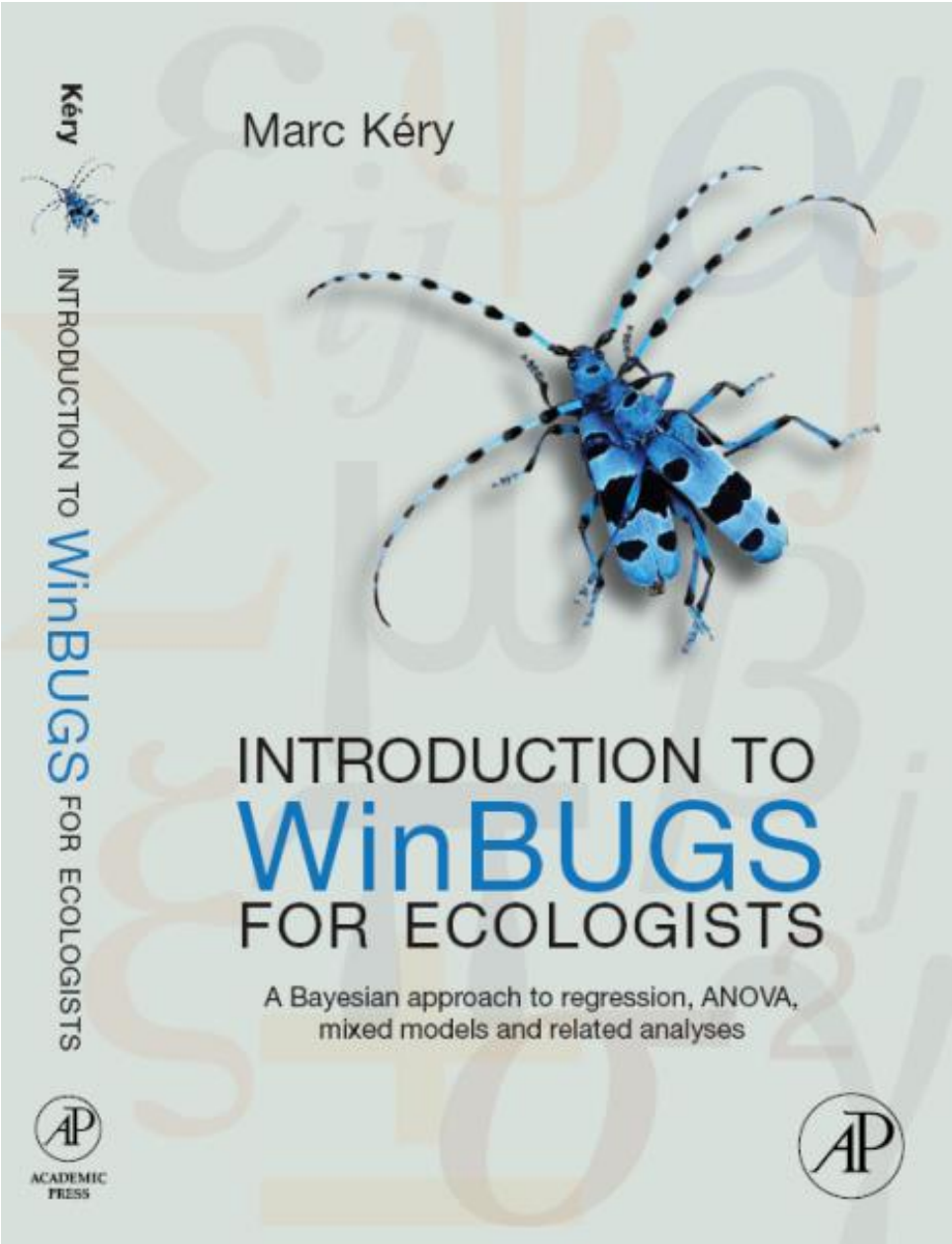
Introduction to Bayesian modeling



5 A first session with JAGS, NIMBLE, Stan and TMB: The "model of the mean"

5.1 Introduction

We will start with an analysis of one of the simplest possible models for a normal response– for what we here call the “model of the mean”. This is a normal linear regression model with just an intercept in the deterministic part of the response. Another description of this model would be that we estimate the mean and the variance of a normal distribution, which we assume as a statistical model for a sample of measurements taken from a population. Our first example will deal with body mass of male peregrines (Fig. 5-1).

