8 Introduction to mixed modelling

8.1 Introduction

This chapter gives a non-technical introduction into mixed modelling. Mixed models are also known as mixed effects models or multilevel models and are used when the data have some sort of hierarchical form such as in longitudinal or panel data, repeated measures, time series and blocked experiments, which can have both fixed and random coefficients together with multiple error terms. It can be an extremely useful tool, but is potentially difficult to understand and to apply. In this chapter, we explain mixed models with random intercept and slope, different variances, and with an auto-correlation structure. In the case study chapters, several examples are presented (Chapters 22, 23, 26, 35 and 37). Some of these chapters are within a linear modelling context, whereas others use smoothing methods leading to additive mixed modelling. In the time series analysis and spatial statistics chapters (16, 17 and 19), temporal and spatial correlation structure is added to the linear regression and smoothing models. All of these techniques can be seen as extensions of mixed modelling.

Good references on mixed modelling and additive mixed modelling are Brown and Prescott (1999), Snijders and Bosker (1999), Pinheiro and Bates (2000), Diggle et al. (2002), Ruppert et al. (2003), Fitzmaurice et al. (2004), Faraway (2006) and Wood (2006). It should be noted that most of the literature on mixed modelling is technical. For non-mathematical text there is Chapter 35 in Crawley (2002) and Twisk (2006).

The easiest way to introduce mixed modelling is by using an example. In Chapters 4 to 7 and 27, we used a marine benthic dataset, referred to as the RIKZ data. Figure 8.1-A shows a scatterplot for these data with NAP plotted against species richness. Richness is the number of species found at each site, and NAP is the height of a site compared with average sea level. The data were sampled at nine beaches (five samples per beach), and the question being asked is whether there are any differences between the NAP-richness relationship at these nine beaches. As discussed in Chapters 5 and 6, species richness is a non-negative integer and we used it to explain both linear regression and generalised linear modelling with a Poisson distribution and log link. The same approach can be followed for mixed modelling. We will explain linear mixed modelling as an extension of linear regression (using the normal distribution) with species richness as the response vari-

able. Generalised linear mixed modelling (GLMM) is the mixed modelling extension of GLM. Although GLMM may be more appropriate for species richness, it is outside the scope of this book. Panel A in Figure 8.1 contains a regression line using all 45 observations from all nine beaches. The regression line in Figure 8.1-A can be written as

Model 1 $$Y_i = \alpha + \beta \text{NAP}_i + \varepsilon_i \quad \text{and} \quad \varepsilon_i \sim N(0, \sigma^2)$$

This model contains three unknown parameters: the two regression parameters (one intercept and one slope) and the variance. The model assumes that the richness-NAP relationship is the same at all nine beaches. Figure 8.1-B contains the same data except that we have added a regression line for each beach. The directions and values of the nine regression lines indicate that the intercepts are different for each beach, and that two beaches have noticeably different slopes. This model can be written as

Model 2 $$Y_{ij} = \alpha_j + \beta_j \text{NAP}_{ij} + \varepsilon_{ij} \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Where j = 1,…,9 and i = 1,..,5 (there are five observations per beach). The regression lines were obtained in one regression analysis using beach as a factor, NAP as a continuous explanatory variable and a beach-NAP interaction term. This can also be called an ANCOVA (Zar 1999). The total number of parameters adds up to 9 × 2 + 1 = 19 (a slope and intercept per beach plus one noise variance term). The total number of observations is only 45, so proportionally this model has a worryingly large number of parameters.

The model with only one regression line (Figure 8.1-A) is the most basic model, and the model with nine regression lines in which slope and intercept are allowed to differ per beach (Figure 8.1-B) is the most complex model we have considered here. Later in this chapter, we will discuss even more complex models by allowing for different variances per beach. There are two intermediate models: one where the intercept is allowed to differ between the beaches but with equal slopes, and one where the intercepts are kept same and the slopes are allowed to differ. This reduces the number of parameters. These models are given by

Model 3 $$Y_{ij} = \alpha_j + \beta \text{NAP}_{ij} + \varepsilon_{ij} \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Model 4 $$Y_{ij} = \alpha + \beta_j \text{NAP}_{ij} + \varepsilon_{ij} \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

The fitted values of models 3 and 4 are given in panels C and D of Figure 8.1, respectively. The number of parameters in both models is 1 + 9 + 1 = 11. In Figure 8.1-D all lines intercept at NAP=0 because the intercepts are identical.

Using either an F-test or the AIC (see Chapter 5), it can be shown that of the four models, the one used in Figure 8.1-B is the best. This is the model using a different intercept and slope at each beach, giving a total of 19 parameters. The nine intercepts tell us which beaches have higher richness or lower richness at NAP = 0, and the nine slopes show the differences in the strength of the NAP-richness relationship.

