Ann. Zool. Fennici 46: 124–137 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 30 April 2009

How to make models add up — a primer on GLMMs

Robert B. O'Hara

Department of Mathematics and Statistics, P.O. Box 68, FI-00014 University of Helsinki, Finland (e-mail: bob.ohara@helsinki.fi)

Received 18 Feb. 2008, revised version received 3 July 2008, accepted 15 July 2008

O'Hara, R. B. 2009: How to make models add up — a primer on GLMMs. — *Ann. Zool. Fennici* 46: 124–137.

Many problems in the analysis of ecological data have the format where there is an observed response that may be predicted by several covariates. Although the response can take several forms (e.g. measurements, counts, observations of presence/absence), and the covariates can also vary (e.g. be measurements themselves, or be grouped according to the treatment applied, the time or location of of sampling, etc.), most of these problems can be handled in a single framework, the Generalized Linear Mixed Model (GLMM). The framework encompasses regression, ANOVA, generalized linear models, and equivalent models with random as well as fixed effects. Here, the different parts of the GLMM are described, building from regression and ANOVA to show how the extra components — the wider range of distributions, and random effects — can be added into the same framework, and how the parameters of the fitted model can be estimated and interpreted. Being able to handle data with GLMMs helps ecologists to analyse the majority of their data.

Introduction

After all the data have been collected for a study (Underwood 2009), they have to be decomposed into something simpler that can be understood and communicated. Doing this is the *raison d'être* of statistics, and the methods used vary from plotting and tabulating the data to fitting complex models, from which statistics can be extracted to answer the questions we are asking.

Many data analysis problems in biology boil down to explaining a response variable by several explanatory variables. Because of the ubiquity of these sorts of problem, there is a set of very general models that can be used to investigate them. These models assume that the predictors have linear effects, i.e. their effects can simply be added together. This means that the models for the means of the data are relatively simple — the complexities are in the random component (i.e. the distribution of the data given the predictors' effects), and in the model fitting. These can, to a large extent, be swept under the table when doing any actual analysis, and instead the focus can be kept on the way the models are built and interpreted.

My purpose here is to review these models, and to outline what they look like and how they can be used. A full explanation of their use can (and does) take up many volumes, so this should be seen as a primer for these models rather than a full explanation. The intent is to help the reader to understand what their final model actually says, rather than to explain how to build the model, and test between different alternative models that may fit the data. Throughout, examples will be used from data analysed by Kotze *et al.* (2003), on abundances and range sizes in Danish carabid beetles. The analyses here are used as examples of the different types of analysis, rather than to provide definitive answers to the questions posed, and hence are sometimes simplified.

Linear models: regression

We begin with a simple linear regression. The purpose of a regression is to fit a model for how a response (y) can be predicted from a covariate (x). The model is:

$$y_i \sim N(\mu_i, \sigma^2) \quad (1)$$
$$\mu_i = \beta_0 + \beta_1 x_i \quad (2)$$

and is shown for some fake data in Fig. 1. Equation 1 describes the random part of the model: y_i (the data) is normally distributed (the $N()$) notation with mean μ_i and variance σ^2. The systematic part of the model is described by Eq. 2. The mean is a function of the covariate (x_i), and two parameters, the intercept (β_0) and slope (β_1): these give a straight line, shown as the dotted line in Fig. 1. This second part of the model is deterministic, so we have separated the model into the stochastic (Eq. 1) and deterministic (Eq. 2) parts.

More generally, if there are several covariates, then multiple regression can be used. If there are s covariates, then Eq. 2 is extended:

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_s x_{is} = \sum_{j=1}^{s} \beta_j x_{ij} \quad (3)$$

the x_{ij}'s are covariates, and the β_j's are the slopes associated with the covariates. It is thus just the sum of the product of the covariates and their regression parameters. In two dimensions this would form a plane, and in more dimensions a hyper-plane. The relationships are therefore all based on straight lines. Critical here is the assumption that the relationship between x_i and μ_i is a straight line. In reality few things are linear, so this may appear to be an arbitrary assumption made for statistical convenience. Indeed in many ways it is, although it often works well in practice. But there is also a justification for using the straight line as an approximation (e.g. Venables 2000). If, in general, we have a relationship $\mu_i = f(x_i)$, where $f()$ is some suitably well behaved function (i.e. a smooth curve that relates x_i to μ_i), then we can write this approximately as

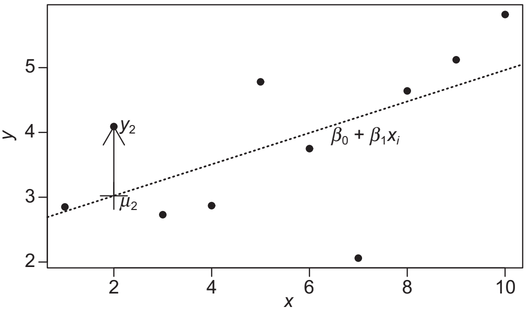

Fig. 1. A regression slope applied to fake data.

$$\mu_i = \beta_0^* + \beta_1^*(x_i - x_0) + \beta_2^*(x_i - x_0)^2 + \ldots + \beta_n^*(x_i - x_0)^n \quad (4)$$

which is called an nth order Taylor series expansion around x_0. $f()$ is being approximated by a polynomial curve, and the higher n, the closer the polynomial to $f()$. This approximation is centred at x_0, which might be the mean of x, but could also be some other sensible value: this is equivalent to placing the origin at x_0. The linear regression (Eq. 2) corresponds to the first order expansion:

$$\mu_i = \beta_0^* + \beta_1^*(x_i - x_0) \quad (5)$$

so we have $\beta_0 = \beta_0^* + \beta_1^* x_0$ and $\beta_1 = \beta_1^*$. If we felt that this was not good enough (e.g. if we thought that the relationship was curved), then we could fit a second order polynomial:

$$\mu_i = \beta_0^* + \beta_1^*(x_i - x_0) + \beta_2^*(x_i - x_0)^2 \quad (6)$$

and obviously higher orders as well. By expanding the parentheses and rearranging the terms, we see that $\beta_0 = \beta_0^* - \beta_1^* x_0 + \beta_2^* x_0^2$ and $\beta_1 = \beta_1^* - 2\beta_2^* x_0$. In other words, the curve is the same whether we place the origin at 0 or x_0, it is just being written in a different way.

