21

MIXED MODELS FOR CONTINUOUS DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Write an equation for a model that contains both fixed and random components.
2. Compute the variance for each level of a multilevel model.
3. Determine how highly correlated observations are within a cluster.
4. Determine if predictors have the same (fixed), or different (random slopes) effects across clusters.
5. Compute the variance of the outcome (a complex function) in models containing random slopes.
6. Determine whether the between-cluster and within-cluster regressions for predictors have different slopes (*ie* whether contextual effects are present in the data).
7. Evaluate the statistical significance of fixed and random effects in a model.
8. Evaluate residuals from a multilevel model.
9. Determine the optimum Box-Cox transformation for the outcome in order to normalise the residuals from a model.

21.1 INTRODUCTION

Mixed models (for continuous data) contain parameters or effects of 2 types:

- **fixed**, or mean effects, such as ordinary regression coefficients in a linear regression model (Chapter 14),
- **random**, or 'variability around the mean' effects, explaining some of the error term.

Mixed models can be used to take into account that the data have a hierarchical, multilevel or nested structure, and are sometimes referred to by these terms. Although other methods exist for analysing hierarchically structured data, the use of mixed models has become a popular choice during the last decade due to advances in computing power. Multilevel models, a special type of mixed model, have been advocated as an appropriate framework for many epidemiological analyses (Diez-Roux, 2000; Greenland, 2000b) we elaborate on this in Section 21.3.4. Mixed models also apply to many other data structures, but our focus in this chapter is on hierarchical data (we discuss repeated measures and spatial data in Chapters 23 and 25-26). Mixed models are also known as **variance component models**. Variance components are the technical/ mathematical constructs used to decompose the variance (variation, variability) in a dataset into (a sum of) several components that can each be given a useful interpretation.

The dataset scc_40 (described in more detail in Chapter 31) is used to illustrate the methods numerically. It is comprised of data from 40 herds selected from a much larger dataset that was collected to study problems related to mastitis and milk yield. We will take the (log) somatic cell count (SCC) as the outcome. The data structure is 3-level hierarchical: 14,357 tests within 2,178 cows within 40 herds. The tests were performed on each cow approximately monthly throughout one lactation, and thus constitute **repeated measures** per cow. In this section, we include only a single test per cow—the first test recorded in the cow's lactation. This gives a 2-level structure of the 2,178 cows in 40 herds; herd sizes in the dataset range from 12 to 105. The 2-level dataset is denoted scc40_2level. Obviously, any inferences to real associations of predictors with the outcome should not be based on results from such subdatasets. The variables used in the examples in this chapter are listed below. For clarity, we use the term season for the quarters of the year without claiming to infer any seasonal effects from 2 years of data.

Table 21.1 Selected variables from the dataset scc_40

Variable	Level of measurement	Description
herdid	3:herd	herd identification
cowid	2:cow	cow identification
test	1:test	approximate month of lactation for test: 0,1,2,...,10
hsize	3:herd	herd size (averaged over study period)
heifer	2:cow	cow parity with values 1 (heifer) and 0 (older cow)
season	1:test	season of test with values 1 (winter: Jan, Feb, Mar), 2 (spring: Apr-Jun), 3 (summer: Jul-Sep) and 4 (fall: Oct-Dec)
dim	1:test	days 'in milk' (since calving) on test day
lnscc	1:test	(natural) log of somatic cell count

21.2 LINEAR MIXED MODEL

Linear mixed models extend the usual linear regression models (Chapter 14) of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i, \quad i = 1, \ldots, n \qquad \textit{Eq 21.1}$$

We will take as our outcome Y the log somatic cell counts and as our regressors $X_1,...,X_k$ the continuous and dummy variables necessary to represent the chosen predictors. Further, the errors $\varepsilon_1,...,\varepsilon_n$ are assumed independent and ~ $N(0,\sigma^2)$. This equation (and its assumptions) would be meaningful if we considered one test per cow and there was no clustering in herds (*eg* we might have data from only one herd). It is worth noting that, in this model, the observations $Y_1,..., Y_n$ are independent and all have the same variance:

$$\text{var}(Y_i) = \text{var}(\varepsilon_i) = \sigma^2$$

So far, the residual variance is the only variance component. However, in reality we have recordings in several (40) herds, and we would like the herds to enter our model as well, because we know that there might be some variation of cell counts across herds. Previously, we have discussed including herds in the model by adding a set of (40-1) indicator variables and estimating a separate β for each of them. A **mixed model** with a **random herd effect** is written:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_{\text{herd}(i)} + \varepsilon_i \qquad \textit{Eq 21.2}$$

The model is often termed a **random intercept model**, for reasons we'll explain later (in Section 21.3.4). **Note** For the sake of simplicity, a single index notation will be used for all multilevel data. The subscript i denotes the individual (lowest level) observation. In the equation above, $u_{\text{herd}(i)}$ refers to the herd containing the i^{th} individual (*eg* u_7 for cows in herd 7). If there are 40 herds, u could have one of 40 values: u_j, j=1,...,40. An alternative notation uses multiple indices such as $u_j + \varepsilon_{ij}$ where j refers to the herd and i to the i^{th} individual in the j^{th} herd.

The explanatory variables and the β-parameters are unchanged from Eq 21.1 to Eq 21.2. These are usually termed the **fixed effects**, in contrast to the last 2 terms which are **random effects**. The only new term in Eq 21.2 is $u_{\text{herd}(i)}$, a random herd effect for the herd of the i^{th} cow. Random simply means that it is modelled as a random variable, in contrast to a fixed parameter (according to a 'frequentist' or non-Bayesian view; see Chapter 24 for the alternative Bayesian approach). Let's defer the question as to why we model herd as a random term for now, and first look at the assumptions for u and ε:

$$u_j \sim N\left(0, \sigma_h^2\right), \quad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

where all u_j and ε_i are independent.

Thus, we assume the impact of each herd to be a random fluctuation with mean zero (and consequently centred at the mean determined by the fixed effects) and standard deviation σ_h. Therefore, the parameter σ_h^2 can be interpreted as the random variation in log cell counts between herds. Furthermore, we could calculate:

$$\text{var}(Y_i) = \text{var}(u_{\text{herd}(i)}) + \text{var}(\varepsilon_i) = \sigma_h^2 + \sigma^2 \qquad \textit{Eq 21.3}$$

In effect, we have decomposed the total variance to a sum of the variance between herds and the error variance (or the variance within herds). The σ^2s are the variance components; Example

21.1 shows how they might be interpreted. **Note** The variation accounted for by the fixed effects is not included here; one way of saying this is that Eq 21.3 is for the **unexplained variance**.

Random effects modelling of herds can be motivated in different ways. Strictly speaking it corresponds to effects (herds) in the model being randomly selected from a population. Sometimes, in a study, this could be the case, but it might be reasonable to assume that the herds are generally representative of the population even if they were not randomly selected. In our example, the 40 herds were randomly selected from the full set of study herds, which constituted all milk-producing herds in a certain geographical area of Denmark. Consequently, these 40 herds were representative of this region. With random effects, the focus shifts from the individual herd to the variability between herds in the population σ_h^2. In a study with only a few herds of particular interest (possibly because they were individually selected for the study), one might prefer to model herds by fixed effects (*ie* β-parameters) instead (as discussed in Section 20.5.2).

Mixed models can be used to take into account more general hierarchical data structures by **inserting random effects for all levels** above the bottom level (which is already present in the model as the error term). For example, a 3-level structure with animals in herds in regions would lead to random effects for both herds and regions and we then split the variation into 3 terms: $var(Y_i)=\sigma_r^2+\sigma_h^2+\sigma^2$. In mixed models, the predictors might reside at any level of the hierarchy. As a particular example, the split-plot design (Section 20.2.1) could be analysed by a mixed model with random effects for the whole-plots. In epidemiology, we often work with datasets in which predictors explain variation at several levels (Section 20.2.2); the mixed model analysis fully takes this into account. Example 21.2 shows some of the possible changes to a linear mixed model when fixed effects are included. Finally, the one exception to the 'random effects for every level' rule is that the top level could be modelled by fixed effects, if (and only if!) there are no predictors at that level. This situation often occurs when the top level (*eg* herd or region) is not a random sample of a larger population and does not have a large number of elements (*eg* Example 21.3). Some 'final' remarks on fixed vs random effects have been collected in Section 21.5.7.