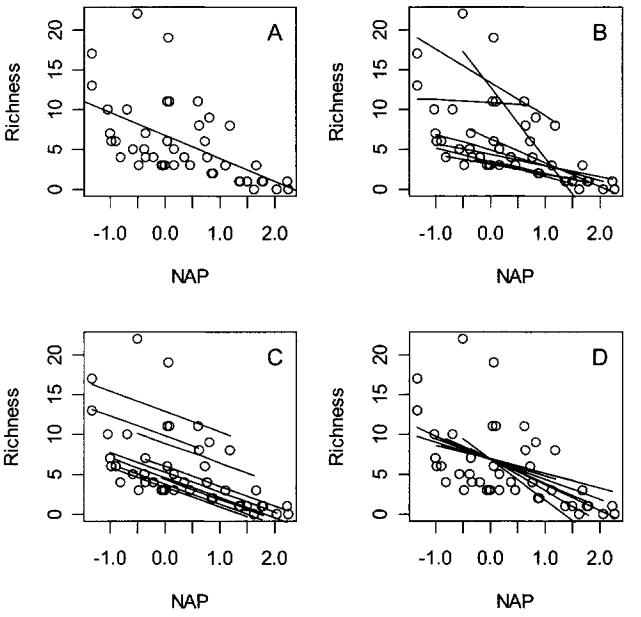

Figure 8.1. Scatterplots of NAP versus richness for the RIKZ data. A: One regression line was added using all data. B: One regression line for each beach was plotted. C: One regression line per beach but with all slopes equal. D: One regression line per beach but with intercepts equal.

If we are only interested in the general relationship between richness and NAP, and do not care about differences between beaches, then we could ignore the nominal variable beach. However, this means that the variance component might contain between-beach variation, and not taking this into account might affect standard errors and p-values of the fixed effects (e.g., suggesting a non-significant relationship between NAP and species richness even when the relationship is significant). But the price of 16 extra regression parameters can be rather large, namely in the loss of *precious degrees of freedom*! To avoid this, mixed modelling can be used. But there is another motivation for using mixed modelling with these data. If beach is used as a fixed term, we can only make a statement of richness-NAP relationships for these particular beaches, whereas if we use it as a random component, we can predict the richness-NAP relationship for all similar beaches.

8.2 The random intercept and slope model

For simplicity, we start with model 3 discussed earlier. In this model, each beach has a different intercept but the same slope. We extend it to:

Model 5

$$Y_{ij} = \alpha + \beta \text{NAP}_{ij} + a_j + \varepsilon_{ij}$$
$$\text{where} \quad a_j \sim N(0, \sigma_a^2) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

The index j (representing beaches) takes values from 1 to 9, and i (representing samples on a beach) from 1 to 5. In model 3, we ended up with nine regression lines. The nine estimated slopes, intercepts and their standard errors tell us which beaches are different. In model 5, we assume there is only one overall regression line with a single intercept and a single slope. The single intercept α and single slope β are called the fixed parameters. Additionally, there is a random intercept a_j, which adds a certain amount of random variation to the intercept at each beach. The random intercept is assumed to follow a normal distribution with expectation 0 and variance σ_a^2. Hence, the unknown parameters in the model are α, β, the variance of the noise σ^2 and the variance of the random intercept σ_a^2, which adds up to only four parameters. So, basically, we have the same type of model fit as in panel C in Figure 8.1, but instead of nine estimated levels for each intercept, we now have nine realisations a_1, .., a_9 from which we assume that they follow a normal distribution. And we only need to estimate the variance of this distribution. The first part of the numerical output for this mixed model is given by: AIC = 247.580, BIC = 254.525 and logLik = −119.740. As with linear regression, GLM and GAM, we can measure how optimal the model is by using a selection criteria like the AIC. An alternative is the BIC, which is more conservative (the number of parameters have a higher penalty than in the AIC). Both terms use the log likelihood as a measure of fit. Further output gives:

Random effects:

	Intercept	Residual
StdDev:	2.944	3.060

The output for the random effects (above) shows that the variance of the noise is equal to $\sigma^2 = 3.06^2$ and the variance of the random intercept to $\sigma_a^2 = 2.944^2$.

Fixed effects:

	Value	Std.Error	df	*t*-value	*p*-value
Intercept	6.582	1.096	35	6.006	<0.001
NAP	−2.568	0.495	35	−5.191	<0.001

For the fixed effects part of the model (above), $\alpha + \beta \text{NAP}_{ij}$, the intercept α is equal to 6.582 and the slope β is –2.568. Both parameters are significantly differently from 0 at the 5% level. The correlation between the estimated intercept

and slope is small. The ANOVA table below also indicates the significance of the fixed slope β.

	numdf	dendf	*F*-value	*p*-value
Intercept	1	35	27.634	<0.001
NAP	1	35	26.952	<0.001

In summary, the model estimates a fixed component of the form: $6.582 - 2.568\text{NAP}_{ij}$. These estimated parameters are significantly differently from 0 at the 5% level. For each beach, the intercept is increased or decreased by a random value. This random value follows a normal distribution with expectation 0 and variance $\sigma_a^2 = 2.944^2$. The unexplained noise has a variance of $\sigma^2 = 3.06^2$. The type of model is called a mixed effects model with random intercept.