The values of β_0^* and β_1^* vary with x_0, and in particular whether they are zero depends on where the origin is placed. So a test of whether the intercept or linear slope is zero depends on the value of x_0. In this sense they are arbitrary:

for example, the results of such a test for a temperature scale might depend on whether it was measured in Celsius or Kelvin. The practical consequence is the advice that, for any model, if a higher order term is included (e.g. a quadratic term), the lower order terms (e.g. the intercept and linear terms) must also be in the model. This is known as the "principle of marginality".

Example

Kotze *et al*. (2003) were interested in explaining the relationship between abundance (defined as the average number of records per grid cell) and range size (defined as the number grids where each species had been observed) in species of Danish carabid beetles. The simplest model for this would be to let y_i be the abundance and x_i be the range size of the *i*th species. Fitting this model to the data gives us the following equation:

$$\mu_i = 0.22 + 0.00087x_i \qquad (7)$$

and then $y_i \sim N(\mu_i, 0.030)$. So, an increase of one grid cell in range size would increase the (log) abundance by 8.7×10^{-4}. This may sound insignificant, but the size of the effect will depend on the variables measured. In this case, the abundance has a standard deviation of 0.19, and the standard deviation of the range size is 77, so y_i has a relatively small range, and x_i has a large range. We can get some feel for the size of the effect of range size by comparing the effect of changing the range by one standard deviation to the standard deviation of the abundance. This is simply $77 \times 8.7 \times 10^{-4} = 0.067$. This is about one third of the standard deviation of the abundance, so the change is not huge. If we look at the R^2 for the model (i.e. the proportion of the variance explained by the covariate), we find that it is fairly small at 13%.

The size of the beetles may also affect the abundance. If we call the size (the length in mm) of the *i*th species z_i, and fit the model with range size and body size, we get the following equation:

$$\mu_i = 0.22 + 0.00087x_i - 0.00012z_i. \qquad (8)$$

The coefficient for the effect of body size is of the same order of magnitude as the range size coefficient. But the standard deviation of body size is about 5 mm, which is much smaller. A change in 1 SD in body size would decrease the log(abundance) by 0.00059, or the actual abundance to 0.9994 times of what it was, which is of no practical significance. The R^2 of the model is 13%, so there is no increase in the amount of variation explained.

The question of whether the sizes of the effects of the variables are meaningful has been approached here in an unusual way. R^2 is a useful summary of the explanatory power of the model, but of the full model rather than specific components. Calculating the change in the response when a covariate is increased by one standard deviation is a simple (if crude) device for getting some feel for the size of an effect. It may not always work, for example if covariates are correlated, so that a change in the covariate of interest would be related to changes in other covariates, then the effect of the change in the ensemble could be very different.

The discussion above has focussed on the interpretation of the coefficients, and has ignored issues of the precision of these estimates. In the second model, the standard error for the range size coefficient is 1.4×10^{-4}, and for the body size effect is 2.1×10^{-3}. The body size coefficient is therefore well within the range that might be expected if the effect was actually zero (the *t* statistic is –0.057, which gives a *p* value of 0.96). In contrast, if the range size coefficient were zero, it would be very unlikely that a value this large would have been observed ($p << 10^{-5}$). Although *p* values are often quoted, all they show is whether a statistic of that magnitude is likely if the true value were zero. They say nothing about the size of the effect (Läärä 2009) — it may be statistically significant, but does an R^2 below 13% mean that it is practically significant? That is a question of biology, not statistics.

Linear models: ANOVA

Another major type of analysis is the Analysis of Variance (ANOVA). In ANOVA, the covariates are discrete factors. Starting with a simple balanced one-way ANOVA, with a groups, and n observations in each group, the model is

$$y_{ij} \sim N(\mu_{ij}, \sigma^2) \tag{9}$$
$$\mu_{ij} = \alpha_j \tag{10}$$

where $i = 1, ..., n$ and $j = 1, ..., a$. Hence, every individual in the same group has the same expected value. This is shown in Fig. 2. For a balanced two-way ANOVA, the model for the mean is this:

$$\mu_{ijk} = \alpha_j + \gamma_k. \tag{11}$$

But there is a problem here: an arbitrary value can be added to every value of α_j, and subtracted from every value of γ_k, and still give the same value of μ_{ijk}. The problem is a general one which can be found in any complex design. The solution is to constrain the parameters in some fashion. This can be done in many ways, the one that is most common in computer packages is the cornerpoint constraint. This has the model

$$\mu_{ijk} = \beta_0 + \alpha_j + \gamma_k \tag{12}$$

with the constraint $\alpha_1 = \gamma_1 = 0$, so that observations with $j = 1$ and $k = 1$ have $\mu_{i11} = \beta_0$, i.e. this is the intercept. Observations with $j = 1$ have $\mu_{i1k} = \beta_0 + \gamma_k$, so γ_k is the difference between the kth level and the first level of γ. The same pattern holds for α. The α's and γ's are therefore the differences between the mean of the level and the mean of the intercept. This would be particularly useful if the intercept were a control treatment, and the other levels were experimental manipulations.

An alternative constraint is to mean-centre the parameters, so the model is the same as in Eq. 12, but the constraints are $\sum\alpha_j = 0$ and $\sum\gamma_k = 0$. The $\sum\alpha$'s and $\sum\gamma$'s are therefore the differences from the mean. Technically they are called contrasts (or at least one type of contrast). Whilst these contrasts have an intuitive interpretation, they are less useful in practice because the grand mean of the data will depend on how the observations have been allocated to the different levels — allocate more to a level with a high effect, and the mean increases. It is therefore harder to make comparisons between studies.