Example 21.1 Variance components and random effects
data = scc40_2level

This dataset contains one observation from each of 2,178 cows from 40 herds. In a 2-level random effects model for -lnscc- with no fixed effects (a 'null' or 'empty' model'), the variance components were estimated at:

$$\sigma_h^2=0.148 \qquad \text{and} \qquad \sigma^2=1.730$$

Thus, the total (unexplained) variance was 0.148+1.730=1.878. It is often useful to compute the fractions at the different levels; here we have 0.148/1.878=7.9% of the variance between herds and 92.1% within herds. We can also give a direct interpretation of 95% of the herd effects should be within an interval of $\pm 1.96\ \sigma_h = \pm 0.754$. As the overall mean ($\beta_0$) was 4.747, this means that most herd mean -lnscc- values in the population lie between 3.993 and 5.501.

21.2.1 Intraclass correlation coefficient

The model assumptions allow us to examine the dependence or correlation between observations from the same herd. In a linear model, all observations are independent, but in mixed models this is no longer so. The correlation between observations within the same group (in our example, herd) is described by the intraclass correlation coefficient (*ICC* or ρ). For a 2-level model (Eq 21.2), the *ICC* equals the proportion of variance at the upper level; from Example 21.1:

Example 21.2 Mixed model estimates for 2-level somatic cell count data
data = scc40_2level

A linear mixed model with herd size, heifer, season and days in milk was fit to the 40-herd, 2-level scc data. The variables are explained in the table below; in addition, the herd size was standardised (-shsize-) by subtracting the approximate mean herd size (45) and dividing by 100, thus effectively measuring herd size in hundreds beyond 45. Similarly, days in milk was standardised (-sdim-) by subtracting 150 and dividing by 100.

	Coef	SE	Z	P	95% CI	
shsize	0.408	0.377	1.08	0.279	-0.331	1.148
heifer	-0.737	0.055	-13.3	0.000	-0.845	-0.628
season = spring	0.161	0.091	1.78	0.076	-0.017	0.339
season = summer	0.002	0.086	0.02	0.986	-0.168	0.171
season = fall	0.001	0.092	0.02	0.987	-0.179	0.182
sdim	0.277	0.050	5.56	0.000	0.179	0.375
constant	5.241	0.114	-	-	5.018	5.464

Note that, because of the random herd effects, the constant refers to the log somatic cell count in an average herd, not to the value of an average cow across the population of cows. As herds differ in size, these means are not necessarily the same. For example, if the highest cell counts were obtained in the largest herds (even if the -shsize- estimate hardly indicates this to be the case), then the cow average would typically be higher than the herd average. The cow and herd averages are analogous to weighted and unweighted averages in multistage sampling (Section 2.8). The other regression coefficients are interpreted in the usual way.

In addition, the estimated variance components (also with standard errors (SEs)) were:

$$\sigma^2_h = 0.149\,(0.044) \quad \text{and} \quad \sigma^2 = 1.557\,(0.048)$$

In a linear regression model, adding predictors always reduces the unexplained variation. Intuitively, one would expect a similar effect in a mixed model at the levels affected by added predictors. But, by comparison, in Example 21.1, we note a reduced value for σ^2 and a slightly increased value for It is not unusual that adding fixed effects to hierarchical models redistributes the variation across the levels and thus increases some of the variance components and, sometimes, even the total variation (the sum of all variance components). No simple intuitive explanation can be offered; see Chapter 7 in (Snijders & Bosker, 1999) for details and ways of defining measures of the variance explained by fixed effects.

$$\rho = \frac{\sigma_h^2}{\sigma_h^2 + \sigma^2} = \frac{0.148}{0.148 + 1.730} = 0.079 \qquad \textit{Eq 21.4}$$

Thus, a low *ICC* means that most of the variation is within the groups (*ie* there is only little clustering), while a high *ICC* means that the variation within a group is small relative to that between groups.

Generally in mixed models with homogeneous variances and independent random effects, correlations are assumed to be the same between any 2 observations in a group and can be computed by a simple rule. Recall (Eq 20.1) that the correlation is the ratio between the covariance of the 2 observations in question and the product of their standard deviations. As all observations have the same variance, the denominator of this ratio is always the total variance, *ie* the sum of all variance components. The numerator is obtained by noting which random effects are at the same level for the 2 observations in question, and summing the respective variance components. For the 2-level model, this rule gives Eq 21.4 for observations in the same group and zero correlation for observations in different groups. If region was added as a third level to the model, the correlation between cows in the same herd (and hence within a region) would be:

$$\rho(\text{cows in same herd}) = \frac{\sigma_r^2 + \sigma_h^2}{\sigma_r^2 + \sigma_h^2 + \sigma^2} \qquad \textit{Eq 21.5}$$

Similarly, the correlation between cows in different herds in the same region would be:

$$\rho(\text{cows in same region, but different herds}) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_h^2 + \sigma^2} \qquad \textit{Eq 21.6}$$

Example 21.3 shows similar computations for a 4-level model. The correlation in Eq 21.6 referred to cows in different herds but an intuitively more appealing value might be the correlation **between herds**—more precisely, **between herd means**. The correlation between means of 2 herds of size m is

$$\rho(\text{herds of size } m \text{ in same region}) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_h^2 + \sigma^2/m} \qquad \textit{Eq 21.7}$$

When m is large, the contribution of σ^2/m to the formula is small and might be ignored (see Example 4.7 of Snijders & Bosker (1999) for further discussion.

21.2.2 Vector-matrix notation

Notation involving vectors and matrices allows us to write the linear and linear mixed models in a compact and clear form. The linear regression model (Eq 21.1) can be written

$$Y = X\beta + \varepsilon$$

where Y, β and ε are (column) vectors and X is the so-called design matrix, comprised of a column of 1s followed by the k columns containing the values of the k predictors of the model. (**Technical Note** Our usage of X_{ji} for the element in the i^{th} row and j^{th} column of X contrasts usual matrix notation but is of no serious consequence because we do not pursue any

Example 21.3 Intraclass correlations in a 4-level mixed model
data = reu_cfs

(Dohoo *et al*, 2001) used 4-level mixed models to analyse the (log) calving to first service intervals for cattle in Reunion Island. Their model had several fixed effects which we denote $X_1,\ldots,X_k$, so that the model could be written:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{\text{cow}(i)} + v_{\text{herd}(i)} + w_{\text{region}(i)} + \varepsilon_i$$

The variance components for the unexplained variation were:

$$\text{region}: \sigma_r^2 = 0.001, \quad \text{herd}: \sigma_h^2 = 0.015, \quad \text{cow}: \sigma_c^2 = 0.020, \quad \text{lactation}: \sigma^2 = 0.132$$

Analysis of a 4-level model for the -reucfs- data with its 2 predictors (-ai- and -heifer-) gives similar estimates. The fact that the first 3 variance components were small once again points out that there is little similarity (in terms of calving to first service interval) between lactations within a cow, between cows within a herd or between herds within a region. In the original study, the authors suggested that management of reproductive performance should focus on individual lactations within individual cows, because this is where most of the unexplained variation resided.

From the estimates we could compute a total variance of 0.168 and the following correlations between observations (lactations):

lactations of the same cow:	ρ=(0.001+0.015+0.020)/0.168=0.214
lactations of different cows in the same herd:	ρ=(0.001+0.015)/0.168=0.095
lactations of cows in different herds in the same region:	ρ=0.001/0.168=0.006

As the study included only 5 regions that could hardly be considered as representative for any population of regions (together they constituted the entire island), it would be more appropriate to model regions by fixed effects. Having noted that there is virtually no variation between regions (it is far from statistical significance), one might also simply remove region effects entirely from the model.

computations with matrix notation.) Similarly, linear mixed models such as Eq 21.2 can generally be written as:

$$Y = X\beta + Zu + \varepsilon \qquad \textit{Eq 21.8}$$

where u is a vector of all random effects (except for ε) and Z is the design matrix for the random part of the model. Our assumptions for the model (in this chapter) are that all random variables are normally distributed with mean zero, and that all the errors are independent, have the same variance and are independent of the random effects.

