Chapter 5
Mixed Effects Modelling for Nested Data

In this chapter, we continue with Gaussian linear and additive mixed modelling methods and discuss their application on nested data. Nested data is also referred to as hierarchical data or multilevel data in other scientific fields (Snijders and Boskers, 1999; Raudenbush and Bryk, 2002).

In the first section of this chapter, we give an outline to mixed effects models for nested data before moving on to a formal introduction in the second section. Several different types of mixed effects models are presented, followed by a section discussing the induced correlation structure between observations. Maximum likelihood and restricted maximum likelihood estimation methods are discussed in Section 5.6. The material presented in Section 5.6 is more technical, and you need only skim through it if you are not interested in the mathematical details. Model selection and model validation tools are presented in Sections 5.7, 5.8, and 5.9. A detailed example is presented in Section 5.10.

5.1 Introduction

Zuur et al. (2007) used marine benthic data from nine inter-tidal areas along the Dutch coast. The data were collected by the Dutch institute RIKZ in the summer of 2002. In each inter-tidal area (denoted by 'beach'), five samples were taken, and the macro-fauna and abiotic variables were measured. Zuur et al. (2007) used species richness (the number of different species) and NAP (the height of a sampling station compared to mean tidal level) from these data to illustrate statistical methods like linear regression and mixed effects modelling. Here, we use the same data, but from a slightly different pedagogical angle. Mixed modelling may not be the optimal statistical technique to analyse these data, but it is a useful data set for our purposes. It is relatively small, and it shows all the characteristics of a data set that needs a mixed effects modelling approach.

The underlying question for these data is whether there is a relationship between species richness, exposure, and NAP. Exposure is an index composed of the following elements: wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer.

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_5,

As species richness is a count (number of different species), a generalised linear model (GLM) with a Poisson distribution may be appropriate. However, we want to keep things simple for now; so we begin with a linear regression model with the Gaussian distribution and leave using Poisson GLMs until later. A first candidate model for the data is

$$R_{ij} = \alpha + \beta_1 \times NAP_{ij} + \beta_2 \times Exposure_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{5.1}$$

R_{ij} is the species richness at site j on beach i, NAP_{ij} the corresponding NAP value, $Exposure_i$ the exposure on beach i, and ε_{ij} the unexplained information. Indeed, this is the familiar linear regression model. The explanatory variable Exposure is nominal and has two[1] classes. However, as we have five sites per beach, the richness values at these five sites are likely to be more related to each other than to the richness values from sites on different beaches. The linear regression model does not take this relatedness into account. The nested structure of the data is visualised in Fig. 5.1.

Many books introduce mixed effects modelling by first presenting an easy to understand technique called 2-stage analysis, conclude that it is not optimal, and then present the underlying model for mixed effects modelling by combining the 2 stages into a single model (e.g. Fitzmaurice et al., 2004). This is a useful way to introduce mixed effects modelling, and we also start with the 2-stage analysis method before moving onto mixed effects modelling.

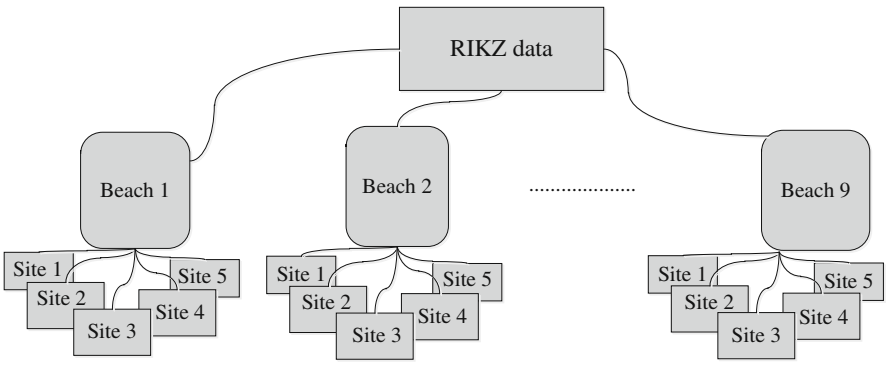

Fig. 5.1 Set up of the RIKZ data. Measurements were taken on 9 beaches, and on each beach 5 sites were sampled. Richness values at sites on the same beach are likely to be more similar to each other than to values from different beaches

[1] Originally, this variable had three classes, but because the lowest level was only observed on one beach, we relabeled, and grouped the two lowest levels into one level called 'a'. The highest level is labeled 'b'.

5.2 2-Stage Analysis Method

In the first step of the 2-stage analysis method, a linear regression model is applied on data of one beach. It models the relationship between species richness and NAP on each beach using

$$R_{ij} = \alpha + \beta_i \times NAP_{ij} + \varepsilon_{ij} \qquad j = 1, \ldots, 5 \tag{5.2}$$

This process is then carried out for data of each beach in turn. In a more abstract matrix notation, we can write the model for the data of beach i as

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i1} \\ 1 & NAP_{i1} \\ 1 & NAP_{i1} \\ 1 & NAP_{i1} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix} \Leftrightarrow \mathbf{R}_i = \mathbf{Z}_i \times \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \tag{5.3}$$

$\mathbf{R}_i$ is now a vector of length 5 containing the species richness values of the 5 sites on beach i: R_{i1} to R_{i5}. The first column of $\mathbf{Z}_i$ contains ones and models the intercept, and the second column contains the five NAP values on beach i. The unknown vector $\boldsymbol{\beta}_i$ contains the regression parameters (intercept and slope) for beach i. This general matrix notation allows for different numbers of observations per beach as the dimension of $\mathbf{R}_i$, $\mathbf{Z}_i$, and $\boldsymbol{\varepsilon}_i$ can easily be adjusted. For example, if beach $i = 2$ has 4 observations instead of 5, $\mathbf{Z}_2$ contains 4 rows and 2 columns, but we still obtain an estimate for the intercept and slope. In this case, Equation (5.3) takes the form

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i5} \end{pmatrix}$$

The model in Equation (5.3) is applied on data of each beach, resulting in nine estimated values for the slope and intercept. The following loop gives the results in the R software.

```
> library(AED); data(RIKZ)
> Beta <- vector(length = 9)
> for (i in 1:9){
      Mi <- summary(lm(Richness ~ NAP,
                       subset = (Beach==i), data=RIKZ))
     Beta[i] <- Mi$coefficients[2, 1]}
```

The `subset` option in the linear regression function `lm` ensures that data from each beach are analysed in a particular iteration of the loop. The last line in the loop

extracts and stores the slope for NAP for each regression analysis. The estimated betas can be obtained by typing `Beta` in R:

```
-0.37 -4.17 -1.75 -1.24 -8.90 -1.38 -1.51 -1.89 -2.96
```

Note that there are considerably differences in the nine estimated slopes for NAP. Instead of the loop in the code above, you can also use the `lmList` command from the `nlme` package to produce the same results. This option also gives a nice graphical presentation of estimated intercepts and slopes (Pinheiro and Bates, 2000).

In the second step, the estimated regression coefficients are modelled as a function of exposure.

$$\hat{\beta}_i = \eta + \tau \times Exposure_i + b_i \qquad i = 1, \ldots, 9 \tag{5.4}$$

This is 'just' a one-way ANOVA. The response variable is the estimated slopes from step 1, Exposure is the (nominal) explanatory variable, τ is the corresponding regression parameter, η is the intercept, and b_i is random noise. The matrix notation for this is below. It looks intimidating, but this is only because exposure is a factor with levels 0 and 1. Level 0 is used as the baseline. The model in Equation (5.4) is written in matrix notation as

$$\begin{pmatrix} -0.37 \\ -4.17 \\ -1.75 \\ -1.24 \\ -8.90 \\ -1.38 \\ -1.51 \\ -1.89 \\ -2.96 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} \eta \\ \tau \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \\ b_8 \\ b_9 \end{pmatrix} \Leftrightarrow \hat{\boldsymbol{\beta}}_i = \mathbf{K}_i \times \boldsymbol{\gamma} + \mathbf{b}_i \qquad i = 1, \ldots, 9 \tag{5.5}$$

The vector $\boldsymbol{\gamma}$ contains the intercept η and slope τ and is not the same thing as $\boldsymbol{\beta}_i$. The following R code was used to apply this model.

```
> fExposure9 <- factor(c(0, 0, 1, 1, 0, 1, 1, 0, 0))
> tmp2 <- lm(Beta ~ fExposure9))
```

As we already mentioned, this linear regression model is also called a one-way analysis of variance (ANOVA). The results of the `anova` command are not presented here, but it shows that the p-value for exposure is 0.22, indicating that there is no significant exposure effect on the nine slopes.

The two formulae of the 2-stage approach are repeated in Equation (5.6).

$$\begin{aligned} \mathbf{R}_i &= \mathbf{Z}_i \times \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \\ \hat{\boldsymbol{\beta}}_i &= \mathbf{K}_i \times \boldsymbol{\gamma} + \mathbf{b}_i \end{aligned} \tag{5.6}$$

It is common to assume that the residuals $\mathbf{b}_i$ are normally distributed with mean 0 and variance $\mathbf{D}$. The second step of the two-stage analysis can be seen as an analysis of a summary statistic; in this case, it is the slope representing the strength of the relationship between species richness and NAP on a beach. The two-stage analysis has various disadvantages. Firstly, we summarise all the data from a beach with one parameter. Secondly, in the second step, we analyse regression parameters, not the observed data. Hence, we are not modelling the variable of interest directly. Finally, the number of observations used to calculate the summary statistic is not used in the second step. In this case, we had five observations for each beach. But if you have 5, 50, or 50,000 observations, you still end up with only one summary statistic.

5.3 The Linear Mixed Effects Model

5.3.1 Introduction

The linear mixed effects model combines both the earlier steps into a single model.

$$\mathbf{R}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{5.7}$$

As before, $\mathbf{R}_i$ contains the richness values for beach i, $i = 1, \ldots, 9$. There are two components in this model that contain explanatory variables; the fixed $\mathbf{X}_i \times \boldsymbol{\beta}$ term and the random $\mathbf{Z}_i \times \mathbf{b}_i$ term. Because we have a fixed and a random component, we call the model a *mixed* effects model. We discuss later how to fill in the $\mathbf{X}_i$ and $\mathbf{Z}_i$. In this case, the $\mathbf{Z}_i \times \mathbf{b}_i$ component represents the Richness–NAP effect for each beach; each beach is allowed to have a different Richness–NAP relationship because there is an index i attached to $\mathbf{b}$. There is no index attached to the parameter $\boldsymbol{\beta}$; hence, it is for all beaches. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices of dimension $n_i \times p$ and $n_i \times q$, respectively, where n_i is the number of observations in $\mathbf{R}_i$ (the number of observations per beach), p the number of explanatory variables in $\mathbf{X}_i$, and q the number of explanatory variables in $\mathbf{Z}_i$.

Many textbooks on linear mixed effects modelling are orientated towards medical science, where i is typically denoted as 'subject' because it represents a patient or person. The component $\mathbf{Z}_i \times \mathbf{b}_i$ is then the subject specific or random effect and $\mathbf{X}_i \times \boldsymbol{\beta}$ the overall or fixed component. The matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ may, or may not, contain the same explanatory variables. This depends on what type of model is fitted. Because the model in Equation (5.7) forms the basis of much of the material to come, we present it one more time, but now with all the assumptions.

$$\begin{aligned}
&\mathbf{Y}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \boldsymbol{\varepsilon}_i \\
&\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D}) \\
&\boldsymbol{\varepsilon}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_i) \\
&\mathbf{b}_1, \cdots \mathbf{b}_\mathrm{N}, \boldsymbol{\varepsilon}_1, \cdots, \boldsymbol{\varepsilon}_\mathrm{N} \text{ independent}
\end{aligned} \tag{5.8}$$

This model is also called the Laird and Ware model formulation after a paper by these two authors in 1982. It is fundamentally important that you understand the model formulation, and therefore we give three examples before continuing with more details. These are the random intercept model, the random intercept and slope model, and the random effects model.

5.3.2 The Random Intercept Model

Suppose we model species richness as a linear function of NAP where the intercept is allowed to change per beach. Within linear regression, we can model this as

$$R_{ij} = \alpha + \beta_1 \times Beach_i + \beta_2 \times NAP_{ij} + \varepsilon_{ij} \tag{5.9}$$

$Beach_i$ is a factor with nine levels, and the first level is used as baseline. The price we pay for including this term is eight regression parameters (which will cost 8 degrees of freedom). However, perhaps we are not interested in knowing the exact nature of the beach effect. In that case, eight regression parameters is a high price! One option is to use beach as a random effect. This means that we include a beach effect in the model, but we assume that the variation around the intercept, for each beach, is normally distributed with a certain variance. A small variance means that differences per beach (in terms of the intercept) are small, whereas a large variance allows for more variation. Such a mixed effects model is defined as follows.