Fig. 5-1: Male peregrine falcon (*Falco peregrinus*) wintering in the French Mediterranean, Sète, 2008 (Photo by Jean-Marc Delaunay).

The model in algebra looks like this, where y_i is the mass measurement for the i-th male peregrine:

$$y_i \sim Normal(\mu, \sigma^2)$$

Here, the mass measurements are defined to be independent draws from a Normal distribution with mean μ and variance σ^2 (and we could alternatively specify the dispersion in terms of the square root of the variance, i.e., in terms of the standard deviation σ).

6 Comparing two groups with equal or unequal variances

6.1 Introduction

In chapter 5, we had a sample of *n* measurements that were assumed to be what statisticians call *iid*: independent and identically distributed. We assumed they had the same mean and variance or standard deviation. In a first step now, we allow our sample to differ such that the units fall into one of two groups. And we will see that we we can assume such a two-group difference both in the mean and in the variance.

6.2 Comparing two groups with equal variances

One possible linear model thay underlies a two-group comparison such as the t-test with equal variances states that:

$$y_i = \alpha + \beta * x_i + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0, \sigma^2)$$

Here, a response y_i is a measurement on a continuous scale taken on individual i in two groups and x_i is an indicator or dummy variable that indexes the individuals in group 2 (see chapter 6 for different parameterizations of this model.) This simple t-test model has three parameters, the mean α for group 1, the difference in the means between groups 2 and 1 (β) and the variance σ^2 of the normal distribution from which the residuals ε_i are assumed to have come from.

6.2.1 Data generation

We first simulate data under this model and for a motivating example return to peregrine falcons. We imagine that we had measured the wingspan of a number of male and female birds and are interested in a sex difference in this measure of size. For Western Europe, Monneret (2006) gives the range of male wingspan as 70–85 cm and that for females as 95–115 cm. Assuming normal distributions for wingspan, this implies means and standard deviations of about 77.5 and 2.5 cm for males, and of 105 and 3 cm for females.

7 Normal linear regression

7.1 Introduction

We have seen in chapter 4 that the linear model underlying the simple normal linear regression is the same as that for a linear-modeling "variant" of a two-group comparison such as implied in a t-test:

$$y_i = \alpha + \beta * x_i + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0, \sigma^2)$$

The only difference is that the variable x_i doesn't just take on just two possible values to indicate membership to one of two groups; rather, variable *x* is a measurement that can take on any possible value, within some bounds and up to measurement accuracy. The geometric representation of this model is a straight line, with α being the intercept and β the slope.