Table 1. Dummy variables for a model with a single factor with three levels.

Level	β_0	β_2	β_3	μ_{ij}
1	1	0	0	β_0
2	1	1	0	$\beta_0 + \beta_2$
3	1	0	1	$\beta_0 + \beta_3$

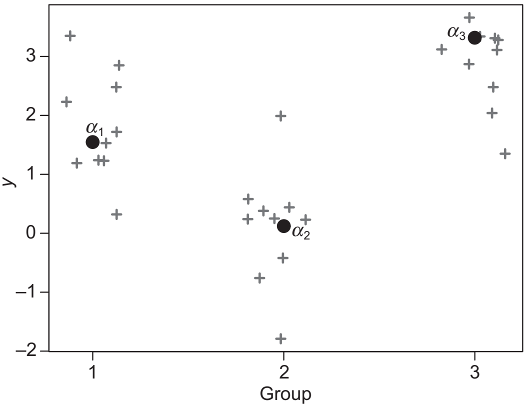

Fig. 2. An ANOVA applied to fake data.

Although it may not be immediately obvious, ANOVA is actually a regression. This can be seen by setting up the ANOVA using dummy variables (which is also how it is done by the computer). For example, consider a one-way classification with 3 levels. The variables can be set up using a corner-point constraint (Table 1). If an observation is at, for example, level 2 of a factor, then for that parameter the dummy variable is set to 1, and is set to 0 for all other levels of the factor. The regression is then done on the dummy variables. Whilst this may appear obscure, it has two consequences. The first is that it unifies the two approaches, so that we can talk about regression and ANOVA (and ANCOVA too!) as being linear models: the distinction between them disappears, and we can focus on how to build the models, rather than learning each type of model as a separate entity. The second consequence is that the models can be fitted to the data in the same way, so we only need a single set of tools to do this. Indeed, this means that computer packages can provide a single command to fit the models.

The models that are fitted can be more complex than outlined above. In particular, variables are allowed to interact. In the two-way ANOVA above, the effect of the second factor (γ_k) is the same regardless of the value of j. This is not always realistic: sometimes we might expect the value to differ (e.g. the effect of habitat preference on abundance may depend on whether a species is a specialist or generalist). In this case, we can simply add more variables, so that the model is

$$\mu_{ijk} = \beta_0 + \alpha_j + \gamma_k + \phi_{jk}. \tag{13}$$

This can be coded using dummy variables (Table 2). Now the differences in the second factor depend on the value of the first factor. The model can be coded in the same way as the simpler model, by making the parameters either contrasts to the mean or to an intercept level of the factors.

Regression and ANOVA are both linear models, so when both discrete factors and continuous covariates are being used, it is natural to combine them in a single model. This can be viewed as a regression with several intercepts (i.e. each combination of levels of the factors is a different intercept), and with different regression slopes for levels of the factors. The same machinery can then be used to fit the model.

Writing a linear model

Using equations to write a linear model can become complex when there are many factors and covariates. A simpler notation was developed by Wilkinson and Rogers (1973), and forms the basis of the way they are written in statistics packages.

An explanatory variable (whether continuous or a discrete factor) can be represented by its name, e.g. Date, Length. A simple model can then be written like this:

$$\text{Weight} \sim \text{Date} + \text{Length}$$

so Weight is the response, and Date and Length are the explanatory variables. Here they appear as separate main effects, with no interaction. The interaction can be written using the 'dot operator', for example Date.Length. When factors in a model are crossed, the interaction will have the main effects in it too, so a model might be Date + Length + Date.Length. Rather than write this out in full, it can be simplified by writing it as Date * Length. The * then says that this should be expanded out as the main effect, and lower order interactions. A term like Date * Length * Height would then be

Date + Length + Height + Date.Length
+ Date.Height + Length.Height
+ Date.Length.Height.

We will also need to denote nested terms. For example, if several samples are taken from and individual, then Sample would be nested within Individual. There would be no point in having a Sample main effect, as they vary between individuals — sample 4 in individual 2 has no relation to sample 4 in individual 7, so a sample 4 parameter is not needed. We therefore only have the Individual main effect and the Individual.Sample interaction. This can be represented conveniently as Individual/Sample. Further nesting can also be done, e.g. Population/Individual/Sample, and could also be crossed with other effects, e.g. Date * Population/Individual/Sample.

Table 2. Dummy variables for a model with two factor with three and two levels, and their interaction.

Level	β_0	β_2	β_3	γ_2	ϕ_{21}	ϕ_{31}	ϕ_{12}	ϕ_{22}	ϕ_{32}	μ_{ij}
1,1	1	0	0	0	0	0	0	0	0	β_0
2,1	1	1	0	0	1	0	0	0	0	$\beta_0 + \beta_2 + \phi_{21}$
3,1	1	0	1	0	0	1	0	0	0	$\beta_0 + \beta_3 + \phi_{31}$
1,2	1	0	0	1	0	0	1	0	0	$\beta_0 + \gamma_2$
2,2	1	1	0	1	0	0	0	1	0	$\beta_0 + \beta_2 + \gamma_2 + \phi_{22}$
3,3	1	0	1	1	0	0	0	0	1	$\beta_0 + \beta_3 + \gamma_2 + \phi_{32}$

Example

Using the same data as before, we can first look at a model to see if the type of wing has an effect on abundance. The Wing variable is a factor with three levels, macropterous (i.e. long wings), brachypterous (only short wings), and dimorphic (both long and short wings). The model is

$$\text{Abundance} \sim \text{Wings},$$

and the estimates are given in Table 3. The model is set up using a corner-point contrast, with brachypterous species as the intercept, so that the estimated difference in abundance between these and the the dimorphic species is 0.092. The estimated abundance of dimorphic species would therefore be 0.28 + 0.092 = 0.37 (or $e^{0.37}$= 1.45 individuals/grid cell).