Extending the mixed model with a random slope

We will now extend model 5 to get the mixed modelling equivalent of model 2, which was illustrated in Figure 8.1-B. Model 5 was a mixed model with a random intercept that allowed the regression line to randomly shift up or down. In model 6, both the intercept and slope are allowed to randomly vary. The model is given by

Model 6

$$Y_{ij} = \alpha + a_j + \beta \text{NAP}_{ij} + b_j NAP_{ij} + \varepsilon_{ij}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $a_j \sim N(0, \sigma_a^2)$ and $b_j \sim N(0, \sigma_b^2)$

This is the same model formulation as model 5, except for the term $b_j\text{NAP}_{ij}$. This new term allows for random variation of the slope at each beach. The model fit will look similar to the one in Figure 8.1-B, except that considerably fewer parameters are used: two for the fixed intercept α and slope β, and three random variances σ^2, σ_a^2 and σ_b^2. In general, one allows for correlation between the estimated variances for σ_a^2 and σ_b^2. The numerical output shows that the AIC is 244.397, which is slightly smaller than the AIC for model 5. As in regression, the same principle of 'the smaller the better' holds, so the lower AIC indicates that model 6 is more optimal than model 5. It should be noted that if the difference is smaller than two, the models are generally thought of as equivalent and the more simple one should be selected (Burnham and Anderson 2002).

Random effects:

	StdDev
Intercept	3.573
NAP	1.758
Residual	2.668

The printout (above) for the random effects component shows that $\sigma^2 = 2.668^2$, $\sigma_a^2 = 3.573^2$, and $\sigma_b^2 = 1.758^2$.

Fixed effects:

	Value	Std.Error	df	*t*-value	*p*-value
(Intercept)	6.612	1.271	35	5.203	<0.001
NAP	−2.829	0.732	35	−3.865	<0.001

The fixed component (above) is given by: 6.612 − 2.829NAP. Both intercept and slope are significant at the 5% level. This can also be inferred from the ANOVA table below.

ANOVA table

	Numdf	dendf	*F*-value	*p*-value
(Intercept)	1	35	12.427	0.001
NAP	1	35	14.939	0.001

8.3 Model selection and validation

We have now applied two mixed models: one where the intercept was allowed to vary randomly (model 5), and one where the intercept and slope were allowed to vary randomly (model 6). The question that arises is which model is better. It is important to realise that the only difference between model 5 and 6 is the random effects component; the fixed components are the same. To compare two models with the same fixed effect, but with different random components, a likelihood ratio test or the AIC can be used. In this case, we get:

Model	df	AIC	BIC	logLik	*L*-Ratio	*p*-value
6	6	244.40	254.96	−116.20		
5	4	247.48	254.53	−119.74	7.08	0.029

The AIC suggests selecting model 6, but the BIC picks model 5. The *p*-value indicates that the more complicated model (containing a random intercept and slope) is more optimal. However, there is one major problem with the likelihood ratio test. In Chapter 5, we used an *F*-test to compare two nested models and the residual sum of squares obtained by ordinary least squares of the two models were used to work out a test statistic. Estimation in mixed modelling is done with maximum likelihood (Chapter 7). The likelihood criteria of the full model L_0 and the nested model L_1 can be used for hypothesis testing. It is in fact the ratio L_0/L_1 that is used, hence the name likelihood ratio test. Taking the log or, more common, −2log gives a test statistic of the form: $L = -2(\log L_0 - \log L_1)$. It can be shown that under the null hypothesis, this test statistic follows *approximately* a Chi-square distribution with v degrees of freedom, where v is the difference in number of parameters in the two models. In linear regression, we can use this procedure to test H_0: $\beta_i = 0$ versus H_1: $\beta_i \neq 0$. So, what are we testing above? The test statistic is $L = 7.08$ with a p-value of 0.029. But what is the null hypothesis? The only difference between models 5 and 6 is the random component b_jNAP where $b_j \sim N(0, \sigma_b^2)$. By comparing models 5 and 6 using a likelihood ratio test, we are testing the null hypothesis H_0: $\sigma_b^2 = 0$ versus H_1: $\sigma_b^2 > 0$. The alternative hypothesis

has to contain a '>' because variance components are supposed to be non-negative. This is called the boundary problem; we are testing whether the variance is equal to null, but null is on the boundary of all allowable values of the variance. The problem is that the theory underlying the likelihood ratio test, which gives us a *p*-value, assumes that we are not on the boundary of the parameter space. For this particular example, comparing a random intercept versus a random intercept plus slope model, it is still possible to derive a valid distribution for the test statistic, see p. 106 in Rupert et al. (2003). But as soon as we compare more complicated variance structures that are on the boundary, the underlying mathematics become rather complicated. Formulated differently, great care is needed with interpreting the *p*-value of the test statistic *L* if we are testing on the boundary. And in most applications, this is what we are interested in. Citing from Wood (2006): 'In practice, the most sensible approach is often to treat the *p*-values from the likelihood ratio test as "very approximate". If the *p*-value is very large, or very small, there is no practical difficulty about interpreting it, but on the borderline of significance, more care is needed'. Faraway (2006) mentions that the *p*-value tends to be too large if we are testing on the boundary. This means that a *p*-value of 0.029 means that we can reasonably trust that it is significant, but 0.07 would be a problem. Pinheiro and Bates (2000), Faraway (2006) and Wood (2006) all discuss bootstrapping as a way to obtain more trustable p-values if testing on the boundary.