Before we further develop the mixed models for hierarchically structured data, let's briefly indicate how mixed models can be set up for **cross-classified** data structures (Section 20.2). In the simplest cross-classified structure, every observation is classified according to 2 groupings, *eg* sires and dams in breeding data. We denote the 2 groupings by A and B. If both sires and dams are taken to represent a population, the natural model has 2 random effects in addition to the error term, as follows:

$$Y_i = (X\beta)_i + u_{\text{A}(i)} + v_{\text{B}(i)} + \varepsilon_i \qquad \textit{Eq 21.9}$$

where ($X\beta$) represents the fixed effects, and the random effects for groupings A and B are drawn from normal distributions with variances σ_A^2 and σ_B^2, respectively. In the context of analysis of experimental design data, this model is known as a 2-way random effects ANOVA model (Dean & Voss, 2000) model (and more commonly written in a 2-index notation with i

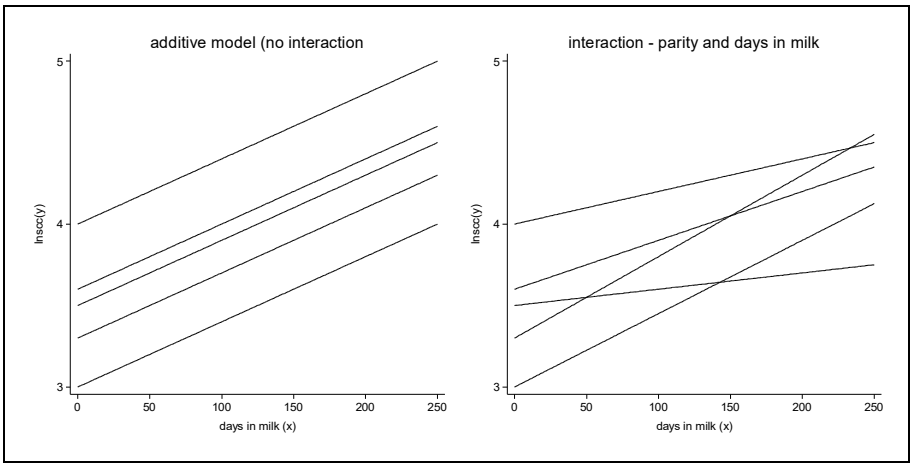

Fig. 21.1 Schematic graphs of additive and non-additive modelling of a continuous predictor (days in milk) for a continuous outcome (lnscc)

and *j* representing the factors A and B). *ICC*s can be computed by the same principles as above; *eg* the *ICC* for observations at the same level of grouping A is computed as:

$$\rho = \sigma_A^2 / (\sigma_A^2 + \sigma_B^2 + \sigma^2)$$

21.3 RANDOM SLOPES

21.3.1 Additive and non-additive modelling

As a prelude to extending the mixed model (Eq 21.2) with a random slope, we consider in more detail one implication of the model assumptions. Let's focus on a quantitative explanatory variable, for instance, days in milk. Assume these values to be in X_1, and assume the model has a linear term for X_1 with a positive regression coefficient (β_1), and no interaction terms with X_2 (parity of the cow). Then the predicted log somatic cell counts from the model for different cows in different parities, as a function of X_1 will be parallel lines, as outlined on the left in Fig. 21.1. Each line represents the predicted value for cows of a specific parity. If an interaction term between parity and days in milk was added, this would produce non-parallel lines (for different parities), as outlined on the right.

Exactly the same interpretation is valid for cows in different herds: in an additive model (Eq 21.2) the regression lines corresponding to different herds are parallel, and the random herd effects can be read as the vertical distances between the lines. This is because Eq 21.2 assumes the impact on the logarithmic cell counts of a change in days in milk (*eg* 10-day increase) to be the same for all cows in all herds (parallel lines).

21.3.2 Random slopes as non-additive herd effects

An assumption of additive herd effects (parallel lines) might not be biologically obvious because other factors such as herd management factors (inherent in the herd effects) could influence the relationship. Adding an interaction between herds and X_1 means that slopes vary

among herds. If herd was included in the model as a set of fixed effects, the interaction term would result in a specific effect being estimated for each herd. With herd as a random effect, the slopes are assumed to vary according to some distribution (in addition to the intercepts varying between herds). A model with random slopes for a single fixed effect (X_1) is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_{\text{herd}(i)} + b_{\text{herd}(i)} X_{1i} + \varepsilon_i \qquad \textit{Eq 21.10}$$

where in addition to the previous assumptions, we assume for the random slopes that the $b_{\text{herd}} \sim N(0,\sigma_1^2)$. The parameter σ_1^2 is interpreted as the variation in slopes among herds. The regression parameter β_1 is now the overall or average slope for X_1, which is then subject to random fluctuations between herds. As a rough rule, with probability 95%, the slope in a given herd would lie in the interval $\beta_1 \pm 2\sigma_1$. The choice of whether the slopes should be modelled as random or fixed effects usually follows the choice for the random effects themselves. That is, if herds are modelled as random, any slopes varying between herds should also be random. (**Note** The random herd effect, u_{herd}, and its variance, σ_h^2, now represent the variation between herds at $X_1=0$; for this to be meaningful it is necessary that zero is a meaningful value of X_1; otherwise it must be centred.)

We have not yet specified the assumptions about the relationship between b_{herd}s and the other random variables, and it is usually undesirable to assume random effects at the same level to be independent. In our example, the 2 random effects at the herd level (u_{herd} and b_{herd}) correspond to intercept and slope for the regression on X_1 at the herd level. Recall that slope and intercept are often strongly negatively correlated (although centring the variable might remove this correlation). Consequently, we usually estimate a correlation or covariance between the herd intercept and slope. It is useful to display the 3 parameters: σ_h^2, σ_1^2 and the covariance σ_{h1}, in a 2X2 matrix as follows:

$$\begin{pmatrix} \sigma_h^2 & \sigma_{hl} \\ \sigma_{hl} & \sigma_1^2 \end{pmatrix}$$

and the correlation between the herd intercepts and slopes is computed as $\sigma_{h1}/(\sigma_h\sigma_1)$. Example 21.4 shows the effect of adding a random slope to the SCC data.

21.3.3 Caveats of random slopes modelling

As intuitively appealing as the random slopes might appear, we must raise a few warning signs in their use. When the main interest is in the fixed effects, it is wise policy not to build models with too many variance parameters. In our experience, it is rarely useful to have more than one or 2 random slopes at each level in a model, and random slopes should usually only be included for statistically significant and clearly interpretable predictors; see also Section 21.3.4 below for a different perspective.

One reason why random slopes should be used cautiously is that the **variance of the model is no longer constant**. To illustrate, we compute the variance components for the random slopes model of Eq 21.10:

$$\begin{aligned} \text{var}(Y_i) &= \text{var}(u_{\text{herd}(i)}) + \text{var}(b_{\text{herd}(i)} X_{1i}) + 2\,\text{cov}(u_{\text{herd}(i)}, b_{\text{herd}(i)} X_{1i}) + \text{var}(\varepsilon_i) \\ &= \sigma_h^2 + X_{1i}^2 \sigma_1^2 + 2 X_{1i} \sigma_{h1} + \sigma^2 \end{aligned} \qquad \textit{Eq 21.11}$$

Example 21.4 Random slopes of -sdim- for somatic cell count data
data = scc40_2level

Adding a random slope of -sdim- to the model of Example 21.2 gave almost the same regression coefficient (0.273) but with a somewhat increased SE (0.061), and the random effect parameters (with SEs) were:

$$\begin{pmatrix} \sigma_h^2 & \sigma_{h1} \\ \sigma_{h1} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.130\,(0.048) & 0.0053\,(0.0246) \\ 0.0053\,(0.0246) & 0.0426\,(0.0259) \end{pmatrix} \quad \text{and} \quad \sigma^2 = 1.541\,(0.048)$$

The value of σ_1^2 suggests that 95% of the slopes for -sdim- lie roughly within 0.27±0.40=-0.13, 0.67. The correlation between intercepts and slopes is small and positive $(0.0053/\sqrt{0.130 * 0.0426} = 0.07)$ so the centring of -sdim- effectively removed the correlation. The value of σ_1^2 is only moderately larger than its SE and σ_{h1} seems totally non-significant, so it is not obvious whether the random slopes add much to the model. We will later see how to compute a statistical test for the random slopes (it is weakly significant). Note finally that a model with random slopes for -shsize- would not be meaningful; random slopes are possible only for variables at a lower level than the random effects themselves in order to be interpreted in the way we have done.