We might want to add the effects of whether a species in Denmark is at the edge of its species range, so the factor has two levels, Yes and No. Including the interaction, the model is

$$\text{Abundance} \sim \text{Wings} * \text{Edge}.$$

The estimated parameters are shown in Table 4. The intercept is for brachypterous species not at the edge of their range. The "main effects" are contrasts to this, so for example the difference between brachypterous and macropterous species not at the edge of their range is 0.026. Similarly the difference between brachypterous species not at the edge of their range and those at the edge is –0.031, i.e. those at the edge have a slightly smaller estimated abundance (although not significantly smaller than might be expected if there was no actual effect). The interactions are slightly more complicated, but again are just sums of the estimates. For example, for macropterous species at the edge of their range, the estimated (log) abundance is 0.30 + 0.026 – 0.031 – 0.25 = 0.045. It is clear that being macropterous at the edge of the range means that the species' abundance is reduced (at least in Denmark).

The analyses above can be combined to demonstrate models with both factors and covariates. A simple model might be

$$\text{Abundance} \sim \text{Range Size} + \text{Wings},$$

where the effect of range size would be the same for each wing type: essentially this would be fitting a different intercept for each wing type. A more complex model would be to allow the slope to vary between different wing morphologies, i.e. the model

$$\text{Abundance} \sim \text{Range Size} * \text{Wings}.$$

The estimates from this model are shown in Table 5. The Range Size effect is for brachypterous species, so the effect for dimorphic species, for example, is $8.7 \times 10^{-4} + 7.7 \times 10^{-5} = 9.5 \times 10^{-4}$.

Table 3. Estimates of parameters from the model Abundance ~ Wings.

Coefficient	Estimate	SE	*t*	Pr(> I*t*I)
(Intercept)	0.28	0.013	22.3	$< 10^{-5}$
Wings, macropterous	0.0011	0.030	0.037	0.97
Wings, dimorphic	0.092	0.037	2.48	0.014

Table 4. Estimates of parameters from the model Abundance ~ Wings * Edge.

Coefficient	Estimate	SE	*t*	Pr(> I*t*I)
(Intercept)	0.30	0.016	19.0	$< 10^{-5}$
Wings, macropterous	0.026	0.033	0.79	0.43
Wings, dimorphic	0.087	0.041	2.139	0.033
EdgeYes	–0.031	0.027	–1.17	0.24
WingsM.EdgeYes	–0.25	0.085	–2.9	0.0039
WingsD.EdgeYes	–0.0032	0.094	–0.034	0.97

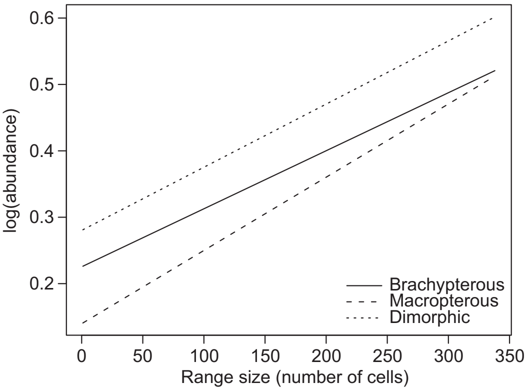

Fig. 3. Fitted regression lines for effects of range size and wing morphology on the abundance of Danish beetles.

The slopes are perhaps best understood by examining them in a plot (Fig. 3). For example, the line for dimorphic species is

$$\mu_i = 0.23 + 0.055 + (8.7 + 0.77) \times 10^{-4} x_i. \quad (14)$$

All of the lines increase, but the increase for brachypterous species is slower.

We can see from these examples that the estimated effects of different predictors are easy to calculate, as they are just sums of different terms. The difficulty at this point comes from extracting the correct terms from the output of the analysis, i.e. understanding what the statistical software has outputted.

Generalized Linear Models: building on distributions

The linear model outlined above is the basis for many analyses. However, it has the shortcoming that it assumes that the variance is constant, and that the data are not constrained. This is not always the case. For example, in the example above the number of records of each species in each grid cell might be better modelled as a count. These must obviously be an integer, and cannot be negative.

A variety of methods were developed to deal with non-normal data (McCullagh & Nelder 1989: pp. 8–17). Eventually it was realised that these all have a common structure, and could be fitted to data using a single approach (iterated weighted least squares, Nelder & Wedderburn 1972).

Taking the carabid data as an example, the linear model described above would generally not be appropriate for the number of records (unless the counts were large), because it assumes that any value is possible. Another problem is that the variance of count data generally increases with the mean — an increase of 10 individuals is more likely when the mean is 1000 than when the mean is 3. A classical solution to this problem is to log-transform the data first. In effect, this makes the model multiplicative (i.e. additive on the log scale), so the covariates say whether the counts double (for example). This solution runs into a problem when there are zero counts, as the log of this is $-\infty$. An *ad hoc* solution is to add 1 to every count, but there is no reason why 1, as opposed to 0.5 or 27, should be added.

There is an alternative approach, which we can explain mechanistically. Count data can generally be viewed as being generated from something like a Poisson process. We can imagine that there is a constant rate at which beetles are observed and recorded. If the rate is λ, and the records are over a time t, then the mean number of individuals recorded is λt. We can actually go

Table 5. Estimates of parameters from the model Abundance ~ Range Size * Wings.