Error terms

So far, we have *assumed* that the error term ε_{ij} has the same variance at each beach, and this is called homogeneity. If residuals plots indicate violation of homogeneity (see Chapter 5 for details), it might be an option to use different variance components per beach. A possible model could be as follows:

Model 7

$$Y_{ij} = \alpha + \beta \text{NAP}_{ij} + a_j + b_j NAP_{ij} + \varepsilon_{ij}$$

where $a_j \sim N(0, \sigma_a^2)$, $b_j \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$

This model is nearly identical to model 6 except that the variance component of the noise now has an index *j*, where *j* = 1, .., 9. The output of model 7 shows that the AIC = 216.154 (BIC = 240.811), which is considerably smaller than model 6, and the variances of the random effects are also lower in model 7 than model 6 (below).

```
Random effects:
             StdDev
(Intercept)  2.940
NAP          0.011
Residual     1.472
```

In model 6 we had $\sigma^2 = 2.668^2$, $\sigma_a^2 = 3.573^2$, and $\sigma_b^2 = 1.758^2$. Both the residual variance and the variance for the random slope are considerably smaller in model

7 than model 6. An informal way to assess the importance of a variance component is its relative size. 0.011^2 is relatively small!

Variance function:
Parameter estimates:

1	2	3	4	5	6	7	8	9
1.0	3.387	0.578	0.269	4.891	0.426	0.450	0.745	2.33

We now have a residual variance component for each beach (above). The largest residual variance is at beaches two, five and nine.

Fixed effects:

	Value	Std.Error	df	*t*-value	*p*-value
(Intercept)	5.75	1.056	35	5.446	<0.001
NAP	−1.42	0.127	35	−11.188	<0.001

The output for the fixed effects part for model 7 (above) shows that it has the form: 5.75 −1.42NAP. The AIC or hypothesis testing with the likelihood ratio test can be used to judge which model is better.

Model	df	AIC	BIC	logLik	*L*-Ratio	*p*-value
6	4	247.480	254.525	−119.740		
7	12	212.154	233.289	−94.077	51.325	<0.001

The likelihood ratio test (above) indicates that model 7 is better than model 6. The effect of introducing nine variances is illustrated in Figure 8.2. Panels A and B show the standardised residuals versus fitted values and QQ-plot for model 6. Panels A and B show the same for model 7. Note that in model 6 there is clear evidence of heterogeneity (see panel A).

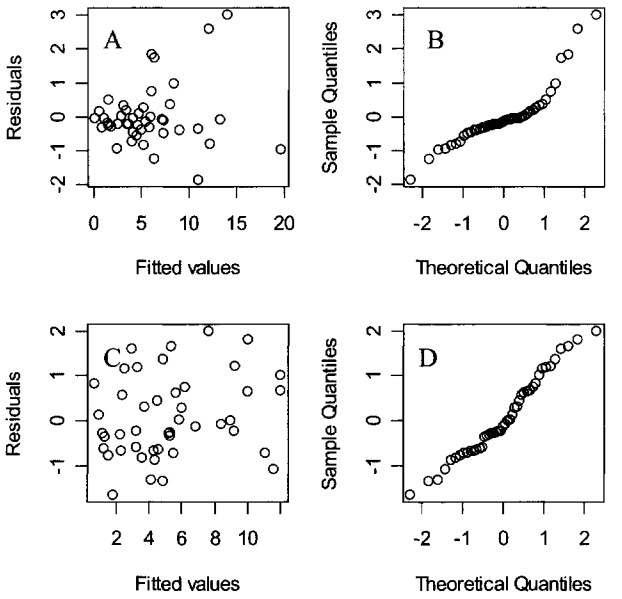

Figure 8.2. A: Standardised residuals versus fitted values for model 6. B: QQ-plot of residuals of model 6. C: Standardised residuals versus fitted values for model 7. D: QQ-plot of residuals of model 7.