This equation involves the values of the explanatory variable X_1. In consequence, the variance is no longer the same for all observations but a function of X_1. Also, there is no longer a unique decomposition of variance in the model. For moderate magnitudes of σ_1^2 and σ_{h1} one might arrive at approximately the same decomposition of variance within the most relevant range of X_1. It is always recommended to plot the resulting variance function from a random slopes model, and if possible, convince yourself that it makes biological sense. Fig. 21.2 shows the variance function of the random slopes model for the SCC data. The dependence of the total variance on X_1 is rather weak because the major portion of the variance is at the cow/test level; nevertheless, the dependence on X_1 is biologically reasonable. Mastitis in cows is more dynamic early in lactation (so we might expect more variance in -lnscc- early in lactation) and rises again late in lactation.

Random slope models have been introduced for continuous predictors (where the relationship between Y and X is a regression). However, interactions between categorical variables and random effects are possible as well, although not interpretable as random slopes. Hence, the more general term **random coefficients** may be used instead of random slopes. As before, an additive model assumes the impact of each categorical predictor to be the same in all herds, and one might want to allow it to vary between herds. It's simplest to specify such models for a dichotomous predictor: treat its 0-1 representation as if it was a continuous variable. If the variable takes several (j) categorical values, one might create (j-1) indicator variables and proceed in the same way. Be aware that such models quickly grow to contain a lot of covariance terms, and that they could produce very different variances for the different categories. In such cases it might be useful to restrict the covariances to zero.

Example 21.5 shows the effect of adding a random slope for a dichotomous predictor in the SCC data.

21.3.4 Random slope models as hierarchical models

So far we have used the term 'hierarchical' only to describe the data structure. A hierarchical

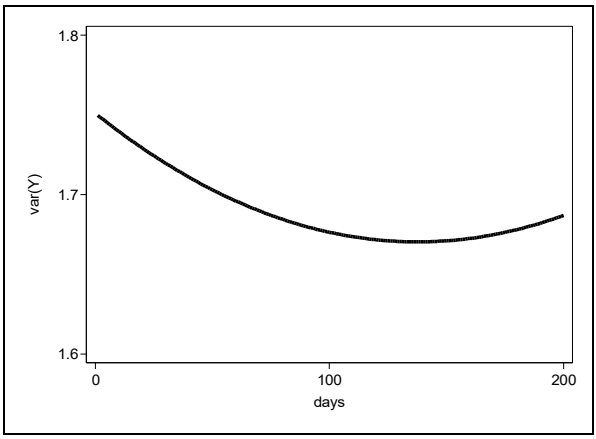

Fig. 21.2 Variance function of random slopes model for somatic cell count data

model has a more specific meaning than a model for hierarchically structured data, namely as a model with multiple hierarchical levels (see Chapter 24 for the Bayesian context). In social science and psychology applications, random slope models are often referred to as hierarchical models (Raudenbush & Bryk, 2002). We will outline the rationale behind the modelling approach by slightly rewriting the random slopes model of Eq 21.10 as:

$$Y_i = 1*(\beta_0 + u_{\text{herd}(i)} + \varepsilon_i) + X_{1i}*(\beta_1 + b_{\text{herd}(i)})$$

This model representation elucidates that every predictor can be included in the model in 3 ways (in a 2-level hierarchy): as a fixed effect, or as random effects at each of the 2 levels in the model. In the equation, the constant (1) corresponds to the intercept, and the term u_{herd} is often termed a random intercept (at the herd level), thus the name **random intercept model** for models such as Eq 21.2. A random slopes model is characterised by the fact that at least one predictor (in addition to the constant) has a higher level random effect. (**Note** Random effects of predictors at their own or lower levels correspond to heterogeneous variance models, discussed in Section 21.5.8). It is common in hierarchical modelling to include (higher level) random effects of all predictors by default, the rationale being that effects at different levels are conceptually relevant. An argument has been made for the use of random coefficient (*ie* random

Example 21.5 Random slopes of -heifer- for somatic cell count data
data = scc40_2level

Adding a random slope (of -heifer-) to the model from Example 21.2 produces a regression coefficient of -0.734(0.067) and the variance parameters (with SEs):

$$\begin{pmatrix} \sigma_h^2 & \sigma_{h1} \\ \sigma_{h1} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 0.202(0.062) & -0.076(0.042) \\ -0.076(0.042) & 0.051(0.039) \end{pmatrix} \quad \text{and} \quad \sigma^2 = 1.546(0.048)$$

The 2 variance contributions at the herd level of this model are 0.202 for non-heifers and 0.202+0.051+2*(-0.076)=0.101 for heifers. We see how the covariance is part of the variance calculation, so it should not be assumed to be zero when dealing with random slopes for categorical predictors. The data thus seem to indicate both a smaller mean and less variation of somatic cell counts for heifers than older cows. This makes biological sense based on our knowledge of mastitis.

slope) models in epidemiology (Greenland, 2000b) as a way to adjust for unmeasured confounders and achieve more realistic assessments of the population-level associations between predictors and outcome. One potential problem with (multiple) random slopes models is a lack of identifiability of variance parameters at the higher levels (where the number of units is typically not very large). Bayesian approaches (Chapter 24) to this problem have been proposed (Gustafson & Greenland, 2006), but at the current state of the methodology the best practical approach may still be a parsimonious modelling of variance (as advocated in the previous section).

21.4 CONTEXTUAL EFFECTS

Our discussion of hierarchical models introduced the idea that a predictor may be modelled with effects at multiple levels. Contextual effects add another facet to the picture, under certain conditions, by allowing for fixed effects of a predictor at higher levels than where it is recorded. The term 'contextual effect of a predictor' originates from social sciences and captures the idea that although the predictor is recorded at an individual level, its effect mostly (or entirely) relates to the group, or context, to which the individual belongs (Snijders & Bosker, 1999). We describe first a contextual effect of a predictor in a random intercept model (Eq 21.2), following (Stryhn *et al,* 2006) and then consider the extension to a random slopes model. The predictor X_1 is said to have a contextual effect if the following 2 conditions are both satisfied:

i. X_1 varies both between and within herds,
ii. the between-herd and within-herd regressions of Y on X_1 have different slopes.

Two situations where condition i. is **not** satisfied are: when X_1 is a herd-level predictor, and when the herd averages ($\overline{X}_{1\,\text{herd}}$) are constant between herds (*eg* in a clinical field trial with treatment groups equally represented within each herd). For condition ii., the within-herd regression of Y on X_1 refers to a regression equation corresponding to different animals within a single herd. Furthermore, the between-herd regression is a regression of herd mean outcomes ($\overline{Y}_{\text{herd}}$) on herd predictor means ($\overline{X}_{1\,\text{herd}}$). Fig. 21.3 illustrates situations where the between-herd and within-herd regressions of Y on a continuous predictor X_1 coincide (left-hand panel) and are completely different (right-hand panel). The within-herd regressions are indicated by straight lines (without showing individual data points), and the between herd regression is obtained by fitting a straight line to the dotted points (herd means of Y and X_1). In the right-hand panel, the within-herd slope is positive whereas the between-herd regression would have a negative slope.