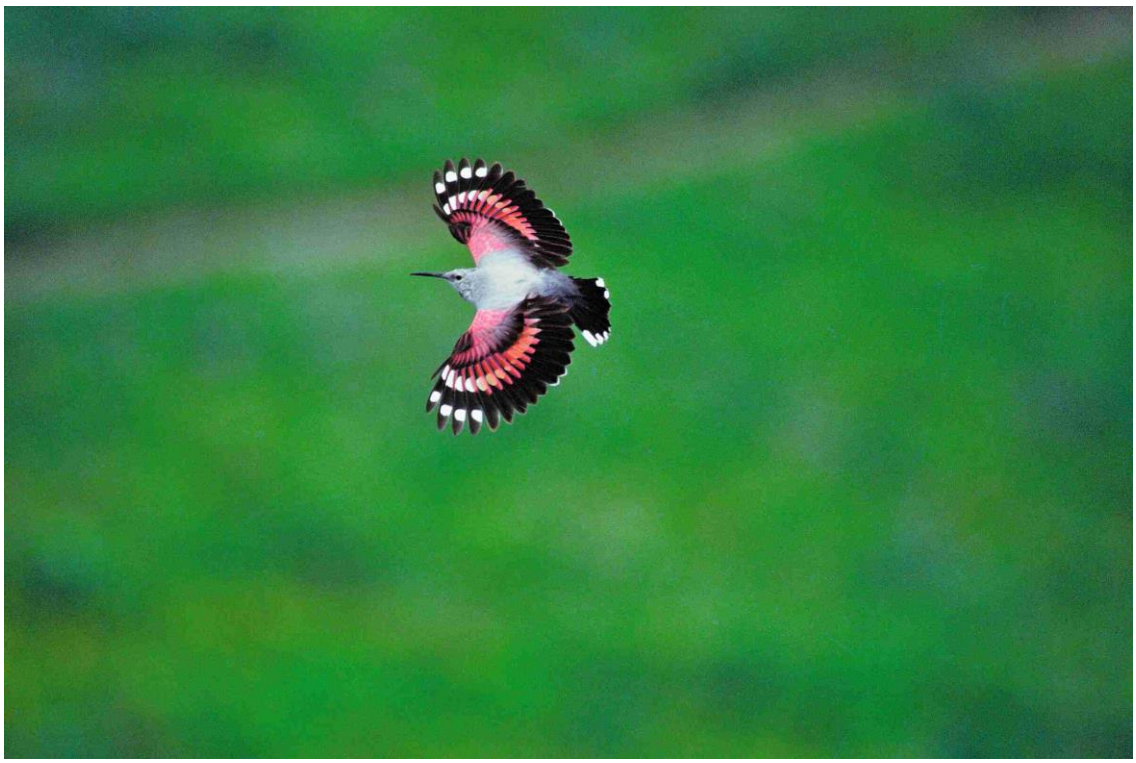


Fig. 7-1: Wallcreeper (*Tichodroma muraria*), Switzerland, 1989 (Photo by Eugen Hüttenmoser).

As a motivating example for a linear regression analysis we take a Swiss survey of the Wallcreeper (Fig. 7-1), a spectacular little cliff-inhabiting bird that appears to have declined greatly in Switzerland in recent years. Assume that we had data on the proportion of sample quadrats in which the species was observed in Switzerland for the years 1990–2005 and that we were willing to assume that the random deviations about a linear time trend were normally distributed. This is for illustration only, usually, we would use logistic regression (chapters 16–18) or a site-occupancy model (see chapter 19) to make inference about such data that have to do with the distribution of a species and represent a proportion (i.e., number occupied/number surveyed).

8 Comparisons in a single classification: Normal one-way ANOVA

As a motivating example for this chapter we assume that we measured snout-vent length (SVL) in five populations of Smooth snakes (Fig. 8-1). We are interested in characterizing the populations in terms of SVL of their snakes and in possible differences between populations.

Fig. 8-1: Smooth snake (*Coronella austriaca*), France, 2006 (Photo by Christophe Berney).

In terms of the underlying linear model a one-way ANOVA can be parameterized in various ways (see 4.3.4.). We here adopt a means parameterization of the linear model for a fixed-effects, one-way ANOVA and write this as follows:

$$y_i = \alpha_{j(i)} + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0, \sigma^2)$$

Here, y_i is the observed SVL of smooth snake i in population j, $\alpha_{j(i)}$ is the expected SVL of a snake in population j, and residual ε_i is the random SVL deviation of snake i from its population mean $\alpha_{j(i)}$. It is assumed to be normally distributed around zero with a constant variance σ^2.

9 Comparisons in two classifications: Normal 2-way ANOVA

In chapter 4 we already saw the linear models for the two-way ANOVA with interaction. In short, the effects parameterization for a model with two factors A and B is this:

$$y_i = \alpha + \beta_{j(i)} * A_i + \delta_{k(i)} * B_i + \gamma_{jk(i)} * A_i * B_i + \varepsilon_i ,$$

In contrast, the means parameterization can be written as this:

$$y_i = \alpha_{jk(i)} * A_i * B_i + \varepsilon_i$$

In both cases, we need to assume a distribution for the residuals to complete the model description:

$$\varepsilon_i \sim Normal(0, \sigma^2) .$$

Fig. 9-1: Mourning cloak (*Nymphalis antiopa*), Switzerland, 2006 (Photo by Thomas Marent)

10 General linear model for a normal response with both continuous and categorical explanatory variables

10.1 Introduction

The inferential situation considered in this chapter is the relationship between body mass and body length of the asp viper (Fig. 10-1) in three populations; Pyrenees, Massif Central, and the Jura mountains. We are particularly interested in population-specific differences of the mass-length relationship, that is, in the interactions between length and population. The means parameterization of the model we will fit can be written as (see 6.3.6.)