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \times b_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix} \Leftrightarrow \mathbf{R}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i \tag{5.10}$$

In this example, five observations are taken on each beach, hence $n_i = 5$ for all i. Therefore, $\mathbf{Z}_i$ is a matrix of dimension 5×1 containing only ones. Now let us have a look at the assumptions. The first assumption is that the random effects b_i are normally distributed: N(0, d^2). The second assumption is that the errors $\boldsymbol{\varepsilon}_i$ (containing the five errors $\boldsymbol{\varepsilon}_{i1}$ to $\boldsymbol{\varepsilon}_{i5}$) are normally distributed with covariance matrix Σ_i. The easiest option is to assume

$$\Sigma_i = \sigma^2 \times \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \vdots \\ \vdots & 0 & 1 & 0 & \vdots \\ \vdots & \vdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

In general, the elements of $\mathbf{\Sigma}_i$ do not depend on i, but this may not always be the case (Verbeke and Molenberghs, 2000; Pinheiro and Bates, 2000). In Chapter 4, we discussed various methods to incorporate heterogeneity into the model, and these will influence the structure of $\mathbf{\Sigma}_i$. But for the moment, we ignore these methods. To apply the random intercept model in R, we need the following code.

```
> library(nlme)
> RIKZ$fBeach <- factor(RIKZ$Beach)
> Mlme1 <- lme(Richness ~ NAP, random = ~1 | fBeach,
               data = RIKZ)
> summary(Mlme1)
```

The mixed effects model is applied using the function `lme`, which stands for linear mixed effects model. The difference with the `lm` command for linear regression is that in the `lme` function, we need to specify the random component. The `~1 |fBeach` bit specifies a random intercept model. The argument on the right hand side of the '|' sign is a nominal variable. The relevant output from the `summary` command is given below.

```
Linear mixed-effects model fit by REML
  AIC       BIC     logLik
  247.48    254.52  -119.74

Random effects:
 Formula: ~1 | fBeach
            (Intercept)     Residual
StdDev:       2.944          3.059

Fixed effects: Richness ~ NAP
             Value Std.Error DF   t-value p-value
(Intercept)  6.58  1.09      35    6.00    <0.001
NAP         -2.56  0.49      35   -5.19    <0.001
```

The first part of the output gives the AIC and BIC. Their definitions and examples on how to use them are given later in this chapter. For the moment, it is sufficient to know that we are using them just as in linear regression to help with model selection. The remaining part of the output is split up in random effects and fixed effects. The residual variance σ^2 is estimated as $3.05^2 = 9.30$, and the variance for the random intercept d^2 is estimated as $2.94^2 = 8.64$. We should not say that $d = 2.94$, as d is a population parameter, and the value of 2.94 is an estimator for it. It is better to put a hat on d, and say that the estimated value for d is

$$\hat{d} = 2.94.$$

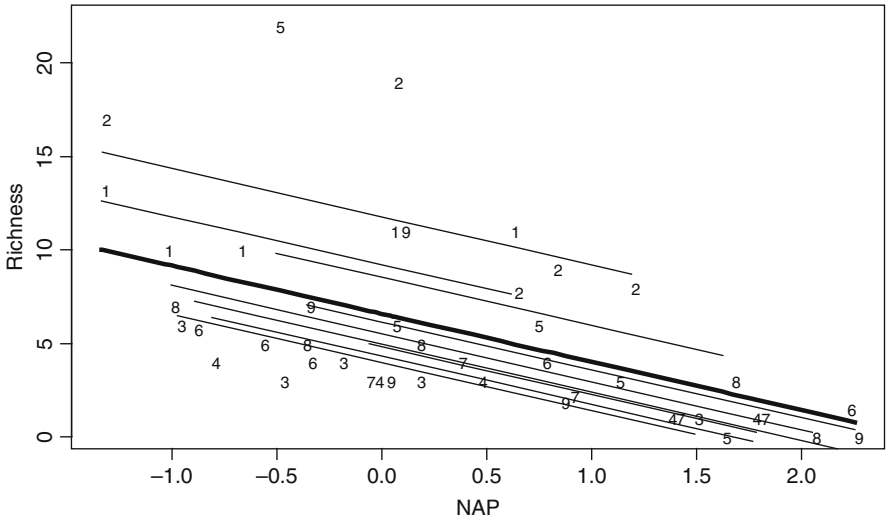

Fig. 5.2 Fitted values obtained by mixed effects modelling. The thick line represents the fitted values for the population and is specified by $6.58 - 2.56 \times NAP_i$, whereas the other lines are obtained by $\mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i$. Numbers represent the beaches

The fixed effects part shows that the intercept α is 6.58 and the slope β is –2.56 (again, we should put hats on parameters as both are estimators). Both parameters are significantly different from 0 at the 5% level. We discuss later how degrees of freedom are obtained.

All this information may look wonderful but what does it mean? The best way to answer this is to plot the fitted values. This raises the question: What are the fitted values? There are two options. We can either consider $\mathbf{X}_i \times \boldsymbol{\beta}$ as the fitted values (again, we should put a hat on the $\boldsymbol{\beta}$), which is $6.58 - 2.56 \times NAP_i$ or use $\mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i$ as the fitted values. Both types of fitted values are presented in Fig. 5.2.

The thick line represents the fitted line obtained by the fixed component $6.58 - 2.56\, NAP_i$, also called the population model. The other lines are obtained by adding the contribution of $\mathbf{b}_i$ for each beach i to the population fitted curve. Hence, the random intercept model implies one average curve (the thick line) that is allowed to be shifted up, or down, for each beach by something that is normally distributed with a certain variance d^2. If d^2 is large, the vertical shifts will be relative large. If $d^2 = 0$, all shifts are zero and coincide with the thick line. The following R code was used to generate Fig. 5.2.

```
> F0 <- fitted(Mlme1, level = 0)
> F1 <- fitted(Mlme1, level = 1)
> I <- order(RIKZ$NAP); NAPs <- sort(RIKZ$NAP)
> plot(NAPs, F0[I], lwd = 4, type = "l",
       ylim = c(0, 22), ylab = "Richness", xlab = "NAP")
```

```
> for (i in 1:9){
    x1 <- RIKZ$NAP[RIKZ$Beach == i]
    y1 <- F1[RIKZ$Beach == i]
    K <- order(x1)
    lines(sort(x1), y1[K])
}
> text(RIKZ$NAP, RIKZ$Richness, RIKZ$Beach, cex = 0.9)
```

The `fitted` command takes as argument the object from the function `lme` plus a level argument. The `level = 0` option means that we take the fitted values obtained by the population model, whereas `level = 1` gives the *within-beach* fitted values. The `order` and `sort` commands avoid spaghetti plots, and the loop draws the nine lines in the same plot as the population curve.

5.3.3 The Random Intercept and Slope Model

The model in Equation (5.10) allows for a random shift around the intercept resulting in fitted lines parallel to the population fitted line (Fig. 5.2). This immediately raises the question whether we can use the same trick for the slope. The answer is yes, but before showing the model and the R code, we first discuss why we want to do this.

Suppose that the relationship between species richness and NAP is different on each beach. This implies that we need to include a NAP–Beach interaction term to the model. Such a model is specified by: R_i = factor(Beach) + NAP × factor(Beach). This is a linear regression model with one nominal variable, one continuous variable, and an interaction between them. A different name is an analysis of covariance (ANCOVA). Because beach has nine levels and one level is used as the baseline, the number of parameters used by this model is excessively high, at 17. And we are not even interested in beach effects! But if there is any between beach variation and a NAP–Beach interaction, then we cannot ignore these terms. If we do, this systematic variation ends up in the residuals, leading to potentially biased inference. To estimate model degrees of freedom more efficiently, we can apply the mixed effects model with a random intercept (as before) *and* a random slope. The required R code is a simple extension of the code we used for the random intercept model.

```
> Mlme2 <- lme(Richness ~ NAP,
               random = ~1 + NAP | fBeach, data = RIKZ)
> summary(Mlme2)

Linear mixed-effects model fit by REML
AIC         BIC         logLik
244.38      254.95      -116.19
```

```
Random effects:
 Formula: ~1 + NAP | fBeach
                  StdDev       Corr
(Intercept)       3.549       (Intr)
NAP               1.714       -0.99
Residual          2.702

Fixed effects: Richness ~ NAP
             Value  Std.Error  DF  t-value p-value
(Intercept)  6.58       1.26   35   5.20     <0.001
NAP          -2.83      0.72   35  -3.91     <0.001
```

Later we discuss how to compare the two models (random intercept and random intercept and slope models) and how to judge which one is better. For the moment, it is sufficient to note that the random intercept and slope model has a lower AIC than the earlier models (the lower the AIC, the better). Later in this chapter, we also dedicate an entire section to the phrase 'Linear mixed-effects model fit by REML'.

The random effects part now has three standard errors and one correlation term. The model that we are fitting is of the form

$$\begin{pmatrix} R_{i1} \\ R_{i2} \\ R_{i3} \\ R_{i4} \\ R_{i5} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix} \tag{5.11}$$

The only difference with this model, compared to the one in Equation (5.10), is the modification of the matrix $\mathbf{Z}_i$; NAP values for beach i have be included. As a result, $\mathbf{b}_i$ is now of dimension 2×1, and its distribution is given by

$$\begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{D}) \qquad \text{where} \qquad \mathbf{D} = \begin{pmatrix} d_{11}^2 & d_{12} \\ d_{12} & d_{22}^2 \end{pmatrix} \tag{5.12}$$

The variance d_{11}^2 plays the same role as d^2 in the random intercept model; it determines the amount of variation around the population intercept α. The numerical output shows that its estimated value is $3.54^2 = 12.5$. The model also allows for random variation around the population slope in a similar way as it does for the intercept. The variance ${d_{22}}^2$ determines the variation in slopes at the nine beaches. The estimated value of $1.71^2 = 2.92$ shows that there is considerably more variation in intercepts than in slopes at the nine beaches. Finally, there is a correlation between the random intercepts and slopes. Its value of –0.99 is rather high (causing potential numerical problems), but indicates that beaches with a high positive intercept also have a high negative slope. This can also be seen from the fitted values in Fig. 5.3.

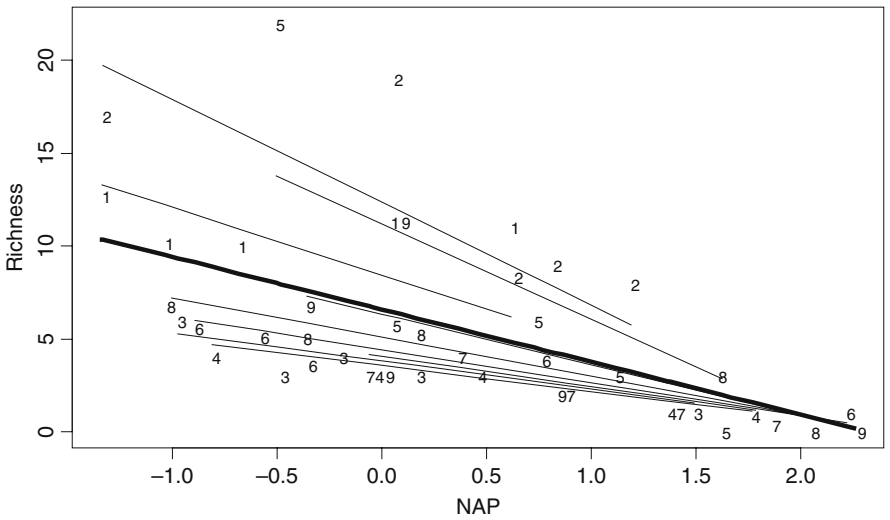

Fig. 5.3 Fitted values obtained by the random intercept and slope model. The thick line represents the fitted values for the population, and the other lines represent the so-called within-group fitted curves. Numbers represent beaches

The thick line is the fitted population curve, and the other lines the within-beach fitted curves. Note the difference with Fig. 5.2.