We can allow for a contextual effect of X_1 in Eq 21.2 by including the herd means ($\overline{X}_{1\,\text{herd}}$) as an additional fixed effects predictor (while retaining the predictor X_1), *ie*:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \overline{X}_{1\,\text{herd}(i)} + u_{\text{herd}(i)} + \varepsilon_i, \qquad \textit{Eq 21.12}$$

where ($\overline{X}_{1\,\text{herd}(i)}$) is the X_1 mean for the herd to which subject i belongs. A contextual effect is (significantly) present when the estimate of the regression coefficient β_2 is statistically significant. If a contextual effect is present, we recommend (in order to reduce collinearity and to obtain more easily interpretable estimates) to reformulate model (21.12) by replacing the original predictor X_1 by its within-herd centred version, $Z_{1i} = X_{1i} - \overline{X}_{1\,\text{herd}(i)}$, as follows:

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \tilde{\beta}_2 \overline{X}_{1\,\text{herd}(i)} + u_{\text{herd}(i)} + \varepsilon_i \;, \qquad \textit{Eq 21.13}$$

The equations (21.12) and (21.13) represent the **same** model, and the coefficients for X_1 and Z_1

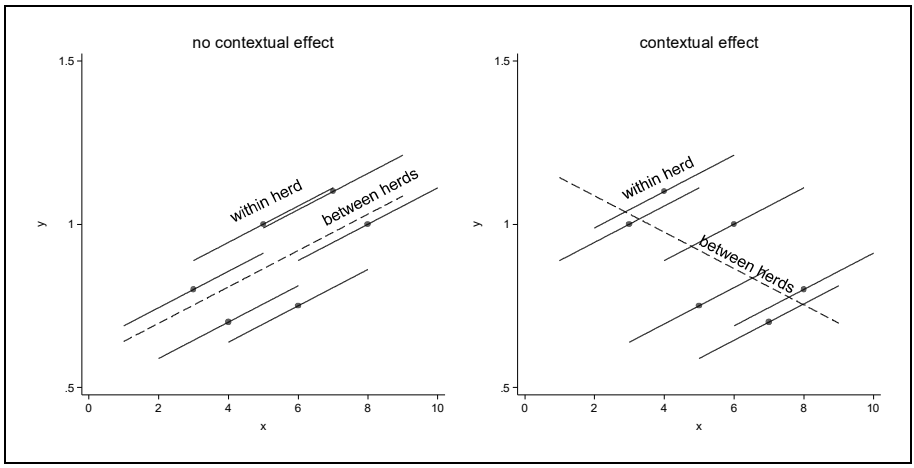

Fig. 21.3 Schematic graphs showing no contextual effect (left) and a strong contextual effect (right) of the predictor X_1

are identical (β_1), whereas $\tilde{\beta}_2 = \beta_1 + \beta_2$. The parameter $\tilde{\beta}_2$ is the slope of the between-herd regression of Y on X_1 (*ie* between the corresponding herd means, as explained above) and the parameter β_1 in models (21.12) or (21.13) is the slope of the within-herd regression of Y on X_1. Example 21.6 shows how these models can be fit in a dataset on somatic cell counts of heifers.

As demonstrated in the example, contextual effects may also be incorporated into random slopes models, by adding the herd-averages of the predictor into the model equation in the same way as we did in Eqs 21.12 and 21.13. One should be aware that the 2 parametrisations above lead to different models if the fixed and random slopes use the same version of the predictor (X_1 and Z_1, respectively). The validity of using the group-mean centred predictor Z_1 has been discussed in the literature (*eg* Hox, 2002, Section 4.3); a practical approach is to explore both models and compare their fit to the actual data.

In summary, it is important to realise the presence of contextual effects for a problem, because the within- and between-group regressions may represent different effects, and therefore often have different interpretations. In the presence of a contextual effect, the single regression coefficient in model (21.2) is a complex function (under certain conditions: a weighted average) of the 2 slopes β_1 and $\tilde{\beta}_2$ and difficult to interpret (see Section 3.6 of Snijders & Bosker, 1999, for details). Failure to account for contextual effects may lead to conclusions based on either ecological or atomistic fallacies (Chapter 29).

21.5 STATISTICAL ANALYSIS OF LINEAR MIXED MODELS

In mixed models there are several methods of analysis, and the principal estimation procedure, which is based on the likelihood function (Section 21.5.1), does not have closed-form expressions for the estimates but involves running several steps of an estimation algorithm. This requires some extra attention to the statistical software by the researcher to ensure that it employs the desired estimation procedure and to ensure that it is capable of analysing the data at hand. Statistical software differ in the range of models that can be analysed, in their ability to handle large data structures (many units at any level beyond the lowest one) and in their user

Example 21.6 Contextual effects for somatic cell count data for heifers
data = scc_heifer

A study on the predictive value of early lactation records of somatic cell counts for outcomes later in lactation was based on data for heifers in Belgian herds collected during the years 2000-01 (De Vliegher *et al*, 2004). Approximately monthly recordings throughout the first lactation (until dry-off) were recorded for each heifer, but we consider only a single record for each heifer (the one obtained during 76-105 days in milk). This subset of the data comprises 10,996 heifers in 3,095 herds. The predictor of primary interest was the (natural) log SCC in early lactation (days in milk 5-14; -lnsccel-), and our focus here is on predicting log SCC (-lnscc-) later in lactation. We show 4 mixed models with this single predictor and herd random effects; the analysis of (De Vliegher *et al*, 2004) included additional predictors such as the yield, season, the breed, and the days in milk at which the early lactation SCC was obtained, and Stryhn *et al* (2006) presented results for contextual effects of both log SCC and yield.

Model	Random intercept		Contextual		Random slopes		Contextual + random slopes	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β_0(intercept)	4.095	0.012	4.094	0.012	4.095	0.012	4.095	0.012
β(lnsccel)	0.262	0.008	0.242	0.009	0.271	0.009	0.248	0.010
β(hlnsccel)	-	-	0.080	0.019	-	-	0.089	0.019
σ^2(herd)	0.121	0.011	0.118	0.011	0.118	0.011	0.115	0.011
σ^2(lnsccel)	-	-	-	-	0.021	0.005	0.021	0.005
σ^2(heifer)	1.037	0.016	1.038	0.016	1.007	0.017	1.007	0.017

The variable -hlnsccel- contains the herd means of -lnsccel-. The strong significance of the contextual effects in both the random intercept and random slopes models is indicated by estimates of -hlnsccel- being much larger than their SEs. In the random intercept model with contextual effects, the within- and between-herd regression slopes are estimated at 0.242 and 0.322 (computed as 0.242+0.080), respectively. The single slope (0.262) in the random intercept model, therefore, mostly represents the within-herd regression, and would most likely also be interpreted as such. The added strength of the between-herd regression can probably be attributed to a herd management effect: in herds with heifers that generally have low early lactation SCCs, the SCCs are also lower later in lactation, in both cases a reflection of good herd management. In this case, the contextual effect adds to the interpretation of the single regression coefficient without altering its (within-herd) interpretation. Stryhn *et al* (2006) presented an example where the contextual effect changed the interpretation.

Adding the strong random slopes for -lnsccel- to the model does not substantially change the estimates or SEs of the within- and between-herd regressions. In summary, the effect of -lnsccel- can therefore be described as a composite of 3 terms, as follows. On average, an increase in -lnsccel- of 1 unit is associated with an increase of 0.248 units for the same animal 76-105 days into the lactation. The 95% range across cows for this effect is, however, fairly wide: $0.248 \pm 2 * \sqrt{0.021} = 0.248 \pm 0.290$. In addition, herds with a 1 unit higher average -lnsccel- could expect an added 0.089 units increase in -lnscc-, for a total of 0.337(0.248+0.089) units remaining of the initial elevation after 76-105 days.

interface. Specialised hierarchical or multilevel software has been developed to deal with huge data structures; a good source of information is the website of the Centre for Multilevel Modelling at the University of Bristol, UK (http://www.cmm.bristol.ac.uk). As of winter 2009,

the main software options (with corresponding texts providing theory, examples and code) were (in unstructured order): Stata (Rabe-Hesketh & Skrondal, 2008), S-Plus/R (Gelman & Hill, 2006; Pinheiro & Bates, 2000), SAS (Littell *et al*, 2006), as well as the 2 multilevel packages MLwiN (with a wealth of material at the above-mentioned website) and HLM (Raudenbush & Bryk, 2002).