Comparing models with different fixed effects

So far, we have only compared models with the same fixed components, but for the RIKZ data we also have exposure per beach. This is a nominal variable with three classes, and one of the prime interests of the project was to know the effects of exposure. As (i) exposure can only fall into one of three pre-defined categories and (ii) one of the prime underlying questions is whether there is an exposure effect, it is modelled as a fixed effect and not as random effect and we have extended model 7 to include this new effect:

Model 8

$$Y_{ij} = \alpha + \beta \text{NAP}_{ij} + \text{exposure}_{ij} + a_j + b_j NAP_{ij} + \varepsilon_{ij}$$

where $a_j \sim N(0, \sigma_a^2)$, $b_j \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$

Exposure is modelled as a factor. There are different ways to include the levels of a nominal variable, and the default value in software packages like R and SPLUS is the treatment option. This means that the first level is set to 0 and is considered as the baseline level (Chapter 5). To assess whether adding exposure results in a better model, we can inspect the individual p-values, compare AIC values or apply a likelihood ratio test. The output is presented below and indicates that model 8 is better than model 7.

Model	df	AIC	BIC	logLik	L.Ratio	*p-value*
8	14	200.978	226.271	−86.488		
7	12	211.684	233.365	−93.842	14.707	<0.001

As we now compare two models with the same random structure but with different fixed terms, the maximum likelihood method is used instead of REML (see the next section).

Model selection strategy

We started this chapter with linear regression, then introduced the random intercept model, extended it to the random intercept and slope model, and finally added more explanatory variables. This order was chosen for reasons of clarity, but we now look at finding the optimal model for both the random and the fixed components. This discussion also applies to additive mixed modelling. Our starting point is either a multiple linear regression or an additive model:

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i \quad \text{or} \quad Y_i = \alpha + f_1(X_{1i}) + \cdots + f_p(X_{pi}) + \varepsilon_i$$

The part $\alpha + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$ or $\alpha + f_1(X_{1i}) + \cdots + f_p(X_{pi})$ is called the fixed component, and ε_i is the random component. We have already seen three forms of random components; random intercepts, random slopes and different variances. More structures will follow for time series and spatial data. The task of the researcher is to find the optimal fixed and random component structure. We assume that the prime interest of the analysis is the fixed component, so we need to get this one as good as possible. However, a poor fixed component structure may result in large residuals and therefore all types of random structures that could have been taken care of by the fixed component. For the RIKZ data, we started with a model using only NAP as an explanatory variable, from a full dataset of 10-15 explanatory variables. The different variances between beaches may be explained with a more appropriate fixed component, and we prefer to have this information as a fixed component, and not in the random structure, if possible. So, the consensus is to start with a model that has a fixed component structure that is as good as possible. Most textbooks advise including every possible explanatory variable and interaction, or a 'just beyond optimal' model. Using this 'just beyond optimal' model, we can select the most optimal random structure and ensure that the random structure does not contain any information that could have been modelled with fixed terms. As we will explain later in this chapter, this process requires one to use the restricted maximum likelihood (REML) estimation procedure. Once we have found the optimal random component structure, then we go to the third step and find the optimal fixed component structure. This requires a backwards selection approach where we drop the explanatory variables that are not significant using maximum likelihood (ML) estimation (which will be explained later). Summarising, the model selection strategy:

1. Start with a model that is close to optimal in terms of fixed components.

2. Search for the optimal random error structure, e.g., allowing for random intercepts and slopes, different variances, auto-correlation, spatial correlation or any of the many options described in Pinheiro and Bates (2000). Use REML.
3. Using the optimal random error structure from step 2, find the optimal fixed components using ML.
4. Present the estimated parameters and standard errors of the optimal model, but use REML estimation!

Needless to say, once the model selection has given us the optimal model, a model validation should be applied. This process is similar to linear regression, except that we now allow for different spread per strata, auto-correlation, etc.

Fixed or random

To find the differences between beaches for the RIKZ data, then clearly you should model these variables as fixed effects rather than as random effects. The advantage of treating the levels of a variable as random effects is that the model then holds for the entire population. Treating beach as a fixed effect in the RIKZ data means that the richness-NAP relationship found by the model only holds for those nine beaches. Treating it as a random component means that the species-richness relationship holds for all beaches, not just the nine beaches we sampled. This means that we need to assume that the nine sampled beaches represent the population of all beaches (with similar values of the other explanatory variables like grain size, etc.). If we cannot make this assumption, then treat them as fixed effects. Another important point is the number of levels. Are there enough levels (> 4, but preferable > 10) in the variable to treat it like a random effect? If there are only two or three levels, treat it as a fixed effect. But if there are more than 10 levels, then treat is as a random effect. In cases where an explanatory variable is a treatment effect, consider it as a fixed effect as in most situations the prime interest of the experiment is in the effects of the treatment (like toxic concentrations in mesocosm experiments).