$y_i = \alpha_{j(i)} + \beta_{j(i)} * x_i + \varepsilon_i$, and # Pop * length

$\varepsilon_i \sim Normal(0, \sigma^2)$,

whre, y_i is body mass of individual i, $\alpha_{j(i)}$ and $\beta_{j(i)}$ are the intercept and the slope, respectively, of the mass-length relationship in population j, x_i is the body length of snake i and as usual, ε_i describes the combined effects of all unmeasured influences on the body mass of snake i and is assumed to behave like a normal random variable whose variance σ^2 we estimate.

Fig. 10-1: Male Asp viper (*Vipera aspis*), Switzerland, 2007 (Photo by Thomas Ott).

The effects parameterization of the same model is this:

$$y_i = \alpha_{Pyr} + \beta_1 * x_{MC(i)} + \beta_2 * x_{Jura(i)} + \beta_3 * x_{body(i)} + \beta_4 * x_{body(i)} * x_{MC(i)} + \beta_5 * x_{body(i)} * x_{Jura(i)} + \varepsilon_i$$

In addition to y_i and ε_i that are as before, α_{Pyr} is the expected mass of snakes in the Pyrenees, β_1 is the

difference between the expected mass of snakes in the Massif Central to that in the Pyrenees and $x_{MC(i)}$ is the indicator for snakes caught in the Massif Central. β_2 is the difference between the expected mass in the Jura to that in the Pyrenees and $x_{Jura(i)}$ is the indicator for snakes in the Jura, β_3 is the slope of the regression of body mass on body length x_{body} in the Pyrenees, β_4 is the difference in that slope between the Massif Central and the Pyrenees and β_5 the difference of slopes between Jura and the Pyrenees. Thus, snakes in the Pyrenees act as baseline with which snakes from the Massif Central and the Jura are compared, but as usual, this choice has no effect on inference.

11 Linear mixed-effects model

11.1 Introduction

We modify our Asp viper (Fig. 11-1) data set from Chapter 9 there just a little bit and assume that we now have measurements from a much larger number of populations, say, 56. A random-effects factor need not possess that many levels (some statisticians suggest to factors with any number of levels as random; see Gelman 2005), but in practice one rarely sees fewer than, say, 5–10 or so parameters fitted as random effects. Estimating a variance with so few values, which are moreover unobserved, will not result in very precise and perhaps even biased estimates (see also Lambert *et al.* 2004).

Fig. 11-1: Gravid female Asp viper (*Vipera aspis*), France, 2008 (Photo by Thomas Ott).

We will re-simulate some Asp viper data using R code fairly similar to that in the previous chapter. However, we will now *constrain the values for at least one set of effects* (either the intercepts and/or the slopes) to come from a normal distribution: this is what the random-effects assumption in a traditional mixed model means. There are at least three sets of assumptions that we could make about the random effects for the intercept or the slope of regression lines that are fitted to grouped (here, population-specific) data:

1. Only intercepts are random, but slopes are identical for all groups,
2. both intercepts and slopes are random, but they are independent, and
3. both intercepts and slopes are random and there is a correlation between them.

(An additional case, where slopes are random and intercepts are fixed, is not a sensible model in most circumstances.) Model No. 1 is often called a random-intercepts model, and both models No. 2 and 3 are also called random-coefficients models. As we will see, model No. 3 is the default in R's function `lmer()` when fitting a random-coefficients model.

Here is one way in which to write the random-coefficients model without correlation between the random effects for mass y_i of snake i in population j:

$$y_i = \alpha_{j(i)} + \beta_{j(i)} * x_i + \varepsilon_i$$
$$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) \quad \text{\# Random effects for intercepts}$$
$$\beta_j \sim \text{Normal}(\mu_\beta, \sigma^2_\beta) \quad \text{\# Random effects for slopes}$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma^2) \quad \text{\# Residual "random" effects}$$

Exactly as in the model in chapter 10, mass y_i is related to body length x_i of snake i in population j by a straight-line relationship with population-specific values for intercept α_j and slope β_j. (These regression parameters vary by individual i according to their membership to population j.) However, both α_j and β_j are now assumed to come from an independent normal distribution, with means μ_α and μ_β and variances of σ^2_α and σ^2_β, respectively. The residuals ε_i for snake i in population j are assumed to come from another independent normal distribution with variance σ^2. As a result, we could also say that when fitting this model, we simply estimate the parameters of three normal distributions simultaneously.