In most ways the mechanics of the analysis of linear mixed models is similar to the analysis of linear models, because the actual estimation procedure is taken care of by the software program, which also outputs many of the same quantities (*eg* estimates and SEs, tests of individual parameters and confidence intervals, as already shown in Example 21.2).

21.5.1 Likelihood-based analysis

Parameter estimation in normal linear mixed models is based on the likelihood function derived from the normal distribution assumptions. Roughly speaking, the likelihood function for any set of parameters gives the 'probability' of the observed data under that set of parameters (see Section 16.4). Then it is intuitively reasonable to seek the set of parameters that maximises this probability—the maximum likelihood estimates. Because of the complicated form of the likelihood function, closed-form formulae for the maximum likelihood estimates generally do not exist. Therefore, parameter estimation employs an **iterative procedure** in which tentative estimates are gradually improved from their starting values to final convergence. As with all iterative procedures, caution must be exercised so that convergence is achieved. The estimation software should take care of this, but any messages that the iterative procedure has not converged are true causes for alarm. If the iterative procedure fails to converge, it sometimes helps to provide sensible starting values of the variance parameters; however, most commonly it signals a misspecified model. The advanced user may also attempt to tune the estimation procedure by some of the settings that control the algorithm. For example, without going into the technical details, several current estimation procedures perform initial iterations by an EM algorithm before switching to Newton-Raphson optimisation, and it could be useful to change the default number of iterations of the EM algorithm before the switch.

Two variants of maximum likelihood estimation are available for mixed linear models: genuine **maximum likelihood** (ML) (also known as **full information maximum likelihood** or FIML) and **restricted maximum likelihood** (REML) estimation. From a theoretical point of view, REML estimates are unbiased, whereas ML estimates often have less variance; the weighting of these properties is not straightforward, but in practice the difference is usually negligible. Both variants give 'asymptotically correct' values (*ie* when the number of observations at all levels of the hierarchy grows very large) and enable a full mixed model statistical inference. Therefore the choice between the 2 is essentially a technicality and a matter of taste; in the authors' experience, REML is the more commonly used. All results shown in this chapter are based on REML estimation unless explicitly stated otherwise.

Before proceeding with the statistical inference based on the likelihood function, it is worth mentioning an estimation approach based on the ANOVA table (Dean & Voss, 2000, Chapter 17). It is simpler to implement and offered by more software packages. By and large, this approach is obsolete by today's standard, but in **balanced datasets** it will give the same estimates for the variance components and similar statistical tests for fixed and random parameters as the REML analysis. A dataset is balanced when every combination of predictor

values ('treatments') occurs the same number of times in the data. While this is frequently the case in experimental, factorial designs, it is rarely so in observational studies (in particular, if the data contain continuous predictors). The idea of the method is to compute variance components as linear functions of the mean squares of the ANOVA table, suitably chosen to make the variance component estimates unbiased. Therefore, closed-form expressions are available and they require little calculation beyond the ANOVA table. Thus, the method is an add-on to a fixed effects analysis rather than a 'real' mixed models analysis, and herein lies its drawback: not all aspects of the statistical inference are managed correctly, *eg* standard errors are not readily available.

One particular example of an ANOVA-based method is still in quite common use—estimation of the *ICC* for a 2-level structure from a one-way ANOVA using the formula:

$$\rho \approx \frac{\text{MSM} - \text{MSE}}{\text{MSM} + (m-1)\,\text{MSE}} \qquad \textit{Eq 21.14}$$

where m is the (average) number of observations per group. If the groups are all of the same size (balanced data), this gives the same value as computing the *ICC* from likelihood-based (REML) variance components using Eq 21.4. When the data are unbalanced, the likelihood-based estimate is preferred. For the 2-level SCC data, the above formula yields ρ=0.076; Eq 21.4 gives a value of 0.079.

21.5.2 Inference for fixed part of model

The reader may have noted a z (standard normal) reference distribution for tests and confidence intervals in Example 21.2, in place of the usual t-distribution in linear models (Chapter 14). This reflects that the statistical inference is no longer exact but approximate, and the approximations are only 'asymptotically exact'. When the number of observations grows very large (at all hierarchical levels), the reference distribution approaches a standard normal distribution—thus one option for the reference distribution. However, with small or moderate numbers of observations at some of the hierarchical levels, a standard normal distribution might be too liberal (or 'anticonservative') as the reference, because it overestimates the degrees of freedom. Some software programs offer a finite sample approximation (*eg* Satterthwaite approximation) based on a t-distribution with degrees of freedom reflecting the design and the parameter under consideration; Schaalje *et al* (2002) studied the performance of several approximate reference distributions available in SAS Proc Mixed. With a reference distribution in place, tests and confidence intervals are computed in the usual manner, *eg* a 95% confidence interval of $\beta_1 \pm t(0.975,\text{df})\text{SE}(\beta_1)$.

Approximate tests computed from the estimate and its SE are usually termed **Wald tests** (see Section 6.5.2), and a multiple version exists for tests involving several parameters, *eg* for several indicator variables of a categorical variable. Tests based on comparing the attained value of the likelihood function (**Note** It is invalid to use the restricted likelihood from REML!) in models with and without the parameter(s) of interest are possible as well but usually offer little advantage over Wald tests, and we leave them to the next section. Pinheiro & Bates (2000), Section 2.4.2 recommend against the use of likelihood-based tests with chi-square reference distributions because of their overestimated degrees of freedom (as discussed above). Example 21.7 illustrates the inference for fixed effects in the SCC data.

Example 21.7 Fixed effects for 2-level somatic cell count data
data = scc40_2level

A multiple Wald test for the combined effect of season gives $\chi^2(3)=6.21$ and a P-value of 0.10; thus, there is no significant differences between seasons (in this subdataset). Analysis by SAS Proc Mixed or R (lme library) with finite sample reference *t*-distributions with about 2,100 degrees of freedom for the coefficients for -heifer-, -sdim- and -season- which corresponds roughly to the residual degrees of freedom at the cow/test level. With such large degrees of freedom there is no difference between *t* and *z* distribution inference.

The finite sample reference distribution for -shsize- is $t(38)$, reflecting that it is a herd-level predictor, and that the 40 herds would leave only 38 degrees of freedom for the herd-level residual. Therefore, the effect of -shsize- is estimated with considerably less precision than the other predictors, and not surprisingly, it shows up clearly non-significant. With the *t* reference distribution, the P-value for Wald test is only increased slightly to 0.286; considerably smaller degrees of freedom (say 10 or less) are needed to substantially affect the inference.

21.5.3 Inference for random part of model

Even though the software usually outputs both variance parameters and their SEs, the latter should not be used to construct Wald-type confidence intervals or tests, because the distribution of the estimate can be highly skewed.

Variance parameters can be tested using likelihood-based (**likelihood ratio**) tests, although we usually retain random effects corresponding to hierarchical levels despite their non-significance (unless the variance is estimated to be zero). To illustrate, a likelihood ratio test in Eq 21.2 for the hypothesis H_0: $\sigma_h=0$ is calculated as $G^2=-2(\ln L_{full}-\ln L_{red})$ where the full and reduced models refer to the models with and without the herd random effects, and L refers to values of the likelihood function. Either ML or REML likelihood functions might be used, provided both models contain the same fixed effects. Generally, the value of G^2 is compared with an approximate χ^2-distribution with the degrees of freedom equal to the reduction in number of parameters between the 2 models. Snijders & Bosker (1999), Section 6.2 note that reference χ^2-distributions are conservative when testing a variance parameter being equal to zero, and recommend **halving the P-value** obtained from the χ^2-distribution to take into account that the alternative hypothesis is one-sided (H_a: $\sigma_h>0$). Most software packages apply this correction by default for testing a random intercept variance. The same procedure (halving the P-value obtained from a nominal χ^2-distribution) applies to tests for random slopes (Berkhof & Snijders, 2001). If there is only a single random slope in the model, the test for the random slope involves 2 parameters (the variance and covariance), so the nominal degrees of freedom is 2. Example 21.8 demonstrates these calculations for the SCC data. If the comparison is to a random slopes model instead of a random intercept model (*eg* for testing one out of 2 random slopes present in the same model), the reference distribution becomes more complicated (see Fitzmaurice *et al*, 2004), Section 8.5, for recommendations and a table of critical values for some settings). The choice of the random part of the model may also be based on model selection statistics such the AIC (Section 15.8.1). The penalty for the model's parameters now include the variance and covariance of the random part. Use of the BIC is not recommended for covariance selection unless one works in a Bayesian framework (Fitzmaurice *et al*, 2004, Section 7.5).