5.3.4 Random Effects Model

A linear mixed effects model that does not contain any $\boldsymbol{\beta}$, except for an intercept is called a random effects model. By dropping the NAP variable in Equation (5.9), we obtain the following random effects model.

$$R_i = \alpha + b_i + \varepsilon_i$$

The term b_i is normally distributed with mean 0 and variance d^2; ε_i is normally distributed with mean 0 and variance σ^2. The index i runs from 1 to 9. The model implies that richness is modelled as an intercept plus a random term b_i that is allowed to differ per beach. The R code to run this model is

```
> Mlme3 <- lme(Richness ~ 1, random = ~1 | fBeach,
            data = RIKZ)
```

The output of the `summary(Mlme3)` command is given below. The estimated values for d and σ are 3.23 and 3.93, respectively.

```
Linear mixed-effects model fit by REML
       AIC       BIC               logLik
  267.11         272.46            -130.55

Random effects:
            (Intercept)      Residual
StdDev:      3.23            3.93

Fixed effects:
             Value  Std.Error  DF  t-value p-value
(Intercept)  5.68   1.22       36  4.63    <0.001
```

Later in this chapter, we discuss how to choose between a random effects model, random intercept model, and random intercept plus slope model. There are also several other issues that we need to discuss such as: What is the correlation between richness values measured at the same beach and measured at different beaches? How do we estimate the parameters? How do we find the optimal model? Finally, once an optimal model has been identified, how do we then validate it? Each of these points is discussed next.

5.4 Induced Correlations

Returning to the RIKZ data discussed earlier in this chapter, we modelled species richness as a function of NAP and a random intercept. The question we now address is: What is the correlation between two observations from the same beach, and from different beaches? To answer this question, we first need to find an expression for the covariance matrix of $\mathbf{Y}_i$. The mathematical notation that we have used so far for the model was $\mathbf{Y}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \mathbf{Z}_i \times \mathbf{b}_i + \boldsymbol{\varepsilon}_i$. This is also called the hierarchical model. The underlying assumptions of this model were given in Equation (5.8). We now derive an expression for the covariance matrix of the $\mathbf{Y}_i$. It is relatively easy to show that $\mathbf{V}_i$ is normally distributed with mean $\mathbf{X}_i \times \boldsymbol{\beta}$ and variance $\mathbf{V}_i$ in mathematical notation:

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \times \boldsymbol{\beta}, \mathbf{V}_i) \qquad \text{where } \mathbf{V}_i = \mathbf{Z}_i \times \mathbf{D} \times \mathbf{Z}_i' + \boldsymbol{\Sigma}_i \tag{5.13}$$

Recall that $\mathbf{D}$ was the covariance matrix of the random effects. So, including random effects has an effect on the structure of the covariance matrix $\mathbf{V}_i$. To illustrate this, we discuss the random intercept model for the RIKZ data we presented in the previous section.

For the random intercept model, $\mathbf{Z}_i$ is a vector of length five containing ones and $\boldsymbol{\Sigma}_i = \sigma^2 \times \mathbf{I}_{5\times 5}$ is a diagonal matrix of dimension 5×5. $\mathbf{I}_{5\times 5}$ is an identity matrix with 5 rows and 5 columns; it has ones on the diagonal and zeros elsewhere. As a result we have

$$\mathbf{V}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \times d^2 \times \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix} + \sigma^2 \times \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 + d^2 & d^2 & d^2 & d^2 & d^2 \\ d^2 & \sigma^2 + d^2 & d^2 & d^2 & d^2 \\ d^2 & d^2 & \sigma^2 + d^2 & d^2 & d^2 \\ d^2 & d^2 & d^2 & \sigma^2 + d^2 & d^2 \\ d^2 & d^2 & d^2 & d^2 & \sigma^2 + d^2 \end{pmatrix}$$

Showing that, the covariance between any two sites on the same beach is d^2, and the variance is $d^2 + \sigma^2$. By definition, the correlation between two observations from the same beach is $d^2/(d^2 + \sigma^2)$. This is irrespective of the identity of the beach (all the $\mathbf{V}_i$s are the same). This is called an induced correlation (or covariance) structure as we did not explicitly specify it. It is the consequence of the random effects structure. The results presented in Section 5.3 show that the estimated value for d is 2.944 and for σ it is 3.06. Giving an induced correlation of $2.94^2/(2.94^2 + 3.06^2) = 0.48$, which is relatively high. This correlation is also called the intraclass correlation and is further discussed at the end of this section.

As to the second question, the model implies that observations from different beaches are uncorrelated.

We can make things a bit more complicated by using the random intercept and slope model. In this case, we get

$$\mathbf{V}_i = \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} \times \begin{pmatrix} d_{11}^2 & d_{21} \\ d_{12} & d_{22}^2 \end{pmatrix} \times \begin{pmatrix} 1 & NAP_{i1} \\ 1 & NAP_{i2} \\ 1 & NAP_{i3} \\ 1 & NAP_{i4} \\ 1 & NAP_{i5} \end{pmatrix} + \sigma^2 \times \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

This is a bit more challenging, but it turns out that the variance of Y_{ij} and covariance of two observations from the same beach, Y_{ij} and Y_{ik}, are given by (Fitzmaurice et al., 2004)

$$\text{var}(Y_{ij}) = d_{11}^2 + 2 \times NAP_{ij} \times d_{12} + NAP_{ij}^2 \times d_{22}^2 + \sigma^2$$

$$\text{cov}(Y_{ij}, Y_{ik}) = d_{11}^2 + (NAP_{ij} + NAP_{ik}) \times d_{12} + NAP_{ij} \times NAP_{ik} \times d_{22}^2$$

This looks complicated, but it tells us that the variance and covariance of Y_{ij} depend not only on the variances and covariances of the random terms, but also on NAP. Fitzmaurice et al. (2004) used time instead of NAP. In that case, the variance and covariance depend on time.

5.4.1 Intraclass Correlation Coefficient

Although outside the scope of the underlying questions raised at the start of this section, it is useful to take some time interpreting the intraclass correlation as it can be used to determine appropriate sample sizes. It is also called the intraclass correlation (Snijders and Bosker, 1999). Recall that we have nine beaches, five observations per beach, and an intraclass correlation of 0.48. If we take a sample of a certain size, the standard error of the mean is given by

$$\text{standard error} = \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$$

Obviously, we want a small standard error and a large sample size may help achieve this as it is the denominator. In this case, we have a sample size of 45. However, these data are nested (hierarchical) and this should be taken somehow into account, especially of the correlation between observations on a beach is relative high. The design effect indicates how much the denominator should be adjusted. For a more formal definition, see Snijders and Bosker (1999). For a two-stage design with equal number of samples per beach ($n = 5$) and intraclass correlation ρ, the design effect is defined as

$$\text{design effect} = 1 + (n - 1) \times \rho = 1 + 4 \times 0.48 = 2.92$$

If this number is larger than 1, and in this case it is 2.92, we should not use 45 in the denominator for the standard error, but an adjusted sample size, also called the effective sample size, should be used. It is given by

$$N_{\text{effective}} = \frac{N \times n}{\text{design effect}} = \frac{9 \times 5}{2.92} = 15.41$$

A high intraclass correlation means that the corrected sample size is considerably lower, and this means less precise standard errors! At the end of the day, this makes sense; if observations on a beach are highly correlated, we cannot treat them as independent observations. Why then bother taking many observations per beach? Perhaps we should sample more beaches with fewer observations per beach? Further examples are given in Chapter 3 in Snijder and Bosker (2000).

5.5 The Marginal Model

In the previous section, we saw how including random effects induces a correlation structure between observations from the same beach. With the random intercept model, the induced correlation structure was fairly simple with the correlation between any two observations from the same beach given as $d^2/(d^2 + \sigma^2)$. Surprisingly, we can get the same correlation structure and estimated parameters in a

different way, and it does not contain any random effects. The model we use is the linear regression model $\mathbf{Y}_i = \mathbf{X}_i \times \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$; but instead of assuming that the five residuals of the same beach, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4}, \varepsilon_{i5})$ are independent of each other, we allow for dependence between them. This is done as follows. We start again with an expression for the covariance matrix of the $\mathbf{Y}_i$. Using the standard linear regression theory, it is easy to show that $\mathbf{Y}_i$ is normally distributed with mean $\mathbf{X}_i \times \boldsymbol{\beta}$ and variance $\mathbf{V}_\mathrm{i}$; in mathematical notation,

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \times \boldsymbol{\beta}, \mathbf{V}_i) \quad \text{where } \mathbf{V}_i = \boldsymbol{\Sigma}_i$$

Note that there is no covariance matrix $\mathbf{D}$ in $\mathbf{V}_i$ as there are no random effects in the model. In linear regression, we use $\boldsymbol{\Sigma}_i = \sigma^2 \times \mathbf{I}_{5\times5}$. $\mathbf{I}_{5\times5}$ is an identity matrix with 5 rows and 5 columns, implying independence between residuals (or observations) of the same beach i. The dependence structure is built in by allowing for non-zero off-diagonal elements in the covariance matrix. One option is the so-called *general correlation* matrix

$$\mathbf{V}_i = \boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma^2 & c_{21} & c_{31} & c_{41} & c_{51} \\ c_{21} & \sigma^2 & c_{32} & c_{42} & c_{52} \\ c_{31} & c_{32} & \sigma^2 & c_{43} & c_{53} \\ c_{41} & c_{42} & c_{43} & \sigma^2 & c_{54} \\ c_{54} & c_{52} & c_{53} & c_{54} & \sigma^2 \end{pmatrix}$$

Because the covariance between observations Y_{i1} and Y_{i2} is the same as that between observations Y_{i2} and Y_{i1}, the covariance matrix $\mathbf{V}_i$ is symmetric. So, in this example, we have to estimate 10 parameters (all the elements in the upper or lower diagonal). But for data sets with larger number of observations per beach, this number increases dramatically. Therefore, we can use more restrictive covariance matrices. The most restrictive correlation structure is the so-called *compound symmetric* structure defined by

$$\mathbf{V}_i = \Sigma_i = \begin{pmatrix} \sigma^2 & \varphi & \varphi & \varphi & \varphi \\ \varphi & \sigma^2 & \varphi & \varphi & \varphi \\ \varphi & \varphi & \sigma^2 & \varphi & \varphi \\ \varphi & \varphi & \varphi & \sigma^2 & \varphi \\ \varphi & \varphi & \varphi & \varphi & \sigma^2 \end{pmatrix}$$

In this case, there is only one unknown parameter, namely φ. So, the covariance between any two observations on the same beach i is given by φ. If it is estimated as 0, then we can assume independence. General correlation and compound symmetry correlation are the two most extreme correlation structures, and there are various intermediate structures that we will see later, which can be applied to spatial and temporal data.

The R code for the marginal model is given below. We also give the command for the equivalent random intercept mixed effects model.

```
> M.mixed <- lme(Richness ~ NAP, random = ~1 | fBeach,
                 method = "REML", data = RIKZ)
> M.gls <- gls(Richness ~ NAP, method = "REML",
          correlation = corCompSymm(form =~ 1 | fBeach),
          data = RIKZ)
```

The argument `corCompSymm(form =~ 1 | fBeach)` for the correlation option in the `gls` function tells R that all observations from the same beach are correlated. The `summary(M.mixed)` and `summary(M.gls)` commands give identical estimated parameters, standard errors, *t*-values, and *p*-values, and these are not shown here (see Section 5.3). We only show the relevant output of the GLS model.

```
Correlation Structure: Compound symmetry
 Formula: ~1 | factor(Beach)
 Parameter estimate(s):
      Rho
0.4807353
...
Residual standard error: 4.246141
```

The estimated Rho is φ divided by the estimated value of σ^2 ($= 4.25^2$) as we expressed $\mathbf{V}_i$ as a covariance matrix and not a correlation matrix.

There are also subtle differences between the hierarchical model and the marginal model with respect to the numerical estimation process (West et al., 2006).

5.6 Maximum Likelihood and REML Estimation*

When applying mixed effects modelling, you will see phrases like 'REML' and 'maximum likelihood' estimation. Unlike linear regression models, where you can get away with not knowing the underlying mathematics, there is no escaping some maths when using REML and maximum likelihood (ML) in mixed effects modelling. So, what does REML mean, and what does it do? The first question is easy; REML stands for restricted maximum likelihood estimation. As to the second question, most books at this point get rather technical or avoid the detail and only present REML as a mystical way to 'correct the degrees of freedom'. We have chosen to try and explain it in more detail and for this we need to use matrix algebra. But, to understand REML you need to first understand the principle of maximum likelihood estimation, and this is where we will begin. If you are not familiar with matrix algebra, or if the mathematical level in this section is too high, we still advise you to skim through this section before reading on.

We start by revising maximum likelihood for linear regression, and then show how REML is used to correct the estimator for the variance.