11.5 The random-coefficients model with correlation between intercept and slope

In a sense, the random-coefficients model with correlation is a simple extension of the previous model. The mass y_i of snake i in population j is assumed to be described by the following relations:

$$y_i = \alpha_{j(i)} + \beta_{j(i)} * x_i + \varepsilon_i$$
$$(\alpha_j, \beta_j) \sim \text{MVN}(\mu, \boldsymbol{\Sigma}) \quad \text{\# Multivariate normal random effects}$$
$$\mu = (\mu_\alpha, \mu_\beta) \quad \text{\# Mean vector}$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_\alpha & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma^2_\beta \end{pmatrix} \quad \text{\# Variance-covariance matrix}$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma^2) \quad \text{\# Residual "random" effects}$$

As before, the mass y_i of snake i in population j is related to its body length x_i by a straight-line relationship with population-specific values for intercept α_j and slope β_j. But now, pairs of α_j and β_j from the same population are assumed to come from a multivariate normal distribution (MVN; actually, here, a bivariate normal) with mean vector μ and variance-covariance matrix Σ. The latter contains the variances of the intercept (σ^2_α) and the slope (σ^2_β) in the diagonal and the covariance between α_j and β_j ($\sigma_{\alpha\beta}$) in the off-diagonals. As before, the residuals ε_i for snake i are assumed to come from an independent (univariate) normal distribution with variance σ^2. The interpretation of the covariance is such that positive values indicate a steeper mass-length relationship for snakes with a greater mass, while negative values show a shallower mass-length relationship for heavier snakes.

12 Introduction to the Generalized linear model (GLM): Comparing two groups in a Poisson regression

Formally, a GLM is described by three components:

1. a *statistical distribution* is used to describe the random variation in the response y; this is the stochastic part of the system description,
2. a so-called *link function* g, that is applied to the expectation of the response $E(y)$, and
3. a *linear predictor*, which is a linear combination of covariate effects that are thought to make up $g(E(y))$; this is the systematic or deterministic part of the system description.

To better see the analogy with the normal linear model, we start by writing the model for the normal two-group comparison (see Chapter 7) in GLM format:

1. Distribution: $y_i \sim Normal(\mu_i, \sigma^2)$

2. Link function: identity, i.e., $\mu_i = E(y_i)$ = linear predictor

3. Linear predictor : $\alpha + \beta * x_i$

Next we generalize this model to count data. The inferential situation considered is that of counts (C) of Brown hares (Fig. 13-1) in a sample of 10 arable and 10 grassland study areas. We wonder whether hare density depends on land-use.

Fig. 12-1: Brown hare (*Lepus europaeus*), Germany, 2008 (Photo by Niklaus Zbinden)

The typical distribution assumed for such counts is a Poisson, which applies when counted things are distributed independently and randomly and samples of equal size are taken randomly. Then, the number of hares counted per study area (C) will be described by a Poisson. The Poisson has a

single parameter, the expected count λ, that is often called the intensity and here represents the mean hare density. In contrast to the normal, the Poisson variance is not a free parameter but is equal to the mean λ. For a Poisson-distributed random variable C we write $C \sim Poisson(\lambda)$.

If hare density depends on land-use, i.e., is different in arable and grassland areas, the assumption of a constant mean density across all 20 study areas is not realistic. And in a "Poisson t-test" we are specifically interested in whether hare density differs between grassland and arable areas. Therefore, here is a model for hare count C_i in area i:

1. Distribution: $C_i \sim Poisson(\lambda_i)$
2. Link function: log, i.e., $\log(\lambda_i) = \log(E(C_i))$ = linear predictor
3. Linear predictor: $\alpha + \beta * x_i$

In words, hare count C_i in area i is distributed as a Poisson random variable with mean $E(C_i) = \lambda_i$. The log-transformation of λ_i is assumed to be a linear function $\alpha + \beta * x_i$, where α and β are unknown constants and x_i is the value of an area-specific covariate. If x_i is an indicator for arable areas, then α becomes the mean hare density on a log scale in grassland areas and β, again on a log-scale, is the difference in mean density between the two land-use types.