Coefficient	Estimate	SE	*t*	Pr(> |*t*|)
(Intercept)	0.23	1.7×10^{-2}	13.2	$< 10^{-5}$
Range Size (cells^{-1})	8.7×10^{-4}	1.7×10^{-4}	5.1	$< 10^{-5}$
WingsM	-8.6×10^{-2}	5.0×10^{-2}	–1.7	0.09
WingsD	5.5×10^{-2}	6.2×10^{-2}	0.88	0.38
Range Size:WingsM (cells^{-1})	2.3×10^{-4}	3.4×10^{-4}	0.67	0.51
Range Size:WingsD (cells^{-1})	7.7×10^{-5}	5.2×10^{-4}	0.15	0.88

further, and say that if the beetles behave independently, then n, the number of beetles caught is Poisson distributed, with mean λt; i.e.

$$\Pr(n = r) = \frac{(\lambda t)^r e^{-\lambda t}}{r!} . \qquad (15)$$

Now, imagine that there are two similar species. For each individual of either species, assume there is a (small) probability p that it is recorded in a unit time interval. Also assume that there are n_1 and n_2 individuals of each species. The expected number of records of species 1 would be ptn_1, and ptn_2 for species 2. Assuming independence between individuals would then imply that the actual number would be Poisson distributed. On the log scale the expected numbers are $\log(p) + \log(t) + \log(n_1)$, and $\log(p) + \log(t) + \log(n_2)$. In other words, the model is linear on the log scale, and the difference between the rate of capture is simply $\log(n_2) - \log(n_1)$. The linear models described above could therefore be used for the log of the mean number of individuals. Notice that here it is the expected value that is log transformed, not the data.

Generalized linear models (GLMs) build on this approach. In general, they can be written as

$$y_i \sim \mathrm{Dist}(\mu_i,\phi) \qquad (16)$$
$$g(\mu_i) = \eta_i \qquad (17)$$
$$\eta_i = \sum_{j=1}^{s} \beta_j x_{ij} \qquad (18)$$

where Dist() is a distribution (technically, this has to be from the exponential family of distributions), with mean μ_i and the variance depending on ϕ. Sometimes, for example for the Poisson distribution, ϕ will be a constant, at other times it will have to be estimated (e.g. for the normal distribution, $\phi = \sigma^2$).

Table 6. Examples of common generalized linear models.

Names	Distribution	Link function
Regression, ANOVA	Normal	Identity
Logistic Regression	Binomial	logit
Probit Analysis	Binomial	probit
Log-linear model	Poisson	log
Ordinal Regression	Multinomial	logit

Equation 18 is the linear part of the model — this is just the same as Eq. 3. So, the covariates affect η_i, rather than μ_i, in a linear way. Equation 18 is therefore the systematic part of the model, and Eq. 16 is the random part. Between these we have Eq. 17, which links the two together. This is done using $g()$, which is a function. For the Poisson distribution, this would be the log function, i.e. $\log(\mu_i) = \eta_i$. For the normal distribution, $g()$ is the identity function, i.e. the trivial $\mu_i = \eta_i$.

Generalized linear models come in several forms, and many common analyses are GLMs, or extensions of them. A list of some common examples is given in Table 6.

Example

If we model the records using a Poisson distribution (we will ignore the problem that there are no zeroes in the data), it is clear that the more grid cells a species is observed in, the more records there are likely to be. Indeed, a naïve expectation is that they should be proportional. If r_i is the number of records of the ith species, with λ_i the expected number, and x_i is again the range size (i.e. the number of grids), then we would expect $\lambda_i = Cx_i$, where C is some constant (possible depending on other predictors). On the log scale this is $\log(\lambda_i) = \log(C) + \log(x_i)$, i.e. a regression equation with a coefficient for Range Size of 1. We can fit the following model (from Eqs. 16–18):

$$r_i \sim \mathrm{Poisson}(\lambda_i) \qquad (19)$$
$$\log(\lambda_i) = \eta_i \qquad (20)$$
$$\eta_i = \beta_0 + \beta_1 x_i \qquad (21)$$

so that we would expect that $\beta_1 = 1$. If we fit the model we find that the coefficient is 1.057, which is close to 1. However the standard error is 0.0091, so the estimate is about 6 standard deviations away from 1, and hence is statistically significantly different from zero. Is it practically significant? To get some idea about this, imagine doubling the number of grid cells, say from 10 to 20. The difference in the expected log number of records would be 1.057 × [log(20) – log(10)] = 1.057 × log(2). The number of records would therefore be expected to increase by a factor of

$2^{1.057} = 2.08$, or about 8% more than would have been expected. This increase is perhaps moderate, although I should demur in this assessment to any carabidologist.

Why make things so complicated? The reason is generality — many models can be written in this way. Generalized linear models therefore provide a single platform for the analysis of many data sets, integrating many types of analysis together. For example, we may sample 100 beetles, and ask how many of those are wingless. If we think that each beetle has the same probability of having wings, and that they are all independent, then the number that are observed to have wings follows a binomial distribution:

$$\Pr\left(n = r|N\right) = \left[\frac{N!}{r!\left(N-r\right)!}\right]p^r\left(1-p\right)^{N-r} \quad (22)$$

After some manipulation we find that this is a generalized linear model with a logit link function, i.e.

$$\eta_i = \mathrm{Bin}(p_i, N^{-1}), \quad (23)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i, \quad (24)$$

$$\eta_i = \sum_{j=1}^{s}\beta_j x_{ij}. \quad (25)$$

The model has the same general structure as Eqs. 19–21, and can be fitted in the same way. A problem is the interpretation of the parameters — for the Poisson example they are relatively simple (being multiplicative), the logit scale is trickier. To help, notice that $e^{\eta_i} = p_i/\left(1-p_i\right)$, and that $O_i = p_i/(1 - p_i)$ is the odds of and event. That is, for every "failure", when the event does not happen, there are on average O_i successes when it does. The logit is then the log of the odds. Although slightly more obscure, using the log scale has the advantages that (1) the scale goes between $-\infty$ and $+\infty$ (i.e. it is not constrained), and (2) it is symmetrical in the following sense. If we have a probability p of a success, then the probability of a failure is $1 - p$. The odds for a success are $o = p/(1 - p)$ and $1/o$. On the log scale, this is $\log(o)$ and $-\log(o)$. In other words, to flip the event of interest from "success" to "failure", we just reflect the log odds around zero. Or, less mathematically, we just change all the signs of the terms.