Assume we have a linear regression model $Y_i = \alpha + \beta \times X_i + \varepsilon_i$, where ε_i is normally distributed with mean 0 and variance σ^2. The unknown parameters in this model are α, β, and σ. Instead of writing these three variables all the time, we can refer to them as $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\alpha, \beta, \sigma)$. One option to estimate $\boldsymbol{\theta}$ is ordinary least squares. It gives an expression for each element of $\boldsymbol{\theta}$, see, for example, Montgomery and Peck (1992), among many other books on linear regression. The expression for the estimated variance obtained by linear regression is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} \times X_i)^2 \tag{5.14}$$

We have put a ˆ on the parameters to indicate that these are the estimated values, and n is the number of observations. It can be shown that $\hat{\sigma}$ is an unbiased estimator of σ; this means that $E[\hat{\sigma}] = \sigma$. Now let us have a look at the maximum likelihood estimation approach. We have used results from Section 2.10 in Montgomery and Peck (1992), which assume that Y_i is normally distributed and its density function is given by

$$f_i(Y_i, X_i, \alpha, \beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{\frac{(Y_i - \alpha - \beta \times X_i)^2}{2\sigma^2}} \tag{5.15}$$

Because we also assume that the Y_i are independent, we can write the joint density function for Y_1, Y_2, …,Y_n as a product of the individual density curves $f_1, f_2, \ldots, f_n$. This is called the likelihood function L. It is a function of the data and $\boldsymbol{\theta}$. The question is how to choose $\boldsymbol{\theta}$ such that L is the highest. To simplify the mathematics, the natural log is taken of L (The log converts the product of the density functions to a sum of log-density functions; it is easier to work with a sum than a product.), resulting in the following log-likelihood equation.

$$\ln L(Y_i, X_i, \alpha, \beta, \sigma) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \alpha - \beta \times X_i)^2 \tag{5.16}$$

We need to maximise this function with respect to α, β, and σ. This is a matter of taking partial derivates of L with respect to each of these parameters, setting them to zero, and solving the equations. It turns out that these equations give simple expressions for the estimators of α and β. Because we can easily calculate them, these equations are called *closed form* solutions. In the generalized linear mixed model chapters, we will see open form solutions, which means there is no direct solution for the parameters.

The formulae for the estimators of α and β are not given here, but for the variance we get

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} \times X_i)^2 \tag{5.17}$$

Note this is nearly the same expression as we found with ordinary least squares in Equation (5.14). In fact, the estimator for the variance obtained by maximum likelihood is biased by a factor $(n - 2)/n$. If the linear regression model contains p explanatory variables, the bias is $(n - p)/n$. The reason that the maximum likelihood estimator is biased is because it ignores the fact that the intercept and slope are estimated as well (as opposed to being known for certain). So, we need a mechanism that gives better ML estimators, and indeed this is what restricted maximum likelihood (REML) does.

REML works as follows. The linear regression model $Y_i = \alpha + \beta \times X_i + \varepsilon_i$ can be written as $Y_i = \mathbf{X}_i \times \boldsymbol{\beta} + \varepsilon_i$. This is based on simple matrix notation using $\mathbf{X}_\mathrm{i} = (1\ X_\mathrm{i})$, and the first element of $\boldsymbol{\beta}$ is the intercept and the second element is the original β. The normality assumption implies that

$$Y_i \sim N(\mathbf{X}_i \times \boldsymbol{\beta}, \sigma^2) \tag{5.18}$$

The problem with the ML estimator is that we have to estimate the intercept and the slope, which are in $\boldsymbol{\beta}$ in Equation (5.18). Obviously, the problem is solved if there is no $\boldsymbol{\beta}$. All that REML does is apply a little trick to avoid having any $\boldsymbol{\beta}$ in Equation (5.18). It does this by finding a special matrix $\mathbf{A}$ of dimension $n \times (n-2)$, and special means 'orthogonal to (or independent of) $\mathbf{X}'$, multiplies $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ with this matrix and continues with ML estimation. Orthogonal means that if $\mathbf{A}$ and $\mathbf{X}$ are multiplied, the result is 0. Hence, we get $\mathbf{A}' \times \mathbf{Y} = \mathbf{A}' \times \mathbf{X} \times \boldsymbol{\beta} + \mathbf{A}' \times \boldsymbol{\varepsilon}$ $= \mathbf{0} + \mathbf{A}' \times \boldsymbol{\varepsilon} = \mathbf{A}' \times \boldsymbol{\varepsilon}$. The distribution for $\mathbf{A}' \times \mathbf{Y}$ is now given by

$$\mathbf{A}' \times \mathbf{Y} \sim N(\mathbf{0}, \sigma^2 \times \mathbf{A}' \times \mathbf{A}) \tag{5.19}$$

which no longer depends on $\boldsymbol{\beta}$. Applying ML on $\mathbf{A}' \times \mathbf{Y}$ gives an unbiased estimator for σ^2 (same expression as in Equation (5.14)). Now we discuss how REML can be used for the mixed effects model. Our starting point is the marginal model

$$\mathbf{Y}_i \sim N(\mathbf{X}_i \times \boldsymbol{\beta}, \mathbf{V}_i) \qquad \mathbf{V}_i = \mathbf{Z}_i \times \mathbf{D} \times \mathbf{Z}_i' + \Sigma_i \tag{5.20}$$

The story now starts all over again. As before, we can formulate a slightly different log-likelihood criteria. The unknown parameters are $\boldsymbol{\beta}$ and the elements of $\mathbf{D}$ and $\boldsymbol{\Sigma}_i$. Again, we denote them all by $\boldsymbol{\theta}$. The log-likelihood function is given by

$$\ln L(\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\theta}) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{n} \ln |\mathbf{V}_i| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})' \times \mathbf{V}_i^{-1} \times (\mathbf{Y}_i - \mathbf{X}_i \times \boldsymbol{\beta})$$

The notation $|\mathbf{V}_i|$ stands for the determinant of $\mathbf{V}_i$. This looks intimidating, but can be found in many introductory statistical textbooks. Just as before, an expression for $\boldsymbol{\beta}$ is obtained by setting the partial derivative of L with respect to $\boldsymbol{\beta}$ equal to zero and solving the equation. Just as in the example discussed on the previous page,

doing the same for the elements of the covariance matrix $\mathbf{V}_i$ gives biased estimates, and therefore we need REML.

For the RIKZ data, we had 9 beaches; hence, $i = 1, \ldots, 9$. In general, the index i runs from 1 to n. We can stack all the vectors $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i5})$ into one long vector of dimension 45×1 (nine beaches, five observations per beach). Let us denote the stacked column by $\mathbf{Y}$. We can also stack all the $\mathbf{X}_i$ into one matrix of dimension $45 \times p$, where p is the number of fixed covariates. Denote it by $\mathbf{X}$. We have to do something slightly different for the covariance matrix. Instead of stacking them, we create a new matrix $\mathbf{V}$ with diagonal blocks $\mathbf{V}_1$ to $\mathbf{V}_9$. The other elements of $\mathbf{V}$ are equal to 0. A similar approach is followed for the $\mathbf{Z}_i$s. Using this new notation, we can write Equation (5.20) as $\mathbf{Y} \sim \text{N}(\mathbf{X} \times \boldsymbol{\beta}, \mathbf{V})$. Just as before, the $\mathbf{Y}$ vector is multiplied with a special matrix $\mathbf{A}$, such that $\mathbf{A}' \times \mathbf{Y} = \mathbf{A}' \times \mathbf{X} \times \boldsymbol{\beta} + \mathbf{A}' \times \mathbf{V} = \mathbf{0} + \mathbf{A}' \times \mathbf{V}$. We can write $\mathbf{A}' \times \mathbf{Y} \sim \text{N}(\mathbf{0}, \mathbf{A}' \times \mathbf{V} \times \mathbf{A})$, and maximum likelihood is used to obtain unbiased estimates for the elements of $\mathbf{V}$. The good news is that the estimators for the variance terms are independent of (not related to) the choice for $\mathbf{A}$. Summarising, REML applies a special matrix multiplication on $\mathbf{Y}$ in such a way that the $\mathbf{X} \times \boldsymbol{\beta}$-bit disappears. It then continues with maximum likelihood estimation and the resulting parameter estimators are unbiased and not related to the specific matrix multiplication. As a consequence, the REML estimators for the $\boldsymbol{\beta}$s are not identical to the maximum likelihood estimators. If the number of fixed covariates is small relative to the number of observations, there are not many differences, but for models with many fixed terms, this may not be the case.

5.6.1 Illustration of Difference Between ML and REML

To illustrate this, we applied two models on the RIKZ data. Both models are random intercept models estimated with ML and REML. In the first model, we used only NAP as fixed covariate and in the second model NAP and exposure. All numerical outputs are given in Table 5.1. For the model that only contains NAP as the fixed term, differences in estimated parameters and variances between REML and ML are relatively small. Adding the nominal variable exposure increases the number of regression parameters by 1. The ML estimated slope for NAP is –2.60 with the REML now –2.58. The R code for the two models is as follows. The `method = "ML"` or `method = "REML"` specifies which estimation method is used. The first three lines define the nominal variable exposure with two levels (instead of 3). The output was obtained with the `summary` command.

```
> RIKZ$fExp <- RIKZ$Exposure
> RIKZ$fExp[RIKZ$fExp == 8] <- 10
> RIKZ$fExp <- factor(RIKZ$fExp, levels = c(10, 11))
> M0.ML <- lme(Richness ~ NAP, data = RIKZ,
               random = ~1 | fBeach, method = "ML")
```

```
> M0.REML <-lme(Richness ~ NAP, random = ~1 | fBeach,
                method = "REML", data = RIKZ)
> M1.ML <- lme(Richness ~ NAP + fExp, data = RIKZ,
               random = ~1 | fBeach, method = "ML")
> M1.REML <- lme(Richness ~NAP + fExp, data = RIKZ,
                 random = ~1 | fBeach, method = "REML")
```

Table 5.1 Results for two models using ML (middle column) and REML (right column) estimation. Numbers between brackets are standard errors. The first model (upper part of the table) uses an intercept and NAP as fixed covariates and a random intercept. The second model (lower part of the table) used the same terms, except that the nominal variable exposure is used as a fixed term as well

Mixed model with NAP as fixed covariate and random intercept		
Parameter	Estimate using ML	Estimate using REML
Fixed intercept	6.58 (1.05)	6.58 (1.09)
Fixed slope NAP	–2.57 (0.49)	–2.56 (0.49)
Variance random intercept	7.50	8.66
Residual variance	9.11	9.36
AIC	249.82	247.48
BIC	257.05	254.52
Mixed model with NAP and exposure as fixed covariate and random intercept		
Fixed intercept	8.60 (0.96)	8.60 (1.05)
Fixed slope NAP	–2.60 (0.49)	–2.58 (0.48)
Fixed Exposure level	–4.53 (1.43)	–4.53 (1.57)
Variance random intercept	2.41	3.63
Residual variance	9.11	9.35
AIC	244.75	240.55
BIC	253.79	249.24

5.7 Model Selection in (Additive) Mixed Effects Modelling

In the earlier sections, we applied a series of models on the species richness for the RIKZ data. Although the original data set contained 10–15 explanatory variables, we have only used NAP and exposure as explanatory variables because our prime aim here is to explain methodology and not to provide the best model for these data. The case studies can be consulted for examples of best possible models. We now use the RIKZ data to explain model selection in mixed effects modelling.

Just as in linear regression, there are two main options for model selection. One option is based on selection tools like the Akaike Information Criteria (AIC), or the Bayesian Information Criteria (BIC). Both the AIC and BIC contain two terms that measure the fit of the model and the complexity of the model. The likelihood value is used in defining the measure of fit, and the number of parameters measures the complexity.

As a measure of fit, we can use the log likelihood function. But there are two likelihood functions: the REML and the ML one. It can be shown (Verbeke and Molenberghs, 2000) that

$$L_{REML}(\boldsymbol{\theta}) = \left| \sum_{i=1}^{n} \mathbf{X}_i' \times V_i^{-1} \times \mathbf{X}_i \right|^{-0.5} \times L_{ML}(\boldsymbol{\theta})$$

The AIC is defined as twice the difference between the value of the likelihood L (measure of fit) and the number of parameters (penalty for model complexity) in $\boldsymbol{\theta}$. For the BIC, the number of observations is also taken into account, which means that more severe increases in the likelihood are required for larger data sets to label a model as better. In the formulae below, p is the number of parameters in $\boldsymbol{\theta}$, L is either the ML or REML likelihood, and for ML, we have $n^* = n$, but for REML, $n^* = n - p$.

$$AIC = -2 \times L(\boldsymbol{\theta}) + 2 \times p$$
$$BIC = -2 \times L(\boldsymbol{\theta}) + 2 \times p \times \ln(n^*)$$

This means that an AIC based on REML is not comparable with an AIC obtained by ML. The same holds for the BIC.