For random effect parameters, symmetric confidence intervals are usually inappropriate. If your software can display the variance estimates at the scale at which they are estimated (behind the

scenes, so to speak), it is better to compute a confidence interval at that scale and transform its endpoints; this may also be the default method in your software. Two alternative methods are suggested in the literature: bootstrapping (Goldstein, 2003, Section 3.6) and profile-likelihood intervals (Longford, 1999). Bootstrapping is a general statistical technique primarily aimed at estimating standard errors and calculation of confidence intervals in situations too complex for analytical methods to be manageable; however, bootstrap confidence intervals require specialised software (*eg* MLwiN). In brief, a profile-likelihood confidence interval (with approximate 95% coverage) includes the values (σ^*) of the parameter, for which twice the log-likelihood with the parameter under consideration fixed at the particular value (*ie* $\sigma=\sigma^*$), drops less than 3.84 (the 95% percentile in $\chi^2(1)$) from twice the log-likelihood value of the model. If your software allows you to fix a variance in the model, a crude search for such parameter values is simple to carry out. Example 21.8 illustrates the inference for random parameters in the SCC data.

21.5.4 Prediction

Even though the random effects in a mixed model are not parameters in the usual sense, it is

Example 21.8 Random effects for 2-level somatic cell count data
data = scc40_2level

The table below gives values for twice the log likelihood function (based on REML) for various somatic cell count models in this chapter and likelihood-ratio test statistics for model comparisons (comparing all models with the one presented in Example 21.2 (random intercept model)). Note that P-values were computed manually by halving the tail probabilities of the respective chi-square distributions.

Model	2lnL	AIC	G^2	df	P-value
no herd random effect	-7328.51	7344.51	97.01	1	0.000
random intercept (Ex 21.2)	-7231.50	7249.50	-	-	-
random slope of -sdim-	-7225.48	7247.48	6.02	2	0.025
random slope of -heifer-	-7225.71	7247.71	5.80	2	0.028

The table shows strong evidence against the hypothesis of no (random) variation between herds, and it also shows that extensions of the model with random slopes for -sdim- and -heifer- are both weakly significant. Based on these results, it would be logical to explore a model with random slopes for both -sdim- and -heifer-, but we stop here.

The 95% confidence interval for σ_h^2 provided by the software for the model of Example 21.2 was (0.084,0.265). It is asymmetric around the estimate (0.149) and based on transformation from log square-root transformed scale. The estimation command does not offer profile-likelihood intervals or to fix parameter values. To illustrate the profile-likelihood method, to assess whether a given value (say 0.20) belongs to the confidence interval, estimate the model with σ_h^2 fixed at 0.20, obtain the model's 2lnL value (-7234.48), which is still within 3.84 of the model's value (-7231.50); therefore, the value 0.20 belongs to the 95% confidence interval. The profile-likelihood CI obtained by a crude search was (0.085,0.269), which is very close to the above interval computed by transformation from estimation scale.

possible to give estimates (more precisely, **predictions**) of their values. These carry the names **best linear unbiased predictors** (**BLUP**s), referring to their inherent statistical properties, and **empirical Bayes estimates** (Greenland, 2000a), referring to an interpretation of the way they are computed. The prediction may be useful *eg* for the purpose of ranking the units with random effects (herds, or schools in education studies, or hospitals in human public health studies), or for identification of extreme values (discussed in the next section). The statistical inference for rankings and comparison of predictions for 2 units (*eg* for the purpose of significance testing) has been described (Goldstein & Spiegelhalter, 1996). Because of the assumed common (normal) distribution of the random effects (in Bayesian terminology a prior distribution, see Chapter 24) the predictions are more regular (*ie* less variable) than the estimates one would obtain from a fixed effects model; this phenomenon is referred to as **shrinkage** (towards the overall mean). The amount of shrinkage depends on the magnitude of the variances and the group sample size: small groups are shrunk more towards the overall mean, and the shrinkage is weaker in datasets with a high *ICC* (because, if the between-group variation is large, the other groups contribute relatively little information about the level of any specific group). Under simplified assumptions (Snijders & Bosker, 1999, Section 4.7), the empirical Bayes estimate is a weighted average of the group mean and the overall mean, and the weight of the group mean (called the shrinkage factor) equals $\sigma_h^2/(\sigma_h^2+\sigma^2/m)$, where m is the group size. It is seen that this formula has the qualitatively behaviour just described; for example, if m is large, the weight is close to one, and the predicted value is close to the group mean (*ie* no shrinkage).

21.5.5 Residuals and diagnostics

Residuals and diagnostics play a similar, crucial role for model-checking in mixed models as they do in ordinary linear models. The mechanics and interpretations are analogous (see Sections 14.8 and 14.9), but the additional model assumptions (for the random effects) should be evaluated critically together with the other assumptions. Accordingly, mixed models contain additional 'residuals'—one set of residuals per random effect in the model. (**Note** Be aware that residuals at the different hierarchical levels contain different numbers of observations; *eg* the SCC dataset has only 40 herd-level residuals.) The residuals include not only the effects for the hierarchical levels but also the random slopes, *ie* in a model with random intercepts and slopes, there are 2 sets of residuals at the corresponding level. These residuals are, in reality, predicted values of the random variables in the model (as discussed in the previous section). In the usual sense, residuals are differences between observed and expected values; however, there are no observed herd values here, so the term predicted values seems preferable. Influence diagnostics are also computed at each hierarchical level and for each random effect. Recent advances in software for multilevel analysis have given access to residuals and diagnostics in many major software packages, although some differences in implementation exist, in particular with respect to the definition of standardised residuals (see Skrondal and Rabe-Hesketh, 2009 for a detailed discussion of this topic). A case study of model-checking using residuals and diagnostics (Langford & Lewis, 1998) recommended to first inspect the residuals at the highest hierarchical level, and then gradually work downwards. Thus, before looking at individual cows being influential or not fitted well by the model, we examine the same questions for the herds. This is because several of the cows being flagged could stem from the same herd, so the 'problem' might be with the herd rather than with the individual cow. Example 21.9 presents herd-level residuals and diagnostics for the SCC data.

Example 21.9 Residuals and diagnostics for somatic cell count data
data = scc40_2level

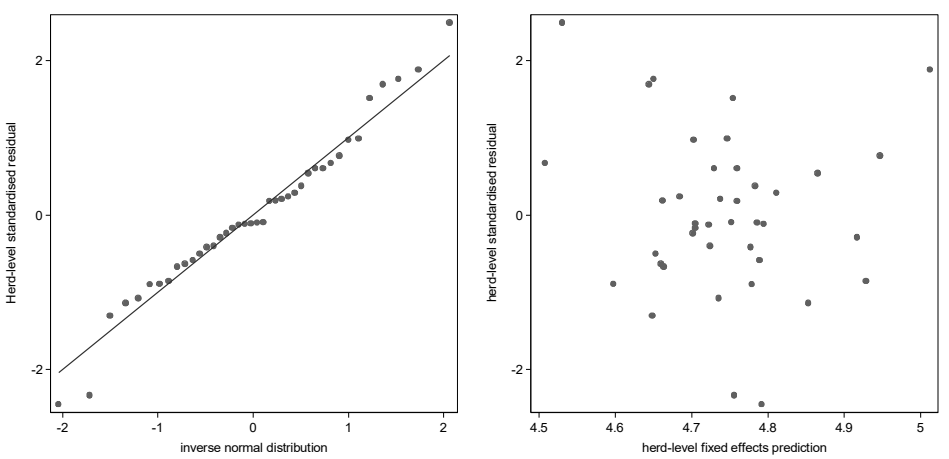

Fig. 21.4 Quantile plot (left) and residual plot (right) for herd-level residuals