Example

Continuing with the beetles, one problem is whether we can infer which traits affect whether a species is declining. Here we use a simplification of the analysis carried out by Kotze and O'Hara (2003). For our purposes, we use a simple binary true/false factor denote whether the species is declining or not. We then ask whether this depends on the size of the beetles, or how specialised they are. Specialisation has been coded as a factor with five levels: level 1 means an extreme specialist, level 5 an extreme generalist.

The model is the same as Eqs. 23–25, and the linear part can be written as

$$\text{Decline} \sim \text{Size} + \text{Specialisation}.$$

After fitting the model, the estimates are obtained (Table 7). Specialisation class 1 is used as the intercept, and the others are contrasts to that. So, for example, the species *Carabus problematicus* is in specialisation class 4, and is 24.5 mm long. The probability that it would be declining can be calculated from

Table 7. Estimates of parameters from the model Decline ~ Size + Specialisation.

Coefficient	Estimate	SE	*t*	Pr(> |*t*|)
(Intercept)	–1.35	0.35	–3.9	10^{-4}
Specialisation 2	0.074	0.34	0.22	0.83
Specialisation 3	–0.63	0.37	–1.7	0.09
Specialisation 4	–2.52	0.67	–3.7	0.0002
Specialisation 5	–17.3	726	–0.024	0.98
Size (mm^{-1})	0.13	0.03	4.14	< 10^{-5}

$$\log[p_i/(1-p_i)] = -1.35 - 2.52 + 0.13 \times 24.5 = -0.61, \quad (26)$$

so the probability is (solving for p_i) $e^{-0.61}/(1 + e^{-0.61}) = 0.35$. In reality, *C. problematicus* is declining. We can look at the effects of it being, say, 1mm smaller. It is simplest to do this directly on the logit scale. The effect is to change the log odds of declining by 0.13 × (–1 mm) = –0.13. Note that this is the same, regardless of the size, or the specialisation class. This is equivalent to a reduction in the odds of 0.88, i.e. approximately 1 – 0.13. If we were to shrink *C. problematicus* by 1 mm, the probability of declining would become 0.32, i.e. a change of 3%. In contrast, if we were to take a species of the same size but in specialisation class 1, the probability would change from 87.0% to 85.5%, i.e. a difference of 1.5%. The point here is that what is constant on the logit scale is not constant on the probability scale.

One final point can be made from this analysis. The estimate for specialisation class 5 (extreme generalist) is –17.3, with a standard error of 726. This might suggest that the effect is poorly estimated. But if we examine the data, we get a different story. None of the extreme generalists are declining. The best estimated probability of declining is therefore 0, which equates to –∞ on the logit scale. In reality, the software used for the estimation stopped at –17.3, and left the estimate as that (–17.3 is evidently almost infinite). These cases are not rare, but are also easy to spot, as they have large estimated values and even larger standard errors.

For distributions like the Poisson and binomial, the mean also determines the variance. In reality, the variance (i.e. the dispersion) is often larger than that assumed, and this is described as over-dispersion. Several methods exist for dealing with this, the most straightforward is to estimate it as a constant from the residual deviance (e.g. McCullagh & Nelder 1989). Some more specific alternatives exist, for example for the Poisson distribution one can model the over-dispersion as following a gamma distribution, in which case one can use a negative binomial distribution in place of the Poisson (e.g. Ver Hoef & Boveng 2007). An alternative is to use a quasi-likelihood approach (McCullagh & Nelder 1989: chapter 9). Only the mean and the variance of the distribution are specified, and the variance is allowed to depend on the mean. This can be particularly useful for the Poisson (Ver Hoef & Boveng 2007) and binomial distributions. Over-dispersion occurs at the level of the individual observation, but extra random effects may be found at different levels of the analysis, and these have to be modelled. Recognition of this lead to the extension of GLMs to GLMMs.

Generalised Linear Mixed Models

A general approach to adding random terms into GLMs had to wait until the technical problem of how to fit the models to data was solved, and was only accomplished in the early 1990s (Breslow & Clayton 1993). This lead to the development of Generalized Linear Mixed Models (GLMMs).

Random effects can be understood in several ways. For example, take the case where beetles have been sampled from several islands in an archipelago. There is a distribution of abundances of the beetles over the whole archipelago, and when we sample beetles from a random selection of these islands, we are sampling from this distribution. Just as a simple regression can be used to model the distribution of abundances, so the distribution of abundances on different islands can also modelled. For example, the variance of the distribution can be estimated.

Mathematically, the random effects are parameters. The interpretation of a random effect as representing the distribution of the abundances (say) is equivalent to modelling the parameters as being drawn from a distribution, in contrast to generalized linear models where the parameters are allowed to be estimated freely. In generalized linear mixed models, the assumption is that the parameters are normally distributed. This implies that a sample from (say) 10 islands will tell us something about the 11th island. This is sensible — if from the first 10 islands around 5 to 10 individuals are captured, it would seem unlikely that 250 individuals would be caught on the 11th.

It is not always obvious when precisely a factor should be set as a random effect. For example, if there are 1000 islands in the archipelago, and 10 are sampled, it makes sense to think that we are sampling from a distribution. But what if there are only 50 islands? Or 20? Or 10? Even with 10 islands, if we sample 9 of them, it may still make sense to think that these tell us something about the 10th island. This sort of argument has been used to suggest that the use of random effects could be expanded (Gelman & Hill 2007). The decision to treat a factor as random can therefore be subjective with grey areas where there is no definite right approach, something not unusual in statistical modelling.

Random effects are typically used for two reasons. Firstly, they may be nuisance variable, i.e. something that causes variation in the data but which is not of direct interest. In this case, the attitude taken towards them is that they are not of direct interest, and may not have to be reported. However, they should still be examined — for example, if their effect is small, it may be better to remove them from the model, as this can improve the estimation of the other parameters. The second reason for using random effects is that the focus is on estimating the amount of variation. A clear example of this comes from quantitative genetics, where the additive genetic variance is important (for example in estimating the speed of phenotypic change due to selection), and the estimated values for the individuals (the breeding values) are often less important. The focus is then on estimating where the variation in the data is occurring.

