5 Linear mixed-effects models and dependency

In this chapter we show how a linear mixed effects model deals with dependency. A mixed-effects model can be used when you have multiple observations from the same site, or from the same animal, or from the same person. These are repeated measurements. Other names are clustered data, hierarchical data, or panel data.

Useful references for mixed models are Pinheiro and Bates (2000), Bolker (2008), or Zuur et al. (2009a; 2013), among many others.

Prerequisite for this chapter: You need to be familiar with R and multiple linear regression.

5.1 White Storks

We use a small part of a data set provided by Boudjéma Samraoui, who studied factors affecting growth parameters of White Stork (*Ciconia ciconia*) nestlings in eastern Algeria; see also Bouriach et al. (2015). A large number of nests in a White Stork breeding colony were sampled. In each nest various chicks were sampled multiple times during the growth period from the first day of life to the maximum age of 54 days. In this section we use beak length measurements sampled in 2012. Figure 5.1 shows a scatterplot of beak length versus age.

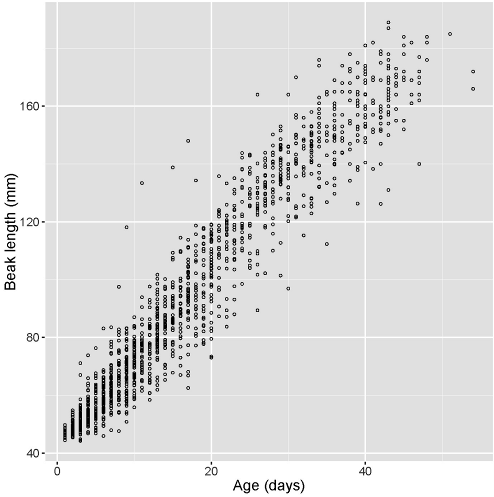

Figure 5.1. Scatterplot of beak length (mm) of White Stork chicks versus age (in days).

5.2 Considering the data (wrongly) as one-way nested

To model the growth pattern of beak length we naively specify a multiple linear regression model of the form

$$\begin{aligned} BL_i &= Intercept + Age_i + Nest_i + Chick_i + \varepsilon_i \\ \varepsilon_i &\sim N\left(0, \sigma^2\right) \end{aligned} \tag{5.1}$$

BL_i is the beak length for observation i, Age_i is the age in days, $Chick_i$ is the identity of the chick, and $Nest_i$ is the nest from which the sample is taken. We have 1438 observations. Age is fitted as a continuous covariate, and Nest (73 levels) and Chick (261 levels) are fitted as categorical covariates.

We only use this model to build up a pedagogical explanation of linear mixed-effects models. For a full analysis we should also consider adding sex of the chick (male, female, and unknown) plus an interaction between age and sex. If we were to do that in this section, the numerical and graphical output would be much larger, too much for a simple explanation of a statistical technique.

Before starting the analysis, we should think carefully what we are about to do. The model has two major problems. Firstly, due to the categorical covariates Nest and Chick we have a large number of regression parameters in the model. These two variables alone consume 334 regression parameters! And image how many more parameters we would have if we also include an interaction between the covariates Age and Chick! Secondly, the model wrongly assumes that we have independent observations; the same chick is repeatedly sampled, and chicks from the same nests are likely to have similar beak lengths. We have pseudoreplication! These two problems are reasons enough not to fit the multiple linear regression model.

One solution is to take the average per nest, but this will reduce the sample size considerably. And it is not always possible to take averages of a covariate (e.g. a categorical covariate that differs for the observations within a nest). And what is the biological relevance of modelling the average beak length as a function of the average age per nest? A mixed-effects model solves both problems.

Model formulation

We will start simple and (wrongly) ignore the chick effect for the moment (this is for pedagogical reasons; we will add the chick effect later in this chapter). A linear mixed-effects model, in which nest is used as a random intercept, is given by