8.4 A bit of theory

A linear regression model can be written as $Y_j = \beta X_j + \varepsilon_j$. The mathematical formulation of a mixed model is as follows:

$$Y_{ij} = \beta X_{ij} + b_j Z_{ij} + \varepsilon_{ij}$$
$$b_j \sim N(0, \Psi) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

The component βX_{ij} contains the fixed effects and bZ_{ij} the random components. The random components b_j and ε_{ij} are assumed to be independent of each other. Covariance between the random effects b_j is allowed (e.g., between random

intercept and slope) using the off-diagonal elements of Ψ. A two-stage algorithm, REML, is used to estimate the fixed regression coefficients and the variance components Ψ and σ. Full details of the REML estimation process can be found in Brown and Prescott (1999) or Pinheiro and Bates (2000). Most textbook discussions on REML and ML use complicated mathematics, but a reasonable non-technical explanation can be found in Fitzmaurice et al. (2004). In maximum likelihood estimation, the maximum likelihood function is specified, and to optimise it, derivates with respect to the regression parameters and variances are derived. The problem is that the estimates for the variance(s) are biased. In REML, a correction is applied so that less biased estimators are obtained. So, in general REML estimators are less biased than ML estimators. For large sample size (relative to the number of regression parameters), this issue is less important.

Covariances and Correlations

Revisiting some of the models we used for the RIKZ data we had:

Model 3 $$Y_{ij} = \alpha_j + \beta_j \text{NAP}_{ij} + \varepsilon_{ij} \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Model 5 $$Y_{ij} = \alpha + \beta \text{NAP}_{ij} + a_j + \varepsilon_{ij}$$
$$\text{where} \quad a_j \sim N(0, \sigma_a^2) \text{ and } \varepsilon_{ij} \sim N(0, \sigma^2)$$

Model 6 $$Y_{ij} = \alpha + a_j + \beta \text{NAP}_{ij} + b_j NAP_{ij} + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim N(0, \sigma^2) \text{ and } a_j \sim N(0, \sigma_a^2) \text{ and } b_j \sim N(0, \sigma_b^2)$$

Model 3 was the ordinary regression model with interaction. Model 5 was called the random intercept model and model 6 the random intercept and slope model. We now focus on the covariance and correlation between two observations from the same beach j: samples Y_{ij} and Y_{kj}. Under model 3, this gives us:

$$Var(Y_{ij}) = Var(\varepsilon_{ij}) = \sigma^2$$
$$Cov(Y_{ij}, Y_{kj}) = \begin{cases} 0 & \text{if } i \neq k \\ \sigma^2 & \text{if } i = k \end{cases}$$

Hence, samples from the same beach are assumed to be independent. For the random intercept model 5, we have:

$$Var(Y_{ij}) = Var(a_j + \varepsilon_{ij}) = \sigma_a^2 + \sigma^2$$
$$Cov(Y_{ij}, Y_{kj}) = Cov(a_j + \varepsilon_{ij}, a_j + \varepsilon_{kj}) = \begin{cases} \sigma_a^2 & \text{if } i \neq k \\ \sigma_a^2 + \sigma^2 & \text{if } i = k \end{cases}$$

Hence, the variance is the same for all samples on a particular beach j. Furthermore, for two different samples i and k on the same beach j, the correlation is equal to (correlation is covariance divided by variance):

$$Cor(Y_{ij}, Y_{kj}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2} \tag{8.1}$$

So, whatever the value of the explanatory variables, the correlation between two different observations from the same beach is defined by the formula in equation (8.1). This type of correlation is also called a compound symmetry structure. Now, let us have a look at the random intercept and slope model 6.

$$Var(Y_{ij}) = Var(a_j + b_j NAP_{ij} + \varepsilon_{ij}) = \sigma_a^2 + 2\,\mathrm{cov}(a_j, b_j) NAP_{ij} + \sigma_b^2 NAP_{ij}^2 + \sigma^2$$

This formula looks a bit more intimidating, but it basically states that the variance of a particular observation depends on the explanatory variables. The same holds for the covariance between two different observations from the same beach. The formula is given by:

$$Cov(Y_{ij}, Y_{kj}) = \sigma_a^2 + \mathrm{cov}(a_j, b_j)(NAP_{ij} + NAP_{kj}) + \sigma_b^2 NAP_{ij} NAP_{kj}$$

This formula also shows why allowing for covariance between a_j and b_j can be useful.

8.5 Another mixed modelling example

The RIKZ example showed how the number of regression parameters in a model can be reduced by using random components instead of fixed components. We now show how mixed modelling can be used to analyse a short time series using some bee data as an example. A full analysis of these data is presented in Chapter 22, but for this exercise we only consider the data for honeybees (foraging on sunflower crops) at five different locations (transects). All locations were sampled over the same ten days.

Figure 8.3 shows a coplot of the bee data; numbers of bees are plotted versus time conditional on location. To aid visual interpretation, a regression line was added for each location, and there seems to be a general downwards trend over time.

Just as for the RIKZ data we apply a mixed model to replace the five regression curves by one curve, plus allowing for random variation in both the intercept and slope. This gives a model of the form:

Model 9

$$Bees_{ij} = \alpha + \beta \mathrm{Time}_{ij} + a_j + b_j \mathrm{Time}_{ij} + \varepsilon_{ij}$$
$$a_j \sim N(0, \sigma_a^2), \quad b_j \sim N(0, \sigma_b^2) \quad \text{and} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

where j = 1, .., 5 as there are five locations, and i = 1, ..,1 0 (10 days). The component $\alpha + \beta \mathrm{Time}_{ij}$ is the fixed component and $a_j + b_j \mathrm{Time}_{ij}$ is the random component ensuring variation around the intercept and slope.