Now we know what a random effect is, how do we deal with it? Breslow and Clayton (1993) developed the model by adding the random effect terms to a GLM, so that Eq. 18 would become

$$\eta_i = \sum_{j=1}^{s} \beta_j x_{ij} + \sum_{k=1}^{r} Z_{ik} u_{ik} \qquad (27)$$

where the u_{ik}'s are the parameters of the random effect that are to be estimated, and the Z_{ik}'s are the levels of the random effect. The u_{ik}'s are assumed to be normally distributed, i.e. $u_{ik} \sim N(0,\sigma_k^2)$. The estimation is more complex because the variances mean that the η_i's are correlated, and the variances (and hence the correlations) have to be estimated.

It is easier to understand what is going on if we extend the notation introduced above. We can denote random effects like this:

$$(1|\text{Island})$$

(the reason for the 1 will become clear). Island is then a random effect, and the assumption is that the Island effects are normally distributed. More formally, if I_k is the effect of the kth island, then $I_k \sim N(0,\sigma_I^2)$. Obviously, we can have more than one random effect. For example, if we had several traps on each island, these might be nested, and we might arrive at a model like this:

$$(1|\text{Island/Trap}).$$

Mathematically, this means that Trap is also normally distributed.

The random effect here is, in essence, affecting the intercept of the model — it is the same for all of the levels and values of the fixed effects. This does not have to be the case, it can vary between different fixed effects. For example, if we have several species sampled, and we want to estimate separate Island variances for each species, it can be written like this:

$$(\text{Species}|\text{Island}).$$

So that $I_k(s) \sim N[0,\sigma_I^2(s)]$, i.e. the variance depends on which species, s, the observation is for. Note that here Species is a discrete factor, so for each level a different variance is estimated.

This approach gives models of great generality, but the more complex models will be harder to fit (if a separate variance is estimated for each species, then each variance is estimated on the basis of fewer data points, so is less precise), and may be harder to interpret.

Similarly, the effects of a continuous variable, such as height above sea level, might be thought to vary between islands. This would mean that the slope would be varying (just as in an ANCOVA model), but would be random. These are sometimes called random regression models. They can be written in the same way as for the factor, i.e.

$$(\text{Height}|\text{Island})$$

and this would mean a model with $\mu_i = \beta_0 + \beta_1(I_i)h_i$, and $\beta_1(I_i) \sim N(0,\sigma_\beta^2)$. A fuller model could therefore be

$$\text{Count} \sim \text{Species} + \text{Height} + (\text{Height}|\text{Island}) + (1 + \text{Species}|\text{Island/Trap})$$

so that the intercept and Species are all random effects.

Example

We can continue with the analysis of the number of records. First, we can see if the abundance varies phylogenetically. We can do this by treating genus as a factor — whilst this is not ideal, it provides an approximate proxy for phylogenetic relatedness. If we treat it as a fixed effect, we need to estimate 55 parameters. Of these, 23 are only estimated from a single species, so their estimates will be poor. Using genus as a random effect can help, because the genera which have several species will help to inform about those with less species (for a deeper discussion, *see* chapter 12 of Gelman & Hill [2007]). We can then fit the following model

$$\text{Abundance} \sim \log(\text{Range Size}) + \text{Wings} + (1|\text{Genus}).$$

When we do this, the standard deviation of the Genus effect is 0.026, which is roughly the same as –0.025, the estimated difference between brachypterous and macropterous species (brachypterous being more abundant). This suggests that the effect is about as important as the effect of wing morphology.

A more complex model would include different random effects between the different wing morphologies, i.e.

$$\text{Abundance} \sim \log(\text{Range Size}) + \text{Wings} + (1 + \text{Wings}|\text{Genus}).$$

The results are shown in Fig. 4. The grey bars are the wing morphology-specific standard deviations. As we can see, only the macropterous species have any appreciable variation between genera, and this variation is larger than the variation between wing morphologies.

It is worth emphasising that the models presented here are to demonstrate the models, rather than to be the best models for the data. In particular, there are other factors, such as habitat preferences, that will also affect the results of these analyses.

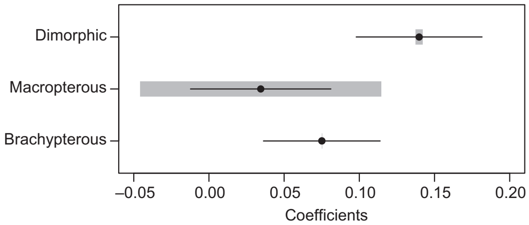

Fig. 4. Estimates of effects of wing morphology on abundance: estimate × standard error (black), standard deviation of genus effects (grey).

Extensions

Generalized linear mixed models are powerful tools, that can be used to solve many problems. Of course, they cannot be used to analyse every data set, but for some problems there are extensions to GLMMs that can be used, and where the ideas behind generalized linear mixed models are useful, both technically in developing these methods and practically in using them. Here, I will point to some of these extensions, without going into details.

One problem with linear models is that they are fairly restrictive about the shape of the effect of a covariate on the expected value of the response — i.e. it has to be a straight line. Adding polynomials can help, but unless a high-order polynomial is used, this restriction is still a problem. One way of easing this is to use GAMs, Generalized Additive Models (e.g. Wood 2006). These work by fitting smoothed curves to the data. The advantage of these models is that they provide a flexible set of curves that can be used, but there are two costs. The first is simply that more data are required in order to fit a curve reasonably well. The second is interpretability — with a straight line, it is clear if the line is increasing or decreasing. Similarly, when fitting complex models, including interactions in a GLM is straightforward, but whilst it can be done in a GAM, it requires more data, and can be difficult to visualise in more than 2 dimensions. In situations where there are data, and it is expected that a curve is not linear, they will be useful, and perhaps they are best viewed as either "black boxes" for prediction, or as providing visualisations of non-linear curves. Random effects can also be added, to form GAMMs, Generalized Additive Mixed Models (Wood 2006).