The second approach to find the optimal model is via hypothesis testing. There are three options here: (i) the t-statistic, the F-statistic, or the likelihood ratio test. In Chapter 1, we discussed how to compare nested linear models using the maximum likelihood ratio test. The problem is that the mixed effects model contains two components: a fixed effect (the explanatory variables) and the random effects. So, we need to select not only an optimal fixed effects structure but also an optimal random effects structure. In most cases, we are interested in the fixed effects. But if the random effects are poorly chosen, then this affects the values (biased) and quality of the fixed effects as the random effects work their way into the standard errors of the slopes for the fixed effects. On the other hand, variation in the response variable not modelled in terms of fixed effects ends up in the random effects. There are two strategies to work your way through the model selection process: the top-down strategy and the step-up strategy (West et al., 2006). The first one is recommended by Diggle et al. (2002) and is the only one discussed here. The protocol for the top-down strategy contains the following steps:

1. Start with a model where the fixed component contains all explanatory variables and as many interactions as possible. This is called the *beyond optimal* model. If this is impractical, e.g. due to a large number of explanatory variables, interactions, or numerical problems, use a selection of explanatory variables that you think are most likely to contribute to the optimal model.
2. Using the beyond optimal model, find the optimal structure of the random component. Because we have as many explanatory variables as possible in the fixed component, the random component (hopefully) does not contain any information that we would like to have in the fixed component. The problem is that comparing

two models with nested random structures cannot be done with ML because the estimators for the variance terms are biased. Therefore, we must use REML estimators to compare these (nested) models. Obtaining valid p-values for such tests is another non-trivial issue due to something called *testing on the boundary*, which we will discuss later in this section. As well as using the (REML) likelihood ratio test, we can also use the AIC or BIC, but again we need to use REML. Using AIC or BIC does not avoid boundary problems.

3. Once the optimal random structure has been found, it is time to find the optimal fixed structure. As mentioned above, we can either use the F-statistic or the t-statistic obtained with REML estimation or compare nested models. To compare models with nested fixed effects (but with the same random structure), ML estimation must be used and not REML. We discuss the details of these tests later in this chapter.
4. Present the final model using REML estimation.

These steps should only be used as a general guidance, and sometimes common sense is required to derive a slightly different approach. For example, sometimes, it is impractical to apply a model with as many explanatory variables as possible, especially in generalised additive modelling.

5.8 RIKZ Data: Good Versus Bad Model Selection

5.8.1 The Wrong Approach

We start with an illustration how not to do a mixed effects model selection. In particular, we show the danger of not starting with a full model. To illustrate this, we take NAP as the only fixed explanatory variable for the fixed component and ignore exposure for the moment. As to the random structure, there are three options: (i) no random term, except for the ordinary residuals; (ii) a random intercept model using beach; and (iii) a random intercept and slope model.

A requirement for the `nlme` function in R is the specification of a random term, and to avoid an error message, the `gls` function can be used instead. The R code for these three models is:

```
> Wrong1 <- gls(Richness ~ 1 + NAP, method = "REML",
                data = RIKZ)
> Wrong2 <- lme(Richness ~ 1 + NAP, random = ~1|fBeach,
                method = "REML", data = RIKZ)
> Wrong3 <- lme(Richness ~ 1 + NAP, method = "REML",
                random = ~1 + NAP | fBeach, data = RIKZ)
```

All models have the same fixed effect structure, but different random components.

5.8.1.1 Step 2 of the Protocol

The second step of the protocol dictates that we judge which of these models is optimal. Note that the only difference is the random structure. Because REML estimation was used, we can compare AICs or BICs. These are obtained with the `AIC` or `BIC` commands:

```
> AIC(Wrong1, Wrong2, Wrong3)

       df      AIC
Wrong1  3 258.2010
Wrong2  4 247.4802
Wrong3  6 244.3839
```

This suggests the model with the random intercept and slope is the best. Note that both the second and third models are considerably better than the model without a random effect. The BIC for the second and third models are similar and both are lower than the BIC of the first model. Instead of the AIC (or BIC), we can also use the likelihood ratio test via the `anova` command as the models are nested.

```
> anova(Wrong1, Wrong2, Wrong3)

     Model df   AIC   BIC    logLik Test L.Ratio p-val.
Wrong1   1  3 258.20 263.48 -126.10
Wrong2   2  4 247.48 254.52 -119.74 1 vs 2 12.72 <0.001
Wrong3   3  6 244.38 254.95 -116.19 2 vs 3  7.09  0.03
```

The second line compares a model without any random effect versus a model with a random intercept. These models are nested with respect to the variances. Unfortunately, there is a little problem here, which is the 'testing on the boundary' mentioned earlier. The null hypothesis of this test is H_0: $\sigma^2 = 0$ versus the alternative H_1: $\sigma^2 > 0$. This is different from how you normally use this test to see whether a regression parameter is equal to zero or not. In that case, we use H_0: $\beta = 0$ versus the alternative H_1: $\beta \neq 0$. Note the subtle difference with respect to the $>$ and $\neq$ symbols. This is called testing on the boundary for the obvious reason that if there is no evidence to reject the null-hypothesis, then $\sigma^2 = 0$ is the lowest possible value as a variance is always non-negative. The p-value provided by the `anova` function is incorrect as this function assumes that twice the differences between the two log-likelihood values, $L = -2 \times (-126.10 + 119.74) = 12.72$, follows a Chi-square distribution with p degrees of freedom; p is the number of extra parameters in the full model (here $p = 1$). The mathematical notation for such a distribution is ${\chi_p}^2$. However, when testing on the boundary, L does not follow this distribution, and therefore the p-value from the table is incorrect. Verbeke and Molenberghs (2000) showed that L follows a $0.5 \times ({\chi_0}^2 + {\chi_1}^2) = 0.5 \times {\chi_1}^2$ distribution. This means that the p-value in the table should be divided by 2. In R, you can get the correct p-value by typing

```
> 0.5 * (1 - pchisq(12.720753, 1))
```

The resulting *p*-value is still smaller than 0.001. This means that adding a random effect beach to the model is a significant improvement. Note that this correction only applies for comparing a model without and with a random intercept! If we want to compare the model with the random intercept and the model with random intercept and slope, then $L = 7.09$ follows a $0.5 \times (\chi_1^2 + \chi_2^2)$ distribution. The resulting *p*-value of 0.018 is calculated by

```
> 0.5 * ((1 - pchisq(7.09, 1)) + (1 - pchisq(7.09, 2)))
```

So, the random structure that contains both the random intercept and slope is significantly better (at least at the 5% level) than the random intercept model. The conclusion of step 2 is that you should proceed to step 3 with the random intercept and slope model.

5.8.1.2 Step 3 of the Protocol

In step 3, we search for the optimal fixed structure for a given random structure. Typing `summary(Wrong3)` gives a slope of –2.83 for NAP, and the associated standard error and *t*-value are 0.72 and –3.91, respectively. The *p*-value of the *t*-statistic is smaller than 0.001, indicating that the slope for NAP is significant. Hence, dropping NAP from the model is not an option. The only thing we can try is adding exposure or adding exposure *and* the interaction between exposure and NAP. We can test the significance of these tests in three ways: either with an *F*-test or *t*-test obtained with REML or by comparing nested models using ML estimation. The first approach is carried out in R as follows. In case you skipped the previous section, the first three lines redefine the nominal variable exposure such that it only has two levels instead of three.

```
> RIKZ$fExp <- RIKZ$Exposure
> RIKZ$fExp[RIKZ$fExp == 8] <- 10
> RIKZ$fExp <- factor(RIKZ$fExp, levels = c(10, 11))
> lmc <- lmeControl(niterEM = 2200, msMaxIter = 2200)
> Wrong4 <- lme(Richness ~1 + NAP * fExp,
              random = ~1 + NAP | fBeach,
              method = "REML", data = RIKZ)
> anova(Wrong4)
```

```
            numDF denDF  F-value p-value
(Intercept)     1    34 34.87139  <.0001
NAP             1    34 18.65502  0.0001
fExp            1     7  5.65495  0.0490
NAP:fExp        1    34  3.32296  0.0771
```

The `anova` command applies sequential testing; the interaction term is the last term to be added, but the order of NAP and exposure depends on how we specified the model. Change the order of NAP and exposure and we may get different *p*-values

for these two terms. The useful bit of this table is the last line, where it is testing whether the NAP × exposure interaction term is significant. The *F*-statistic is 3.32, and the *p*-value suggests it is not significant at the 5% level.

We can use the *t*-statistic as an alternative to the *F*-statistic. They are calculated in the same way as in linear regression, namely, the estimated value divided by its standard error. They are obtained with the `summary(Wrong4)` command, and the relevant output is given below.

```
Fixed effects: Richness ~ 1 + NAP * fExp
                Value Std.Error DF   t-value p-value
(Intercept)  9.118945 1.2242357 34  7.448684  0.0000
NAP         -3.879203 0.8816476 34 -4.399947  0.0001
fExp11      -5.534743 1.8510032  7 -2.990132  0.0202
NAP:fExp11   2.429496 1.3327641 34  1.822900  0.0771
```

The *t*-statistic also shows that we can drop the interaction term. Rerunning the model without the interaction term gives a *t*-statistic of $t = -2.44$ ($p = 0.04$), which is not convincing neither. The new output is given below.

```
Fixed effects: Richness ~ 1 + NAP + fExp
                Value Std.Error DF   t-value p-value
(Intercept)  8.407714  1.183419 35  7.104595  0.0000
NAP         -2.808422  0.759642 35 -3.697034  0.0007
fExp11      -3.704917  1.517669  7 -2.441189  0.0447
```

Both the *F*-statistic and the *t*-statistic indicate a strong NAP effect, but a weak exposure effect and no significant interaction. Both these test are approximate. This means that we should not take them too literally. Hence, $p = 0.04$ is not convincing evidence of an exposure effect.

Before moving on to the ML testing procedure, we first need to address the issue of degrees of freedom. Within the mixed effects modelling literature, explanatory variables are divided into level 1 and level 2 variables. An explanatory variable that has the same value for all observations within the levels of the random effect is called a level 2 variable. An example is exposure; it has the same value for all observations on a beach. NAP, on the other hand, has a different value for each observation within a beach; it is called a level 1 variable. The degrees of freedom for a level 1 variable (NAP) in R is calculated as the number of level 1 observations (= 45) minus the number of level 2 clusters (= 9 levels in the random variable beach) minus the number of level 1 fixed effects (1, namely NAP). This explains the 35 degrees of freedom for NAP. For a level 2 variable, the equation is slightly different. It is calculated as the number of level 2 clusters (= 9 levels in the random variable beach) minus the number of level 2 fixed variables (In case only exposure) minus 1 if there is an intercept. This explains why the degrees of freedom are equal to 7 for exposure. Further details can be found on page 111 in West et al. (2006), Verbeke and Molenberghs (2000), or Pinheiro and Bates (2000).

So far, we only discussed the use of the approximate F-statistic and t-statistic. Estimation was done with REML. We now show the third hypothesis testing approach: the likelihood ratio test using ML estimation. In this approach, we fit two models with the same random effects structure using ML estimation and compare the likelihood criteria. The code below compares the model with a fixed structure containing NAP versus NAP + exposure. It also compares the model with NAP + exposure versus the model that also contains the interaction between both explanatory variables.

```
> lmc <- lmeControl(niterEM = 5200, msMaxIter = 5200)
> Wrong4A <- lme(Richness ~1 + NAP, method="ML",
                 control = lmc, data = RIKZ,
                 random = ~1 + NAP | fBeach)
> Wrong4B <- lme(Richness ~ 1 + NAP + fExposure,
                 random = ~1 + NAP | fBeach, method="ML",
                 data = RIKZ)
> Wrong4C <- lme(Richness ~1 + NAP * fExposure,
                 random = ~1 + NAP | fBeach, data = RIKZ,
                 method = "ML", control = lmc)
> anova(Wrong4A, Wrong4B, Wrong4C)
```

To avoid an error message related to convergence, we used the `control` option, which basically tells R to use more iterations. The output from the `anova` command is given below.

```
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Wrong4A     1  6 246.6578 257.4977 -117.3289
Wrong4B     2  7 245.3353 257.9820 -115.6677 1 vs 2 3.322437  0.0683
Wrong4C     3  8 243.2228 257.6761 -113.6114 2 vs 3 4.112574  0.0426
```

The comparison of model 1 versus 2 (`Wrong4A` versus `Wrong4B`) shows that exposure is not significant at the 5% level. Adding the interaction to a model that already contains exposure gives a log ratio statistic of $L = 4.11$, which is borderline significant ($p = 0.04$). So the ML testing procedure also indicates the interaction and exposure effects may be dropped from the model. This means that the optimal model, according to our model selection strategy, contains NAP as a fixed effect with a random slope and intercept. This means that the NAP effect is changing per beach (but in a random fashion).