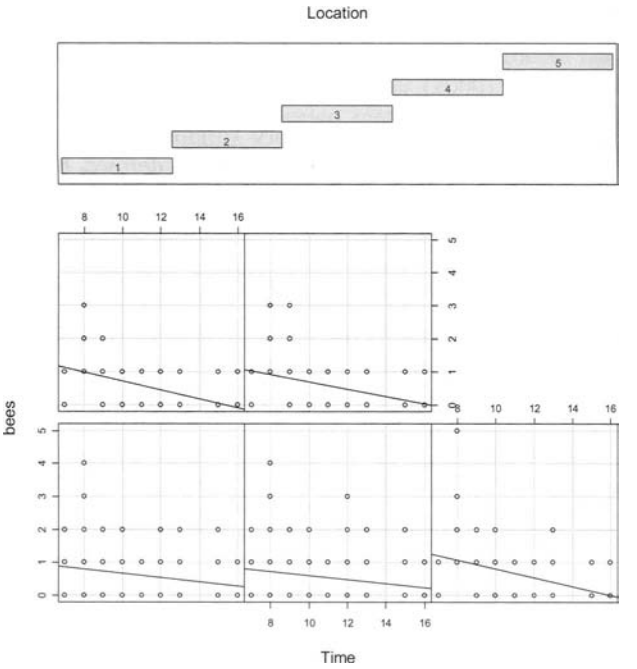

Figure 8.3. Coplot of bee numbers versus time conditional on location (transect). In each panel a linear regression line was added. The lower left panel shows the relationship between bees and time at location 1, and the upper right at location 5.

One of the assumptions in model 9 is that the errors ε_{ij} are independent. Hence, the residuals of two sequential days are not correlated to each other:

$$cor(\varepsilon_{ij}, \varepsilon_{i+1,j}) = 0$$

However, the observations are made sequentially over time and therefore we may be violating the independence assumption. One option to deal with this is by adding a correlation structure on the errors ε_{ij}:

$$cor(\varepsilon_{ij}, \varepsilon_{i+h,j}) = \rho_k$$

The problem is now how to model this correlation structure, and there are several ways of doing this. One option is to assume that the noise at day i is related to the noise at day $i - 1$, i - 2, etc. This is a so-called an auto-regressive model of order p:

$$\varepsilon_{ij} = \phi_1 \varepsilon_{i-1,j} + \phi_2 \varepsilon_{i-2,j} + ... + \phi_p \varepsilon_{i-p,j} + \eta_{ij}$$

The error term η_{ij} is independently normally distributed. The notation for this model is 'AR(p)' and the model allows for auto-correlation between the residuals. The coefficients $\phi_1,...,\phi_p$ are all smaller than 1 in the absolute sense. The unexplained information on day i is modelled as a function of the unexplained information on day $i-1$, $i-2$, etc. The AR(1) model is given by

$$\varepsilon_{ij} = \phi_1 \varepsilon_{i-1,j} + \eta_{ij}$$

Using some basic mathematics, it can be shown that the correlation between ε_{ij} and $\varepsilon_{i-k,j}$ is given by ϕ_1^k. This means that the further away two days are, the lower their correlation; the correlation between day i and $i-1$ is ϕ_1, between day i and $i-2$ is ϕ_1^2, etc. An auto-correlation function (Chapter 16) can be used to assess this assumption. For the bee data, an AR(1) structure seems to be a sensible choice, as what happens today is closely related to what happened yesterday, but much less to what happened two days ago. Another option is the so-called 'compound symmetry'. In this case the correlation between two error components is modelled as

$$cor(\varepsilon_{ij}, \varepsilon_{i-k,j}) = \rho$$

By definition if $k = 0$, the correlation is 1. This modelling approach assumes that the correlation between two days is always equal to ρ however far apart the days are. So, the correlation between the errors at time i and $i-k$ is the same whatever the value of k. Yet another option is to assume a general correlation structure. All days separated with k days have the same correlation ρ_k:

$$cor(\varepsilon_{ij}, \varepsilon_{i-k,j}) = \rho_k$$

All these options can be compared with each other using the AIC. We use the bee data to demonstrate mixed modelling and temporal auto-correlation. First, we fit the model with a random intercept and fixed slope but with no auto-correlation on the errors. The relevant output is given below.