13 Overdispersion, zero-inflation and offsets in a Poisson GLM

14 Poisson regression with continuous and categorical explanatory variables

14.1 Introduction

We assume that instead of measuring body mass in Asp vipers in three populations in the Pyrenees, Massif Central and the Jura mountains, leading to a normal model, we had instead assessed ectoparasite load in a dragonfly, the Sombre Goldenring (Fig. 14-1), leading to a Poisson model. We are particularly interested in whether there are more or less little red mites on dragonflies of different size (expressed as wing length) and whether this relationship differs among the three mountain ranges.

Fig. 14-1: Sombre goldenring (*Cordulegaster bidentata)*, Kleinlützel, Switzerland, 1995 (Photo by Felix Labhardt)

We will fit the following model to mite count C_i on individual i:

1. Distribution: $C_i \sim Poisson(\lambda_i)$
2. Link function: log, i.e., $\log(\lambda_i)$ = $\log(E(C_i))$ = linear predictor
3. Lin. predictor:

$$\log(\lambda_i) = \alpha_{Pyr} + \beta_1 * x_{MC} + \beta_2 * x_{Jura} + \beta_3 * x_{wing} + \beta_4 * x_{wing} * x_{MC} + \beta_5 * x_{wing} * x_{Jura}$$

Note the great similarity between this model and the one we fitted to the mass of asps in chapter 11. Apart from the link function, the main other difference is simply that for this model we don't have a dispersion term; the Poisson already comes with a built-in variability. We could model overdispersion in the mite counts by using a Poisson-lognormal formulation as in the previous chapter, but we omit such added complexity here.

15 Poisson Generalized linear model or Poisson GLMM

15.1. Introduction

Here, we adopt a Poisson GLMM to analyze a set of long-term population surveys of Red-backed shrikes (Fig. 15-1).

Fig. 15-1: Male Red-backed shrike (*Lanius collurio*), Switzerland, 2004 (Photo A. Saunier).

We assume that pair counts over 30 years were available in each of 16 shrike populations (again, the balanced design is for convenience only). Our intent is to model population trends. First, we write down the random-coefficients model without correlation between the intercepts and slopes. This model is very similar to that for the normal linear case that we examined extensively in chapter 12. Thanks to how we specify models in the BUGS language, this similarity is more evident than when fitting the model with a canned routine such as in R.

We model C_i, the number pairs of Red-backed shrikes counted in year i in study area j:

1. Distribution: $C_i \sim Poisson(\lambda_i)$
2. Link function: log, i.e., $\log(\lambda_i) = \log(E(C_i))$ = linear predictor
3. Linear predictor: $\alpha_{j(i)} + \beta_{j(i)} * x_i$
4. Submodel for parameters/Distribution of random effects:

$$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_j \sim \text{Normal}(\mu_\beta, \sigma_\beta^2)$$

16 Comparing two groups in a Binomial regression

16.1 Introduction

As our inferential setting of this chapter, we consider a plant inventory on calcareous grasslands in the Jura mountains. A total of 50 sites were visited by experienced botanists who recorded whether they saw a species or not. The Cross-leaved gentian (Fig. 16-1) was found at 14 sites and the Chiltern gentian (see chapter 19) at 31 sites. We wonder whether this is enough evidence, given the variation inherent in binomial sampling, to claim that the Cross-leaved gentian has a more restricted distribution in the Jura mountains.

Fig. 16-1: Cross-leaved gentian (*Gentiana cruciata*), Spanish Pyrenees, 2006 (Photo by Marc Kéry).