5.8.1.3 Step 4 of the Protocol

As a last step, we need to present the numerical output of the optimal model using REML estimation. The code for the optimal model is

```
> Wrong5 <- lme(Richness~1+NAP,
                random = ~1 + NAP | fBeach,
                method = "REML", data = RIKZ)
> summary(Wrong5)

Random effects:
 Formula: ~1 + NAP | fBeach
            StdDev   Corr
(Intercept) 3.55    (Intr)
NAP         1.71     -0.988
Residual    2.69
Fixed effects: Richness ~ 1 + NAP
               Value Std.Error  DF   t-value p-value
(Intercept)     6.59      1.26  35      5.20  <0.001
NAP            -2.83      0.72  35     -3.90  <0.001
```

5.8.2 The Good Approach

5.8.2.1 Step 1 of the Protocol

The top-down strategy specified earlier in this chapter indicated that we should start with as many explanatory variables as possible in the fixed component. So, we should start with a model that contains as fixed effects NAP, exposure, and their interaction. The starting point, therefore, is

```
> B1 <- gls(Richness ~ 1 + NAP * fExp,
            method = "REML", data = RIKZ)
> B2 <- lme(Richness ~1 + NAP * fExp, data = RIKZ,
            random = ~1 | fBeach, method = "REML")
> B3 <- lme(Richness ~ 1 + NAP * fExp, data = RIKZ,
            random = ~1 + NAP | fBeach, method="REML")
```

5.8.2.2 Step 2 of the Protocol

The AIC values of these three models are 238.53, 236.49, and 237.13. The random intercept model is therefore the preferred option.

5.8.2.3 Step 3 of the Protocol

The `summary(B2)` command indicates that all parameters in this model are significant as can be seen from the table below.

```
Fixed effects: Richness ~ 1 + NAP * fExp
                Value Std.Error DF   t-value p-value
(Intercept)  8.861084 1.0208449 34  8.680147  0.0000
NAP         -3.463651 0.6278583 34 -5.516613  0.0000
fExp11      -5.255617 1.5452292  7 -3.401190  0.0114
NAP:fExp11   2.000464 0.9461260 34  2.114374  0.0419
```

If we use the same argument as above that a p-value of 0.04 is unconvincing, we could drop the interaction and refit the model. In that case, the p-value for exposure is 0.01, which is probably small enough to keep it in.

5.8.2.4 Step 4 of the Protocol

The results of the optimal model are given below.

```
Linear mixed-effects model fit by REML
Data: RIKZ
       AIC      BIC    logLik
  240.5538 249.2422 -115.2769

Random effects:
Formula: ~1 | fBeach
        (Intercept) Residual
StdDev:    1.907175 3.059089

Fixed effects: Richness ~ 1 + NAP + fExp
                Value Std.Error DF   t-value p-value
(Intercept)  8.601088 1.0594876 35  8.118158  0.0000
NAP         -2.581708 0.4883901 35 -5.286160  0.0000
fExp11      -4.532777 1.5755610  7 -2.876929  0.0238
```

Note that we end up with a fundamentally different model compared to our first approach above. The biological conclusion is also very different as this model suggests there is a strong NAP effect, a weak exposure effect, and absolute values differ per beach in a random way (as modelled by the random intercept).

The reason we ended up with a different model is because in the previous example, part of the information that we want to have in the fixed effects ended up in the random effects. This is due to starting with a fixed component that only contained NAP.

5.9 Model Validation

As with linear regression and additive modelling, the prime tool to validate the model is the normalised residuals based on the REML fit in step 4 of the protocol. These were defined in Chapter 4. Residuals should be plotted against fitted val-

ues to identify violation of homogeneity, indicated by differences in spread. If you do see an increase in spread for larger fitted values, then there are several options: (i) apply a transformation, (ii) check whether the increase in spread is due to a covariate, and (iii) apply generalised linear mixed modelling with a Poisson distribution (if the data are counts). If the increase in spread is due to a covariate, use the methods described in Chapter 4. These can easily be combined with a random effect.

You should also plot the residuals against each explanatory variable. Again, you do not want to see any patterns in the spread. Nor do you want to see a pattern in the residuals as it indicates the wrong model was applied. If this happens, consider adding more explanatory variables, interactions, quadratic terms, and if this does not help, use additive mixed modelling.

To verify normality, make histograms of the residuals. We recommend assessing normality (and homogeneity) using graphical tools. However, some software packages provide normality tests like the Shapiro-Wilks test, and these offer an alternative approach.

Examples of the model validation are given in the case studies and in the next section.

5.10 Begging Behaviour of Nestling Barn Owls

For those readers who enjoy television shows with many people in a house and cameras all over the place, here is the ecological version of it. Roulin and Bersier (2007) analysed the begging behaviour of nestling barn owls.

They looked at how nestlings responded to the presence of the father and of the mother. Using microphones inside and a video outside the nests, they sampled 27 nests and studied vocal begging behaviour when the parents brought prey. The number of nestlings was between 2 and 7 per nest.

Different response variables were defined in the paper: the amount of time spent on the perch by a parent, the amount of time in the nestbox, sibling negotiation, and begging. Here, we analyse sibling negotiation[2], which is defined as follows. Using the recorded footage, the number of calls made by all offspring in the absence of the parents was counted during 30-s time intervals every 15 min. To allocate a number of calls to a visit from a parent, the counted number of calls from the preceding 15 min of the arrival was used. This number was then divided by the number of nestlings. You may need to read this last sentence more than once, but in summary, the sibling

[2]When the need for food varies between the young owls, the calls used in the absence of parent birds have been shown to communicate the different levels of hunger between the chicks. This pre-parental arrival behaviour then seems to influence competitive behaviour between chicks when the parent bird arrives. Using information from this sibling communication, the least hungry chick avoids competing for food against the hungriest chick, which is the more likely to succeed in winning the food from the parent bird. Thus saving energy to only compete for food when there is the highest probability of successfully winning it.

negotiation is just the number of counted calls in the nearest 30-s interval before the arrival of a parent divided by the number of nestlings. In Chapters 12 and 13, we return to these data and analyse the number of calls using a Poisson distribution. We also use the data in Chapter 6 to model a more detailed auto-correlation structure.

The explanatory variables are sex of the parent, treatment of food, and arrival time of the parent. Half of the nests were given extra prey, and from the other half, prey (remaining) were removed. These were called 'food-satiated' and 'food-deprived', respectively. Measurements took place on two nights, and the food treatment was swapped on the second night. Note that the original paper contains an ethical note stating that food treatment did not have an effect on survival of the chicks. Measurements took place between 21.30 h and 05.30 h and the variable ArrivalTime reflects the time when a parent arrived at the perch with a prey. Further biological information and a description of the fascinating behaviour of barn owl nestlings can be found in the Roulin and Bersier (2007).

How should we analyse these data? Ok, given the fact that this is a section in a mixed effects modelling chapter, it should not be difficult to guess that nest will be used as a random effect. The reasons for this are as follows. Firstly, there were multiple observations from the same nests so these observations will be correlated. Secondly, there are 27 nests and using nest as a fixed effect would be rather expensive in terms of degrees of freedom. Furthermore, we would like to make a statement on relationships for barn owl nests in general and not just on these 27. If we use nest as a random effect, we allow for correlation between multiple observations from the same nest, and we only need to estimate one variance, and our statements will hold for all similar nests. Instead of starting immediately with a model that contains nest as a random effect, we will follow one of the protocols described earlier. We can either use the four-step protocol presented in Section 5.7 or the ten-step protocol discussed in Chapter 4. The later one has more intermediate steps, but basically does the same thing. Because the protocol from Chapter 4 is easier to follow (more detail, less chance to make mistakes), we use it here.

5.10.1 Step 1 of the Protocol: Linear Regression

We start with a linear regression model. Nestling negotiation is modelled as a function of sex of the parents, arrival time, and food treatment. Because one of the prime aims of the analysis is to find a sex effect, we also include the interaction between sex and each of the other variables. See also Appendix A for a discussion on interactions. The following R code imports the data, applies the linear regression model, and produces the graph in Fig. 5.4.

```
> library(AED) ; data(Owls)
> M.lm <- lm(NegPerChick ~ SexParent * FoodTreatment +
             SexParent * ArrivalTime, data = Owls)
> plot(M.lm, select = c(1))
```

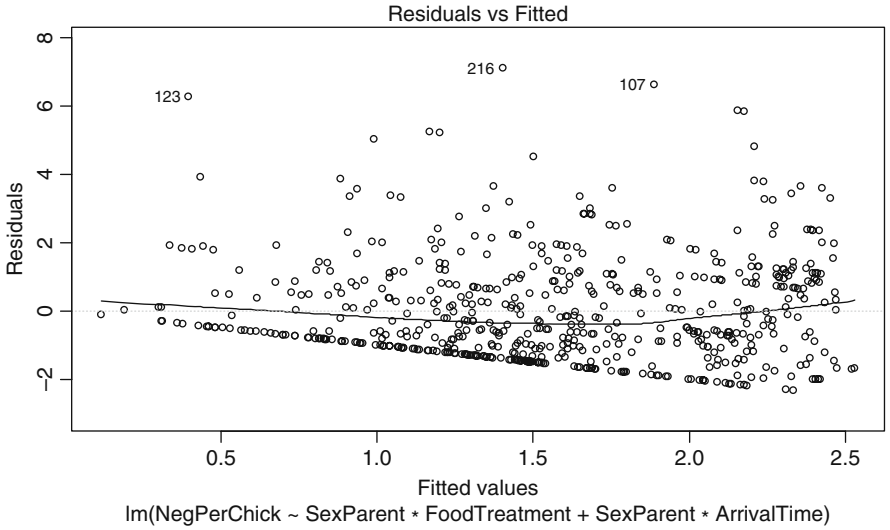

Fig. 5.4 Residuals versus fitted values for the linear regression model. Note that the residual spread increases for larger fitted values, indicating heterogeneity

The graph indicates heterogeneity because the residual spread increases along the horizontal axis. To understand why we have heterogeneity, we plotted residuals versus sex of the parents, food treatment, and arrival time.

However, as there is no clear pattern in any of these graphs, we cannot easily model the heterogeneity the way we did in Chapter 4. For this reason, we went for plan B and applied a $\log_{10}(Y + 1)$ transformation on the sibling negotiation data. This transformation was also used in Roulin and Bersier (2007). The code below applies the $\log_{10}$ transformation, refits the model, and plots the residuals versus the nominal variable nest (Fig. 5.5).

```
> Owls$LogNeg <- log10(Owls$NegPerChick + 1)
> M2.lm <- lm(LogNeg ~ SexParent * FoodTreatment +
              SexParent * ArrivalTime, data = Owls)
> E <- rstandard(M2.lm)
> boxplot(E ~ Nest, data = Owls, axes = FALSE,
          ylim = c(-3, 3))
> abline(0,0); axis(2)
> text(1:27, -2.5, levels(Owls$Nest), cex=0.75, srt=65)
```

The `abline(0, 0)` command adds a horizontal line at $y = 0$. The `axes = FALSE` and `text` commands are used to add fancy labels along the horizontal axis. In a perfect world, the residuals should lie in a cloud around this line without any patterns. However, for some nests, all residuals are above or below the zero line, indicating that the term 'nest' has to be included in the model. We can do this as

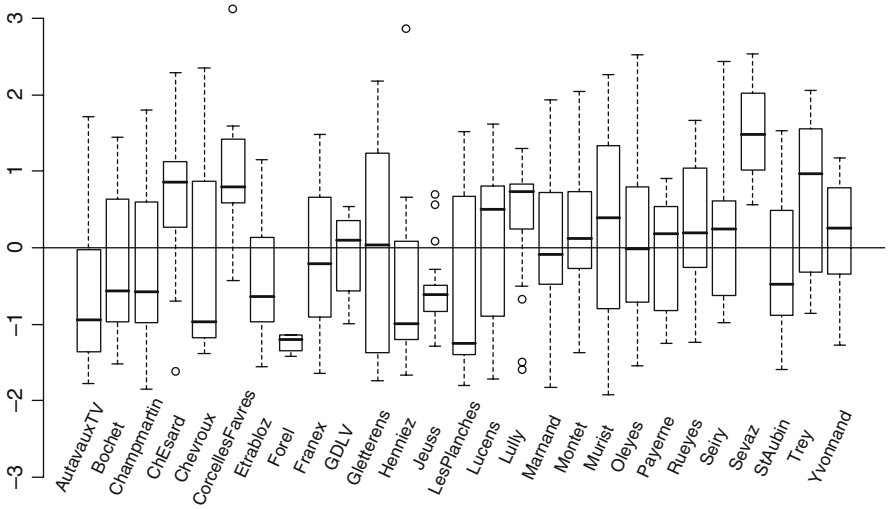

Fig. 5.5 Boxplot of standardised residuals obtained by a linear regression model applied on the log-transformed sibling negotiation data. The *y*-axis shows the values of the residuals and the horizontal axis the nests. Note that some nests have residuals that are above or below the zero line, indicating the need for a random effect

a fixed term or as a random term, but we already discussed that this has to be as a random term.