The random effects were developed above using a hierarchical description — the data are functions of parameters, which themselves are modelled by being functions of further parameters. This approach can be extended. At its most general, it leads to hierarchical models (e.g. Gelman & Hill 2007), where complex models can be built from simpler blocks. Although this approach is very powerful, beyond a certain point the model fitting becomes difficult, and it would be best to switch to a Bayesian approach to the fitting (Läärä 2009). All of the models discussed above are hierarchical models (even if the hierarchy is flat for many of them).

Spatial data can be modelled using hierarchical models which are very similar to GLMMs (e.g. Banerjee *et al*. 2004). The difference is that the spatial term is a more complex form of a random effect (the exact form will depend on the type of data).

One very powerful class of hierarchical models are HGLMs, Hierarchical Generalized Linear Models (Lee *et al*. 2006). These build on GLMM ideas, but provide more flexibility. For example, they provide an approach to modelling the variance in the same way as the mean, using a doubly hierarchical linear model.

Table 8. Common software tools for analysing classes of models. R packages for analyses given in brackets.

Models	Programme		
	R (2.6.1)	SAS (ver. 9)	Genstat (10th ed.)
Linear Models	Y	Y	Y
GLMs	Y	Y	Y
GLMMs	Y (lme4)	Y	Y
GAMs	Y (mgcv)	Y	Y
GAMMs	Y (mgcv)	N	N
HGLMs	N	N	Y

Conclusions

The models outlined above can be used to analyse many of the data sets that are generated in the ecological and evolutionary sciences. The focus here has been on describing how the models work, and how they should be interpreted. Much of the actual practice of using these models has been ignored, for reasons of simplicity and space. For example, variable selection has been ignored (e.g. Burnham & Anderson 2002), as have other matters such model checking. Many books are available that explain the processes of data analysis with the sorts of models discussed above. Gelman and Hill (2007) is one recent tome that gives a comprehensive coverage of the issues involved in fitting and interpreting GLMs, GLMMs and hierarchical models.

In practice, these models will be fitted to the data using statistical software. This again means that we can sidestep some of the complexities. A summary of what analyses can be done using which software packages is given in Table 8.

It should be clear that the models are relatively simple, as they are just made up of sums of terms. The mathematical complexities arise from the form of the distributions, and more critically, from the methods used to fit the models to the data. However, with the ready availability of statistical packages that can be used to carry out the computations, this is less of a problem. One issue is the reliability of these packages, and of the fitting methods. Perhaps the main problems are seen in the application of generalized linear mixed models to data with a binary response, where the typical methods perform poorly (e.g. Lin & Breslow 1996).

The models explained here are being regularly used by biologists. Unfortunately, because of the way statistics has been taught, many analyses are presented simply as hypothesis tests (e.g. ANOVA tables), without the model that gives the tables being examined (Läärä 2009). The models are themselves not difficult to understand (perhaps the main problem is understanding the scale on which the linear model is measured), and by understanding them we get closer to a description of the data being analysed, and hence closer to the underlying biology — the model is a description of the data, and what the organisms were doing, only in numbers rather than words. For those who find the numbers abstract, we can also draw graphs, and describe the data in pictures.

Acknowldgements

This research was supported by the Academy of Finland through grant 205371.

References

Banerjee, S., Carlin, B. P. & Gelfand, A. E. 2004: *Hierarchical modeling and analysis for spatial data*. — Chapman and Hall/CRC Press, Boca Raton, Fl., USA.

Breslow, N. E. & Clayton, D. G. 1993: Approximate inference in generalized linear mixed models. — *Journal of the American Statistical Association* 88: 9–25.

Burnham, K. P. & Anderson, D. R. 2002: *Model selection and multimodel inference. a practical information-theoretic approach*, 2nd ed. — Springer, New York, New York, USA.

Gelman, A. & Hill, J. 2007: *Data analysis using regression and multilevel/hierarchical models*. — Cambridge University Press, New York, U.S.A.

Kotze, D. J. & O'Hara, R. B. 2003: Species decline — but why? Explanations of carabid beetle (*Coleoptera, Carabidae*) declines in Europe. — *Oecologia* 135: 138–148.

Kotze, D. J., Niemelä, J., O'Hara, R. B. & Turin, H. 2003: Testing abundance–range size relationships in European carabid beetles (Coleoptera, Carabidae). — *Ecography* 26: 553–566.

Läärä, E. 2009: Statistics: reasoning on uncertainty, and the insignificance of testing null. — *Annales Zoologici Fennici* 46: 138–157.

Lin, X. & Breslow, N. E. 1996: Bias correction in generalized linear mixed models with multiple components of dispersion. — *Journal of the American Statistical Association* 91: 1007–1016.

McCullagh, P. & Nelder, J. A. 1989: *Generalized linear models*, 2nd ed. — Chapman and Hall, London.

Nelder, J. A. & Wedderburn, R. W. M. 1972: Generalized linear models. — *Journal of the Royal Statistical Society A* 135: 370–384.

Underwood, A. J. 2009: Components of design in ecological field experiments. — *Annales Zoologici Fennici* 46: 93–111.

Venables, W. N. 2000. *Exegeses on linear models*. — Available at http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf.

Ver Hoef, J. M. & Boveng P. L. 2007: Quasi-Poisson *vs*. negative binomial regression: how should we model overdispersed count data? — *Ecology* 88: 2766–2772.

Wilkinson, G. N. & Rogers, C. E. 1973: Symbolic description of factorial models for analysis of variance. — *Applied Statistics* 22: 392–399.

Wood, S. N. 2006: *Generalized additive models*. — Chapman and Hall/CRC Press, Boca Raton, Fl., USA.