For gentian species i, let C_i be the number of sites it was detected. A simple model for C_i is this:

1. (Statistical) Distribution: $C_i \sim Binomial(N, p_i)$

2. Link function: logit, i.e., $\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$ = linear predictor

3. Linear predictor: $\alpha + \beta * x_i$

If x is an indicator for the Chiltern gentian, then α can be interpreted as a logit-scale parameter for the probability of occurrence of the Cross-leaved gentian in the Jura mountains and β is the difference, on a logit-scale, between the probability of occurrence of the Chiltern gentian and that of the Cross-leaved gentian.

17 Binomial GLM with continuous and categorical explanatory variables

We will model a count governed by an underlying probability; specifically, we model the proportion of black individuals in Adder populations. The adder has an all-black and a zigzag morph, where females are brown and males grey (Fig. 17-1).

Fig. 17-1: Male adder (*Vipera berus*) of the zigzag morph, Germany, 2007 (Photo by Thomas Ott*)*

It has been hypothesized that the black color confers a thermal advantage and therefore the proportion of black individuals should be greater in cooler or wetter habitats. We will simulate data that bear on this question and "study", by simulation, 10 adder populations each in the Jura mountains, the Black Forest and the Alps. We will capture a number of snakes in these populations and record the proportion of black adders. Then, we relate these proportions to the mountain range as well as to a combined index of low temperature, wetness and northerliness of the site. Our expectation will of course be that there are relatively more black adders at cool and wet sites. As always, a count is the result of a true number (here, of black and zigzag adders) and a detection probability. Hence, in the following analyses, we make the implicit assumption that the detectability of black and zigzag adders neither differs between each other, nor among populations.

We will model the number of black adders C_i among N_i captured animals in population i. Here is the description of the model.

1. Distribution: $C_i \sim Binomial(p_i, N_i)$

2. Link function: logit, i.e., $\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$ = linear predictor

3. Linear predictor: $\alpha_{Jura} + \beta_1 * x_{BlackF} + \beta_2 * x_{Alps} + \beta_3 * x_{wet} + \beta_4 * x_{wet} * x_{BlackF} + \beta_5 * x_{wet} * x_{Alps}$

Note that the number of animals captured, N_i, is not a parameter of the binomial distribution in this example but is known (in contrast to the model of chapter 20 where N_i will be estimated). The link function is the logit, as is customary for a binomial distribution, though other links are possible (e.g., the complementary log-log, which is asymmetrical; see GLM textbooks such as McCullagh and Nelder 1989). Finally, the linear predictor is made up of an intercept α_{Jura} for the expected proportion, on the logit scale, of black adders in the Jura, parameters β_1 and β_2 for the difference in the intercept between Black Forest and the Jura and the Alps and the Jura, respectively. The parameters β_3, β_4 and β_5 specify the slope of the (logit-linear) relationship between the proportion of black adders and the wetness indicator of a site in the Jura and the difference from β_3 of these slopes in the Black Forest and the Alps, respectively.

18 Binomial mixed-effects model (Binomial GLMM)

As in a Poisson GLMM, we can also add into a binomial GLM random variation beyond what is stipulated by the binomial distribution. We illustrate this for a slight modification of the Red-backed shrike example from chapter 15. Instead of counting the number of pairs, which naturally leads to the adoption of a Poisson model, we now imagine that we study the reproductive success (success or failure) of its much rarer cousin, the glorious Woodchat shrike (Fig. 18-1). We examine the relationship between precipitation during the breeding season and reproductive success; wet springs are likely to depress the proportion of successful nests. We assemble data from 16 populations studied over 10 years.

Fig. 18-1: Woodchat shrike (*Lanius senator*), Catalonia, 2008 (Photo by Jordi Rojals).

First, we write down the random-coefficients model (without intercept-slope correlation) for a binomial response. We model C_i, the number successful pairs among N_i studied pairs in year i and study area j:

1. Distribution:	$C_i \sim Binomial(p_i, N_i)$
2. Link function:	logit, i.e., $\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$ = linear predictor
3. Linear predictor:	$\alpha_{j(i)} + \beta_{j(i)} * x_i$
4. Submodel for parameters:	$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha)$
	$\beta_j \sim \text{Normal}(\mu_\beta, \sigma^2_\beta)$

Except for a different distribution and link function, the additional kind of data provided by the binomial totals N_i, and a different interpretation and indexing of the covariate, this model looks exactly like the Poisson GLMM in chapter 15 ! The linear predictor, $\alpha_{j(i)} + \beta_{j(i)} * x_i$, specifies a

population-specific, logit-linear relationship between breeding success and precipitation. Furthermore, populations are assumed to be related in the sense that both intercepts (α_j) and slopes (β_j) of those relationships come from two normal distributions whose hyperparameters we estimate.