5.10.2 Step 2 of the Protocol: Fit the Model with GLS

In this step we fit the model using the `gls` function. It allows us to compare the linear regression model with the mixed effects model that we will calculate using the `lme` function in a moment.

```
> library(nlme)
> Form <- formula(LogNeg ~ SexParent * FoodTreatment +
                  SexParent * ArrivalTime)
> M.gls <- gls(Form, data = Owls)
```

To reduce the code, we have used the formula expression. The numerical output in the object `M.gls` is identical to that of the `lm` function.

5.10.3 Step 3 of the Protocol: Choose a Variance Structure

In Chapter 4, this step consisted of finding the optimal variance structure in terms of heterogeneity. We can still do that here, but adding the random component nest is our first priority. Note that the random intercept is also part of the ‘choose a

variance structure' process. This means that the following random intercept mixed effects model is fitted.

$$\begin{aligned} LogNeg_{ij} = \alpha &+ \beta_1 \times SexParent_{ij} + \beta_2 \times Foodtreatment_{ij} \\ &+ \beta_3 \times ArrivalTime_{ij} + \beta_4 \times SexParent_{ij} \times FoodTreatment_{ij} \\ &+ \beta_5 \times SexParent_{ij} \times ArrivalTime_{ij} + a_i + \varepsilon_{ij} \end{aligned}$$

LogNeg_{ij} is the log-10 transformed sibling negotiation for observation j at nest i. SexParent_{ij} and $\text{FoodTreatment}_{ij}$ are nominal variables with two levels, and ArrivalTime_{ij} is a continuous variable. The second line contains interactions. The term a_i is a random intercept and is assumed to be normally distributed with mean 0 and variance d^2. The residual ε_{ij} is assumed to be normally distributed with mean 0 and variance σ^2. Both random terms are assumed to be independent of each other.

5.10.4 Step 4: Fit the Model

The linear mixed effects model is applied in R with the following code.

```
> M1.lme <- lme(Form, random = ~ 1 | Nest,
                method = "REML", data = Owls)
```

5.10.5 Step 5 of the Protocol: Compare New Model with Old Model

We use the `anova` command to compare the models `M.gls` and `M1.lme`. Note that the models were estimated with REML, which allows us to apply the likelihood ratio test to see whether we need the random intercept.

```
> anova(M.gls, M1.lme)
      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
M.gls     1  7 64.37422 95.07058 -25.18711
M1.lme    2  8 37.71547 72.79702 -10.85773 1 vs 2 28.65875  <.0001
```

The likelihood ratio test indicates that the model with the random intercept is considerably better. You would quote this statistic in a paper as $L = 28.65$ ($df = 1$, $p < 0.001$). Recall from Section 5.8 that we are testing on the boundary here. If we did the correction for testing on the boundary, the p-value would get even smaller. Because the random intercept is highly significant, testing on the boundary is not a problem here.

The AIC of the model with the random intercept is also considerably smaller, confirming the results of the likelihood ratio test. As well as the random intercept, it is also an option to use a random intercept and random slope model. In this case, you assume that the strength of the relationship between sibling negotiation and arrival time changes randomly between the nests. We leave this as an exercise to the reader.

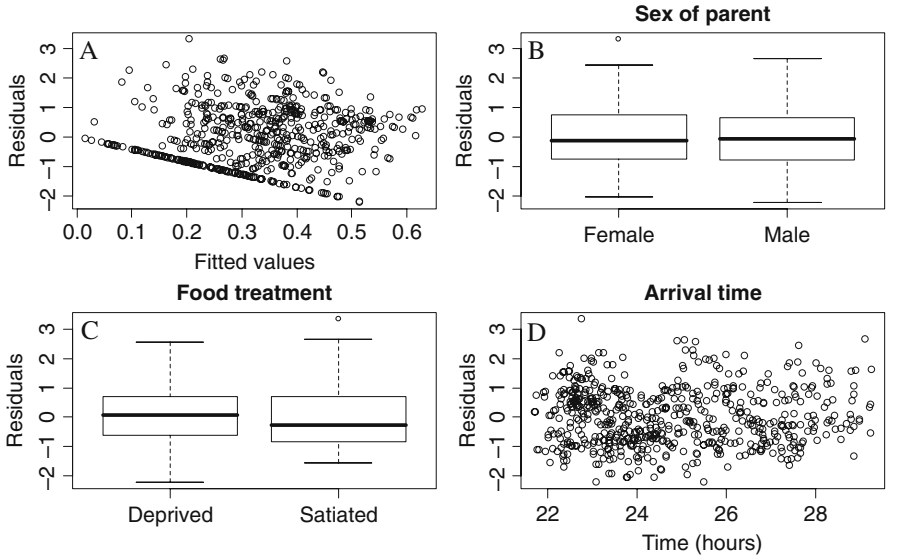

Fig. 5.6 Model validation graphs for the random intercept mixed effects model. Residuals are plotted versus fitted values (**A**), sex of the parent (**B**), food treatment (**C**), and arrival time (**D**)

5.10.6 Step 6 of the Protocol: Everything Ok?

The next thing we should think of is whether we have homogeneity of variance in the model and independence. Before doing anything, ask yourself whether you expect different residual spread per sex or per treatment (or over time). We have a large data set and blindly following some test statistics may not be wise. The large number of observations means that even small differences in spread may cause a significant variance covariate and we prefer to judge homogeneity by eye. Figure 5.6 shows residuals versus fitted values, sex, food treatment, and arrival time. These graphs do not show any clear violation of heterogeneity. There may be a violation of independence along arrival time, but Fig. 5.6D is not very clear. For the moment, we ignore any potential independence problems, and return to this issue later in this section. The R code to make Fig. 5.6 is as follows. Residuals and fitted values are extracted, a graph with four panels is set up, and the rest is a matter of trivial `boxplot` and `plot` commands.

```
> E2 <- resid(M1.lme, type = "normalized")
> F2 <- fitted(M1.lme)
> op <- par(mfrow = c(2, 2), mar = c(4, 4, 3, 2))
> MyYlab <- "Residuals"
> plot(x = F2, y = E2, xlab = "Fitted values", ylab = MyYlab)
> boxplot(E2 ~ SexParent, data = Owls,
          main = "Sex of parent", ylab = MyYlab)
```

```
> boxplot(E2 ~ FoodTreatment, data = Owls,
          main = "Food treatment", ylab = MyYlab)
> plot(x = Owls$ArrivalTime, y = E, ylab = MyYlab,
       main = "Arrival time", xlab = "Time (hours)")
> par(op)
```

5.10.7 Steps 7 and 8 of the Protocol: The Optimal Fixed Structure

In this step, we look at the optimal model in terms of the explanatory variables sex, food treatment, arrival time, and the selected interaction terms. The first thing we should do is to type `summary(M1.lme)` and inspect the significance of the regression parameters.

```
                              Value  Std.Error  DF   t-value p-value
(Intercept)                1.1236414 0.19522087 567  5.755744  0.0000
SexParentMale              0.1082138 0.25456854 567  0.425087  0.6709
FoodTreatmentSatiated     -0.1818952 0.03062840 567 -5.938776  0.0000
ArrivalTime               -0.0290079 0.00781832 567 -3.710251  0.0002
SexParMale:FoodTSatiated   0.0140178 0.03971071 567  0.352998  0.7242
SexParMale:ArrivalTime    -0.0038358 0.01019764 567 -0.376144  0.7070
```

Note, the interaction terms have been edited, to let the R printout fit on the page. Neither interaction term is significant. We could drop the least significant term, and reapply the model. Note that you should not use the `anova (M1.lme)` command as it applies sequential testing (which depends on the order of the two-way interaction terms). Its output is given below.

```
> anova(M1.lme)
                          numDF denDF   F-value p-value
(Intercept)                   1   567 252.64611  <.0001
SexParent                     1   567   1.52859  0.2168
FoodTreatment                 1   567  71.43972  <.0001
ArrivalTime                   1   567  37.13833  <.0001
SeParent:FoodTreatment        1   567   0.13472  0.7137
SexParent:ArrivalTime         1   567   0.14148  0.7070
```

The *p*-value of the last interaction term is the same as that obtained by the `summary` command. The third option, and our preferred one, is the likelihood ratio test. We need to fit the same model again, but now with ML. Both interaction terms can be dropped from the model. Using the likelihood ratio test, the significance of the dropped term is determined.

```
> M1.Full <- lme(Form, random =~ 1 | Nest,
                 method = "ML", data = Owls)
> M1.A <- update(M1.Full, .~. -SexParent:FoodTreatment)
> M1.B <- update(M1.Full, .~. -SexParent:ArrivalTime)
```

```
> anova(M1.Full, M1.A)

        Model df        AIC      BIC   logLik   Test  L.Ratio p-value
M1.Full     1  8 -0.7484292 34.41366 8.374215
M1.A        2  7 -2.6246932 28.14214 8.312347 1 vs 2 0.123736   0.725
> anova(M1.Full, M1.B)

       Model df        AIC      BIC   logLik   Test   L.Ratio p-value
M1.Full    1  8 -0.7484292 34.41366 8.374215
M1.B       2  7 -2.6103305 28.15650 8.305165 1 vs 2 0.1380986  0.7102
```

Recall that the update command takes all settings from the original `lme` command, and `-SexParent:FoodTreatment` means that this term is dropped from the model. We decided to omit the sex–food treatment interaction as it is the least significant. In the second round, we have a model that contains sex, food treatment, arrival time, and the interaction between sex and arrival time. There are two more terms that can be dropped from this model, the interaction term and food treatment.

```
> Form2 <-formula(LogNeg ~ SexParent + FoodTreatment +
                          SexParent * ArrivalTime)
> M2.Full <- lme(Form2, random= ~1| Nest, method= "ML",
                 data = Owls)
> M2.A <- update(M2.Full, .~. -FoodTreatment)
> M2.B <- update(M2.Full, .~. -SexParent:ArrivalTime)
> anova(M2.Full, M2.A)

        Model df      AIC      BIC     logLik   Test L.Ratio p-value
M2.Full     1  7 -2.62469 28.14214   8.312347
M2.A        2  6 65.52071 91.89228 -26.760355 1 vs 2 70.1454  <.0001

> anova(M2.Full, M2.B)

        Model df       AIC      BIC   logLik   Test   L.Ratio p-value
M2.Full     1  7 -2.624693 28.14214 8.312347
M2.B        2  6 -4.476920 21.89465 8.238460 1 vs 2 0.1477732  0.7007
```

The interaction term sex–arrival time is not significant so this was also omitted. The new model contains the main terms sex, food treatment, and arrival time. We dropped them each in turn and applied the likelihood ratio test.

```
> Form3 <- formula(LogNeg ~ Sex-Parent + FoodTreatment +
                            ArrivalTime)
> M3.Full <- lme(Form3, random= ~1 | Nest,
                 method = "ML", data = Owls)
> M3.A <- update(M3.Full, .~. -FoodTreatment)
> M3.B <- update(M3.Full, .~. -SexParent)
> M3.C <- update(M3.Full, .~. -ArrivalTime)
```

```
> anova(M3.Full, M3.A)

        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
M3.Full     1  6 -4.47692 21.89465   8.23846
M3.A        2  5 63.56865 85.54496 -26.78433 1 vs 2 70.04557  <.0001

> anova(M3.Full, M3.B)

        Model df       AIC      BIC   logLik   Test   L.Ratio p-value
M3.Full     1  6 -4.476920 21.89465 8.238460
M3.B        2  5 -5.545145 16.43116 7.772572 1 vs 2 0.9317755  0.3344

> anova(M3.Full, M3.C)

        Model df      AIC      BIC   logLik   Test  L.Ratio p-value
M3.Full     1  6 -4.47692 21.89465  8.23846
M3.C        2  5 29.71756 51.69387 -9.85878 1 vs 2 36.19448  <.0001
```

The term sex of the parent is not significant, and we omitted it from the model. In the next round, the model has the terms food treatment and arrival time. It is fitted with the following code. Each term is dropped in turn.

```
> Form4 <- formula(LogNeg ~ FoodTreatment +  ArrivalTime)
> M4.Full <- lme(Form4, random= ~1 | Nest,
                 method = "ML", data = Owls)
> M4.A <- update(M4.Full, .~. -FoodTreatment)
> M4.B <- update(M4.Full, .~. -ArrivalTime)

> anova(M4.Full, M4.A)

        Model df      AIC      BIC     logLik   Test  L.Ratio p-value
M4.Full     1  5 -5.54514 16.43116   7.772572
M4.A        2  4 64.03857 81.61962 -28.019286 1 vs 2 71.58372  <.0001

> anova(M4.Full,M4.B)

       Model df       AIC      BIC     logLik   Test  L.Ratio p-value
M4.Full    1  5 -5.545145 16.43116   7.772572
M4.B       2  4 28.177833 45.75888 -10.088917 1 vs 2 35.72298  <.0001
```

Both food treatment and arrival time are significant at the 5% level and we have reached the end of the model selection process.

