by a factor of 3.3 relative to 1995 (1.18 [0.72,1.6] log units); in 1997 density decreased by 38% (-0.98 [-1.49,-0.46] log units) relative to 1995. Tick density increased by approximately 2% per meter above sea level (-0.024 [-0.03,-0.017] log-units), decreasing by half for every 30 (log(2)/0.024) m of altitude."

7 Conclusions

I hope you are convinced by now that GLMMs are a widely useful tool for the statistical exploration of ecological data. Once you get your head around the multi-faceted concept of random effects, you can see how handy it is to have a modeling framework that naturally combines flexibility in the response distribution (GLMs) with the ability to handle data with a variety of sampling units with uneven and sometimes small sample sizes (mixed models).

GLMMs cannot do everything; especially for very small data sets, they may be overkill (Murtaugh 2007). Ecologists will probably always be faced with data sets

which are too small to fit as sophisticated a model as they would like, as in the first two examples in this chapter (coral symbionts and gopher tortoises), but one can often find a sensible middle ground.

In this chapter I have slightly neglected the other end of the spectrum, very large data sets. Ecologists dealing with Big Data from remote sensing, telemetry, or citizen science, may have tens or hundreds of thousands of observations rather than the dozens to hundreds represented in the examples here (although telemetry data often contains huge amounts of detail about a very small number of individuals; in this case a fixed-effect or two-stage (Murtaugh 2007) model may work as well as a GLMM). The good news is that some of the computational techniques described here scale well to very large data sets. In addition, finite-size corrections, and the associated computationally intensive recipes such as parametric bootstrapping, become essentially irrelevant when all the grouping variables have more than 50 levels.

Also neglected here has been a variety of useful GLM extensions: offsets (for managing data sampled over different temporal and spatial extents); non-standard link functions (for fitting specific nonlinear models such as the Beverton-Holt or Ricker functions); methods for handling multinomial or ordinal data; and zero-inflation. The good news is that most of these tricks are at least in principle extendable to GLMMs, but your choice of software may be more limited (see e.g. Bolker et al 2013 and the associated web resources).

Unfortunately, GLMMs do come with considerable terminological, philosophical and technical baggage, which I have tried to clarify as much as possible. As GLMM software, and computational power, continues to improve, many of the technical difficulties will fade, and GLMMs will continue their growth in popularity; a firm grasp of the *conceptual* basis of GLMMs will be an increasingly important part of the quantitative ecologist's toolbox.

9 Neglected

- What to do when you can't or don't want to do a mixed model: two-stage models (approx of variance); analyze residuals and hope for non-significant/small block effects; computing $\sum(\beta - \overline{\beta})^2 / n$ as an approximation of variance when there are too few RE levels
- GEEs; marginal vs conditional slopes of effects
- Regularization/priors

References

Florin Vaida and Suzette Blanchard. Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2):351-370, June 2005.

N. E. Breslow. Whither PQL? In Danyu Y. Lin and P. J. Heagerty, editors, *Proceedings of the second Seattle symposium in biostatistics: Analysis of correlated data*, pages 1-22. Springer, 2004.

José C. Pinheiro and Douglas M. Bates. *Mixed-effects models in S and S-PLUS*. Springer, New York, 2000.

D. A. Elston, R. Moss, T. Boulinier, C. Arrowsmith, and X. Lambin. Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology*, 122(5):563-569, 2001.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

Arpat Ozgul, Madan K Oli, Benjamin M Bolker, and Carolina Perez-Heydrich. Upper respiratory tract disease, force of infection, and effects on survival of gopher tortoises. *Ecological Applications*, 19(3):786-798, April 2009. PMID: 19425439.

Andrew Gelman. Analysis of variance: why it is more important than ever. *Annals of Statistics*, 33(1):1-53, 2005.

M. G Kenward and J. H Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983-997, 1997.

Sonja Greven and Thomas Kneib. On the behaviour of marginal and conditional Akaike information criteria in linear mixed models. *Biometrika*, 97(4):773-789, 2010.

D. J. Spiegelhalter, N. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64:583-640, 2002.

Yun Ju Sung. Monte carlo likelihood inference for missing data models. *The Annals of Statistics*, 35(3):990-1011, July 2007.

James G. Booth and James P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo EM algorithm. *Journal of the Royal Statistical Society. Series B*, 61(1):265-285, 1999.

Nicholas J. Gotelli and Aaron M. Ellison. *A Primer of Ecological Statistics*. Sinauer, Sunderland, MA, 2004.

Michael J. Crawley. *Statistical Computing: An Introduction to Data Analysis using S-PLUS*. John Wiley & Sons, 2002.

Paul A Murtaugh. Simplicity and complexity in ecological data analysis. *Ecology*, 88(1):56-62, 2007.

José Miguel Ponciano, Mark L. Taper, Brian Dennis, and Subhash R. Lele. Hierarchical models in ecology: Confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology*, 90(2):356-362, February 2009.

C. Seabird McKeon, Adrian Stier, Shelby McIlroy, and Benjamin Bolker. Multiple defender effects: synergistic coral defense by mutualist crustaceans. *Oecologia*, 169(4):1095-1103, 2012.

Gerry P. Quinn and Michael J. Keough. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, England, 2002.

Holger Schielzeth and Wolfgang Forstmeier. Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, 20(2):416-420, March 2009.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255-278, April 2013

Alan Agresti. *Categorical Data Analysis*. Wiley, Hoboken, NJ, 2d edition, 2002.

Ruggero Bellio and Alessandra R. Brazzale. Restricted likelihood inference for generalized linear mixed models. *Statistics and Computing*, 21(2):173-183, April 2011.

M. McCarthy. *Bayesian methods for ecology*. Cambridge University Press, Cambridge, England, 2007.
Marc Kéry. *Introduction to WinBUGS for ecologists Bayesian approach to regression, ANOVA, mixed models and related analyses*. Elsevier, Amsterdam; Boston, 2010.

Péter Sólymos. dclone: Data cloning in R. *The R Journal*, 2(2):29-37, 2010.

Ulrich Halekoh Søren Højsgaard <sorenh@math.aau.dk>. *pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison*, 2013. R package version 0.3-7.

Fabian Scheipl, Sonja Greven, and Helmut Kuechenhoff. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52(7):3283-3299, 2008.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference*. Springer, New York, 2d edition, 2002.

Samuel Müller, J. L. Scealy, and A. H. Welsh. Model selection in linear mixed models. *Statistical Science*, 28(2):135-167, May 2013.