AIC	BIC	logLik
2705.759	2726.308	-1348.880

Random effects:

	Intercept	Residual
StdDev:	0.013	0.702

Fixed effects:

	Value	Std.Error	df	*t*-value	*p*-value
Intercept	1.583	0.079	1254	19.934	<0.001
Time	-0.090	0.007	1254	-13.171	<0.001

The fixed effects are significantly differently from 0 at the 5% level. Refitting the model but specifying an AR(1) structure on the residuals gives:

AIC	BIC	logLik
2583.155	2608.841	-1286.577

Random effects:

	Intercept	Residual
StdDev:	0.004	0.703

Correlation Structure: AR(1)
Parameter estimate(s):
Phi
0.308

Fixed effects:

	Value	Std.Error	df	*t*-value	*p*-value
(Intercept)	1.577	0.108	1254	14.562	<0.001
Time	-0.090	0.009	1254	-9.585	<0.001

The coefficient ϕ_1 is estimated as 0.308, indicating that the correlation (in the error) between day i and $i-1$ is 0.308. And between days i and $i-1$ it is 0.308^2. To assess whether the AR(1) structure has improved the model, the likelihood ratio test can be used (note that this is not a boundary problem), or AIC values can be compared:

Model	df	AIC	BIC	logLik	*L*-Ratio	*p*-value
9	4	2705.76	2726.31	-1348.88		
9+AR(1)	5	2583.15	2608.84	-1286.57	124.604	<0.001

Both the AIC and the likelihood ratio test indicate that the second model (including the AR structure) is more optimal. Other options we tried were the compound symmetry correlation, but the AIC was 2707.758. In Chapter 22 additional explanatory variables and random components are used for the full analysis. Further discussion on auto-correlation and examples are given in Chapters 16, 26, 36 and 37.

8.6 Additive mixed modelling

Recall that in Section 8.3 we ended up with the following mixed model for the RIKZ data:

Model 7
$$Y_{ij} = \alpha + \beta \mathrm{NAP}_{ij} + a_j + b_j NAP_{ij} + \varepsilon_{ij}$$
where $a_j \sim N(0, \sigma_a^2)$, $b_j \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$

It models the species richness as a linear function of NAP, with the intercept and slope varying randomly across the beaches. This model also dealt with the violation of homogeneity by allowing for different variation in the noise per beach. Although this is a complicated model, it is still fundamentally a linear relationship between richness and NAP. And, in the same way as the linear regression

model was extended to additive models, we can extend mixed models into additive mixed models (Wood 2004, 2006; Ruppert et al. 2003). Two possible models are:

Model 10

$$Y_{ij} = \alpha + f(\text{NAP}_{ij}) + a_j + \varepsilon_{ij}$$

where $a_j \sim N(0, \sigma_a^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$

Model 11

$$Y_{ij} = \alpha + f(\text{NAP}_{ij}) + a_j + \varepsilon_{ij}$$

where $a_j \sim N(0, \sigma_a^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$

The linear component βNAP has been replaced by a smoothing function of NAP, denoted by f(NAP). The difference between models 10 and 11 is that the latter allows for different spread of the residuals per beach. So, the only difference between models 9 and 11 is the way the NAP effect is modelled. In model 9, we used a linear component plus random variation around the slope. In model 10, we have a smoothing function but we cannot have random variation around this smoother. The smoother itself can be estimated using, for example splines, and cross-validation can be used to estimate the optimal amount of smoothing. In model 11 we allow for different spread of the data per beach (heterogeneity). Both models 10 and 11 were applied on the RIKZ data and the AIC can be used to choose which one is best. In this example the AIC for model 11 was 211.71 and the AIC for model 10 was versus 249.52, suggesting model 11 to be the more optimal. The smoothing function for NAP is given in Figure 8.4.

In the same way as the linear regression model was extended to generalised linear modelling to analyse count data, presence–absence data or proportional data, so can the mixed model be extended to generalised linear mixed modelling. It is also possible to add auto-correlation structure to these methods. However, generalised linear mixed modelling and generalised additive mixed modelling are outside the scope of this book, and the interested reader is referred to Wood (2006) and Ruppert et al. (2003).

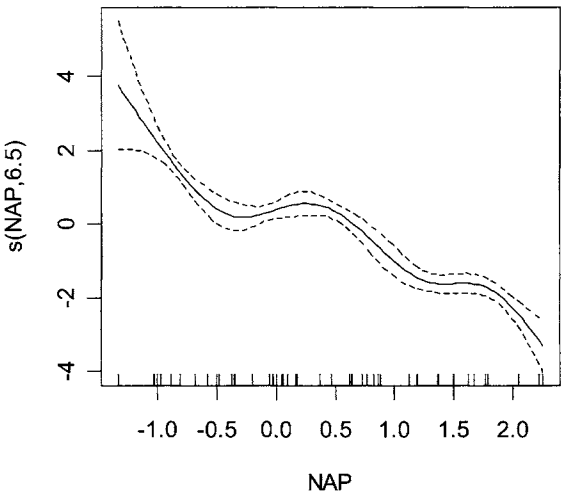

Figure 8.4. Smoothing function for NAP obtained by the additive mixed model 11. The amount of smoothing was estimated using cross-validation.