5.10.8 Step 9 of the Protocol: Refit with REML and Validate the Model

The model that we have selected above is of the form

$$LogNeg_{ij} = \alpha + \beta_2 \times FoodTreatment_{ij} + \beta_3 \times ArrivalTime_{ij} + a_i + \varepsilon_{ij}$$

The estimated parameters are obtained by the following R code.

```
> M5 <- lme(LogNeg ~ FoodTreatment + ArrivalTime,
       random= ~1 | Nest, method = "REML", data = Owls)
> summary(M5)

Linear mixed-effects model fit by REML
       AIC      BIC    logLik
  15.07383 37.02503 -2.536915

Random effects:
 Formula: ~1 | Nest
        (Intercept)  Residual
StdDev:   0.0946877 0.2316398

Fixed effects: LogNeg ~ FoodTreatment + ArrivalTime
                    Value  Std.Error  DF     t-val p-val
(Intercept)     1.1821386 0.12897491 570  9.165648     0
FoodTrSatiated -0.1750754 0.01996606 570 -8.768650     0
ArrivalTime    -0.0310214 0.00511232 570 -6.067954     0

Correlation:
                      (Intr) FdTrtS
FoodTreatmentSatiated -0.112
ArrivalTime           -0.984  0.039

Number of Observations: 599. Number of Groups: 27
```

The slope for food treatment is -0.175. This means that sibling negotiation for an observation from an owl that was food satiated is -0.175 lower (on the log-10 scale) than a food deprived sibling. Indicating that siblings are quieter if they have more food. The slope for arrival time is -0.03, which means that the later in the night the parents arrive, the lower the level of sibling negotiation.

As to the random effects, the random intercept a_i is normally distributed with mean 0 and variance 0.09^2. The residual term ε_{ij} is normally distributed with mean 0 and variance 0.23^2. These two variances can be used to calculate the correlation between observations from the same nest: $0.09^2/(0.09^2 + 0.23^2) = 0.13$. This is relatively low, but significant (as shown by the likelihood ratio test above).

Note that there is a high correlation between the intercept and the slope for arrival. This is because all arrival values are between 22 and 30 (recall that 30 is 06.00 AM). The intercept is the value of the response if all explanatory variables are 0 (or have the baseline value for a nominal variable), which is obviously far outside the range of the sampled arrival time values. A small change in the slope can therefore have a large change on the intercept, hence the high correlation. It would be better to centre arrival time around 0 and refit all models. Something like

```
> Owls$CArrivalTime <- Owls$ArrivalTime −
                              mean(Owls$ArrivalTime)
```

will do the job. You can also use the `scale` function (with `center = TRUE` and `scale = FALSE`. In all the analyses presented in this section, you should then use `CArrivalTime`. We leave this as an exercise for the reader.

To validate the model, you should make a graph like Fig. 5.6. It is not presented here, but homogeneity seems a fair assumption. Independence will be discussed in a moment.

5.10.9 Step 10 of the Protocol

A biological discussion can be found in Roulin and Bersier (2007).

5.10.10 Sorry, We are Not Done Yet

Our optimal model contained food treatment as a nominal variable and arrival time as a continuous variable. We assumed independence because we cannot see a clear pattern if residuals are plotted versus arrival time; see also Fig. 5.6D. In Fig. 5.7, we made a multipanel plot with the `xyplot` from the lattice package. It shows the residuals of the optimal mixed effects model versus arrival time for each sex–food treatment combination. A LOESS smoother was added. This smoother should not show any pattern. Unfortunately, it raises some suspicion about a possible pattern. So, how do we know for sure there is no pattern in the residuals? The answer is to fit an additive mixed model. However, before we do this, we present the R code to make Fig. 5.7.

```
> library(lattice)
> xyplot(E2 ~ ArrivalTime | SexParent * FoodTreatment,
        data = Owls, ylab = "Residuals",
        xlab = "Arrival time (hours)",
        panel = function(x,y){
          panel.grid(h = -1, v = 2)
          panel.points(x, y, col = 1)
          panel.loess(x, y, span = 0.5, col = 1,lwd=2)})
```

The R code to make multiple panel graphs with smoothers is discussed in various case studies, e.g. Chapters 13, 14, 15, 16, 17, and 18. Note that the argument(s) on the right hand side of the '|' symbol are nominal variables. Due to the way we coded them in the data files, they are indeed nominal variables. If you coded them as numbers, use the `factor` command.

Before fitting the additive mixed model, we give the underlying equation.

$$LogNeg_{ij} = \alpha + \beta_2 \times FoodTreatment_{ij} + f(ArrivalTime_{ij}) + a_i + \varepsilon_{ij}$$

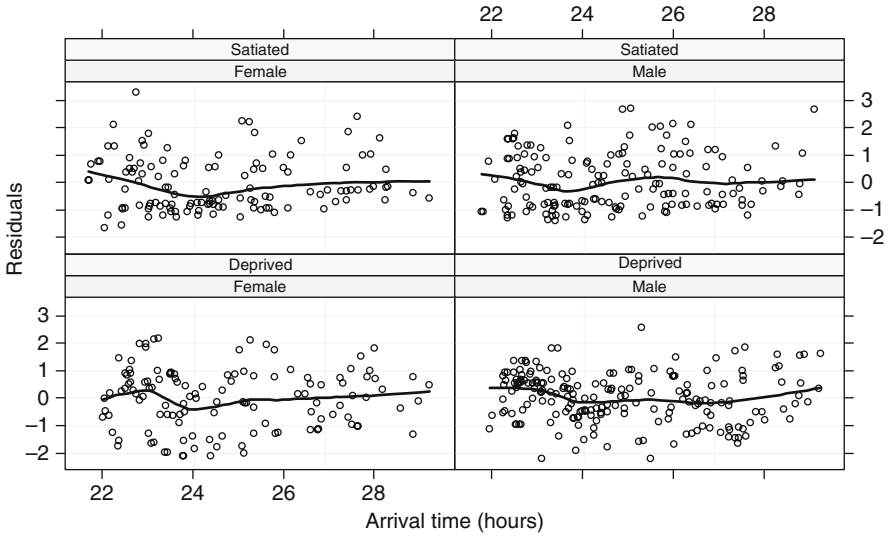

Fig. 5.7 Residuals versus arrival time for each sex–food treatment combination. A LOESS smoother with a span of 0.5 was fitted to aid visual interpretation

The term $\beta_3 \times \text{ArrivalTime}_{ij}$ has been replaced by $f(\text{ArrivalTime}_{ij})$, which is now a smoother (smoothing spline); see also Chapter 3. If the resulting shape of the smoother is a straight line, we know that in the model presented in step 9 of the protocol, arrival time has indeed a linear effect. However, if the smoother is not a straight line, the linear mixed effects model is wrong!

The following R code fits the additive mixed model.

```
> library(mgcv)
> M6 <- gamm(LogNeg ~ FoodTreatment + s(ArrivalTime),
        random = list(Nest =~ 1), data = Owls)
```

Formulation of the random intercept is slightly different and uses the list argument. Just do it, it's better not to ask why at this stage. Because no `family` argument is specified, the `gamm` function uses the Gaussian distribution. Other options are the Poisson, binomial, negative binomial, etc., and these will be discussed in Chapter 9. The output from `gamm` is slightly confusing. If you type `summary(M6)`, R gives:

```
  Length Class Mode
lme 18   lme   list
gam 25   gam   list
```

The object `M6` has an `lme` component and a `gam` component. You can use the following commands:

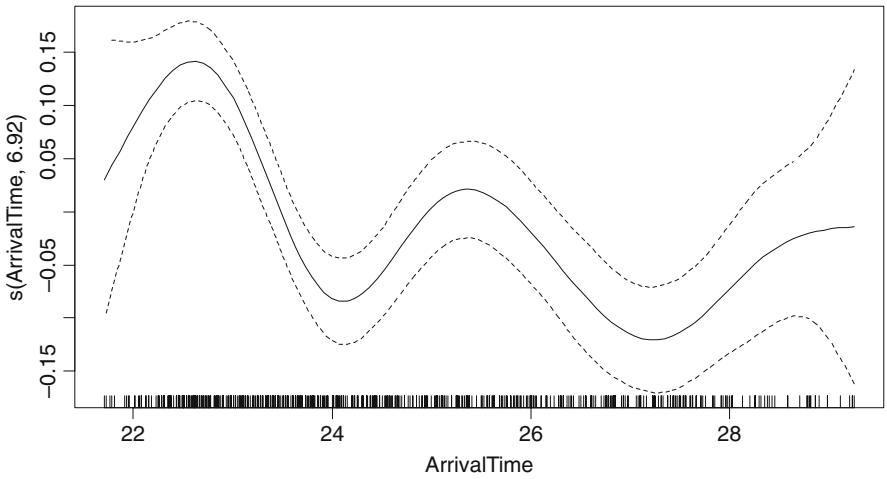

Fig. 5.8 Estimated smoother for the additive mixed model. The solid line is the estimated smoother and the dotted lines are 95% point-wise confidence bands. The horizontal axis shows the arrival time in hours (25 is 01.00 AM) and the vertical axis the contribution of the smoother to the fitted values. The smoother is centred around 0

- `summary(M6$gam)`. This gives detailed output on the smoothers and parametric terms in the models.
- `anova(M6$gam)`. This command gives a more compact presentation of the results as compared to the `summary(M6 $gam)` command. The anova table is not doing sequential testing!
- `plot(M6$gam)`. This command plots the smoothers.
- `plot(M6$lme)`. This command plots the normalised residuals versus fitted values and can be used to assess homogeneity.
- `summary(M6$lme)`. Detailed output on the estimated variances. Not everything is relevant.

Good, let us now have a look at the shape of the smoother and see whether it is a straight line or not. The command `plot(M6$gam)` produces the smoother in Fig. 5.8 and indicates that it is bad news for the linear mixed effects model; the effect of arrival time is non-linear. The `summary(M6$gam)` gives the following output.

```
Family: Gaussian. Link function: identity
Formula: LogNeg ~ FoodTreatment + s(ArrivalTime)

Parametric coefficients:
                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)       0.41379    0.02222  18.622   <2e-16
FoodTreaSatiated -0.17496    0.01937  -9.035   <2e-16
```

```
Approximate significance of smooth terms:
                 edf Est.rank     F  p-value
s(ArrivalTime) 6.921        9 10.26 8.93e-15

R-sq.(adj) =  0.184  Scale est. = 0.049715  n = 599
```

The estimated regression parameter for food treatment is the similar to the one obtained by the linear mixed effects model. The smoother is significant and has nearly seven degrees of freedom! A straight line would have had one degree of freedom.

We also tried models with two smoothers using the `by` command (one smoother per sex or one smoother per treatment), but the AIC indicated that the model with one smoother was the best.

So, it seems that there is a lot of sibling negotiation at around 23.00 hours and a second (though smaller) peak at about 01.00–02.00 hours.