19 Non-GLMM hierarchical models 1: Site-occupancy species distribution model

Fig. 19-1: The wonderful Chiltern gentian (*Gentianella germanica*), Slovenia, 2007 (Photo by Milan Vogrin).

As a motivating example we consider an inventory of the beautiful Chiltern gentian (Fig. 19-1) conducted in 150 patches of calcareous grassland in the wild and wonderful Jura mountains. Our aim is to estimate the number or the proportion of occupied sites and to identify environmental factors related to the presence or the absence of the gentian at a site..

The genesis, and therefore the analysis, of detection/nondetection observation $y_{i,t}$ at site i during survey t is naturally described by a hierarchical model that contains one submodel for the only partially observed true state (occurrence, the result of the biological process), and another submodel for the actual observations. The actual observations result from both the particular realization of the biological process and of the observation process and thus from two random processes in sequence, which makes the resulting model hierarchical.

$z_i \sim Bernoulli(\psi)$ Biological process yields true state

$y_{i,t} \sim Bernoulli(z_i \times p_{i,t})$ Observation process yields observations

Hence, true occurrence z_i of *G. germanica* at site i is a Bernoulli random variable governed by the parameter ψ which is occurrence probability: this is exactly the parameter that most distribution modelers wish they were modeling. The actual gentian observation $y_{i,t}$, detection or not at site i during survey t (or "presence-absence" datum $y_{i,t}$), is another Bernoulli random variable with a success probability that can be expressed as the product of the actual occurrence of *G. germanica* at that site, z_i, and detection probability $p_{i,t}$ at site i during survey t. Hence, at a site where the gentian doesn't occur, $z = 0$, and y must be 0. Conversely, at an occupied site we have $z = 1$, and *G. germanica* is detected with probability $p_{i,t}$. That is, in the site-occupancy model, detection probability is expressed *conditional on occurrence*, and the two parameters ψ and p are separately estimable if replicate visits are available.

20 Non-GLMM hierarchical models 2: Binomial N-mixture model to model abundance

Fig. 20-1: Pair of sand lizards (*Lacerta agilis*), Switzerland, 2006 (Photo by Thomas Ott).

We assume that a count $y_{i,t}$ at site i and made during survey t comes from a two-stage stochastic process. The first stochastic process is the biological process that distributes the animals among the sites. This process generates the site-specific abundance that we would like to model directly but cannot because we hardly ever see all individuals. The standard statistical model for such data is the Poisson distribution, governed by the intensity (density) parameter λ, which is typically expressed conditional on habitat covariates. The result of this first stochastic process is the local, site-specific abundance N_i. Given that true state N_i of a site, the second stochastic process is the observation process which, together with N_i, determines the data actually observed, i.e., the counts $y_{i,t}$. A natural model for the observation process in the presence of imperfect detection, but in the absence of double counts, is the binomial distribution; given that there are N_i sand lizards present and that each has a probability of $p_{i,t}$ to be observed at site i during replicate survey t, the number of lizards actually observed is binomially distributed. Two important consequences are that (1) we typically observe fewer than N_i lizards, and (2) the counts $y_{i,t}$ will vary automatically from survey to survey even under identical conditions (Kéry and Schmidt 2008). Three important assumptions of the binomial mixture model are that of population closure, independent and identical detection probability for all individuals at site i and during survey t except insofar as differences among sites or surveys are modeled by covariates, and the absence of double counts and other false positive errors. The effects of violations of these assumptions are still being investigated (e.g., Joseph *et al.* 2009).

In summary, the binomial mixture model to estimate abundance from temporally and spatially replicated counts can be written succinctly in just two lines:

$N_i \sim Poisson(\lambda)$	Biological process yields true state
$y_{i,t} \sim Binomial(N_i, p_{i,t})$	Observation process yields observations