We present here herd-level residual plots and a listing of the residuals and diagnostics for the 6 most extreme herds; the analysis of cow-level residuals and diagnostics follows similar lines as in Chapter 14. The computations were done mostly using Stata software; the leverages and DFITS values were computed by the MLwiN software which also gave slightly different standardised residuals (not shown).

herd number	herd size	raw residual	standardised residual	Cook's distance	leverage	DFITS
40	37.7	-0.831	-2.453	1.287	0.113	0.405
7	37.4	-0.787	-2.335	1.361	0.117	0.389
8	35.1	-0.445	-1.309	0.441	0.114	0.204
…		…	…	…	…	…
32	43.3	0.600	1.760	0.549	0.103	0.264
6	84.0	0.666	1.880	1.985	0.130	0.344
18	10.3	0.688	2.488	1.570	0.300	0.712

The quantile plot of the standardised residuals did not indicate any serious deviations from the normal distribution, nor did the residual plot reveal any concerns. Based on the residuals and diagnostics, herd 18 stands out somewhat with the highest values of residuals, leverage and DFITS. The magnitude of the residuals is hardly anything to worry about, but the influence seems appreciable. When analysing the data without this herd, the effect of -h_size- increases by more than 50% and approaches significance. Herd 18 turns out to have the smallest value of -hsize-, but the highest average -lnscc-. The high value of Cook's distance for herd 6 is related to this herd being among the largest and also having the second highest average -lnscc-; without herd 6 the coefficient for -hsize- drops to about 1/3 of its value. The strong influences of single herds on the coefficient for -hsize- can be attributed to the fairly small sample size at the herd level.

21.5.6 Box-Cox transformation for linear mixed models

In Section 14.9.3, we discussed the Box-Cox method of choosing the 'best' power (λ) transformation of our data to match the assumptions of a linear model. We assumed the method to be implemented was available software and did not go into details with how the optimal λ was calculated. A Box-Cox analysis is however, to our knowledge, not readily available elsewhere for mixed models, so we give the necessary details to enable the analysis for transformation of the outcome. The Box-Cox transformation in principle takes all model assumptions into account, but in our experience it is most sensitive to the assumptions at the lowest level.

Recall that we confine the analysis to a set of 'nice' λ-values, *eg* for a right-skewed distribution, we might search for the best value among λ=1, 1/2, 1/3, 1/4, 0, -1/4, -1/3, -1/2, -1, -2. Among these, λ=1 corresponds to no transformation, λ=0 to natural log transformation, and λ=-1 to reciprocal transformation. Finding the approximate optimal λ-value involves the following steps:

1. compute the mean of the ln(Y)-values and denote this value by $\overline{\ln(Y)}$; also denote the total number of observations as n,
2. for each candidate λ-value, compute for each observation i the transformed value

$$Y_i^{(\lambda)} = \begin{cases} (Y_i^{\lambda} - 1)/\lambda & \text{for } \lambda \neq 0 \\ \ln(Y_i) & \text{for } \lambda = 0 \end{cases}$$

   and analyse these $Y(\lambda)$-values by the same mixed model as the untransformed values, and record the model's attained log-likelihood (lnL(λ)) value using ML estimation (not REML!),

3. compute the value of the profile log-likelihood function as:

$$\text{pl}(\lambda) = \ln(L^{(\lambda)}) + n(\lambda - 1)\overline{\ln(Y)} \qquad \textit{Eq 21.15}$$

   and plot the function to identify approximately the λ where pl(λ) is maximal. This is the optimal power transformation of the outcome. An approximate 95% confidence interval for λ consists of those λ-values with a value of pl(λ) within 1.92 of the optimal pl-value.

We demonstrate the procedure in Example 21.10 using the SCC data.

Recall (from Chapter 14) that the optimal Box-Cox value does not guarantee 'well-behaved' residuals (at all hierarchical levels), and that transformation could shift problems from one model assumption to another (*eg* from skewed residuals to heteroscedasticity). Therefore, even after transformation, all the residuals should be examined. If well-behaved residuals at some hierarchical level cannot be achieved by transformation, one might turn instead to models with non-normal random effects; such models are available within the Bayesian framework for hierarchical models (Chapter 24), or rely on the robustness of the linear mixed model procedures to model misspecification (Section 21.5.8).

21.5.7 Model specification: fixed versus random effects

In this section, we will discuss a test to compare estimates based on fixed and random effects, and summarise the choice between these 2 models. In econometry, it is a commonly used procedure to assess the adequacy of a random effects model by a 'Hausman specification test'.

Example 21.10 Box-Cox analysis for somatic cell count data
data=scc40_2level

The data contain n=2,178 observations and the mean (natural) logarithmic cell count is 4.757. The following table and graph give a Box-Cox analysis:

λ	1	0.5	0.33	0.25	0
ln(L) for $Y^{(\lambda)}$	-17247.28	-9807.27	-7585.75	-6532.35	-3604.68
pl(λ) from(21.15)	-17247.28	-140987.63	-14493.24	-14302.89	-13965.84
λ	-0.10	-0.25	-0.33	-0.5	-1
ln(L) for $Y^{(\lambda)}$	-2530.05	-1014.68	-219.64	1282.26	5236.52
pl(λ) from(21.15)	-13926.84	-13965.57	-14033.58	-14258.82	-15484.91

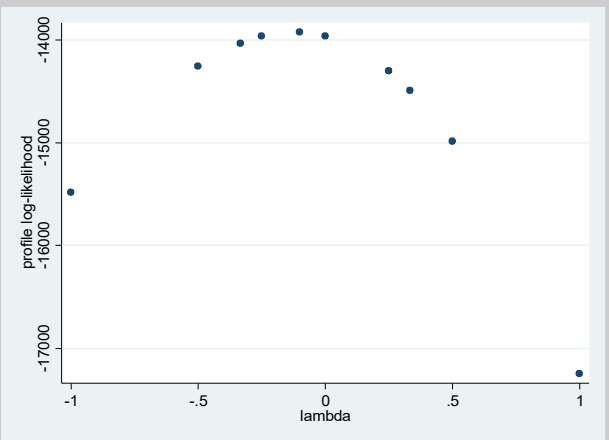

Fig. 21.5: Profile-likelihood function for Box-Cox analysis of SCC data

The table and figure indicate the optimal value of λ to be close to, and slightly less than, zero, but a 95% CI for λ does not include zero; the large number of lowest-level observations causes the CI to be very narrow. With the optimal transformation so close to the log-transformation, the Box-Cox analysis supports our choice of analysing the log somatic cell counts, in the sense that no power transformation improves the compliance with model assumptions substantially. Analysis of the power transformed (λ=-0.10) SCC values instead of -lnscc- reduced the skewness of the cow-level residuals, but did not substantially change the inference (results not shown).

The Hausman test is a general procedure for comparing 2 estimates where one is asymptotically valid under more general conditions. The rationale for preferring a fixed effects model would be that one of the (implicit) assumptions of the random effects model, that the random effects are independent of the predictors (X), is invalidated (the predictor in question is then termed 'endogenous'). However, Skrondal and Rabe-Hesketh state this to be misguided because the test is really for a contextual effect of one of the predictors, and if the test is significant one should instead insert the missing contextual effect into the random effects model (Rabe-Hesketh & Skrondal, 2008, Section 3.2.1). Moreover, the Wald test for contextual effects discussed in

Section 21.4 remains valid. To illustrate, a Hausman specification test for the model of Example 21.2 gave $\chi^2(5)=4.33$, which is absolutely non-significant in a χ^2 distribution with 5 df (P=0.50). We leave it as an exercise for the reader to verify that none of the predictors in the model have a contextual effect.

In our view, random effects for hierarchical levels are usually preferable, but fixed effects modelling is occasionally a useful approach to account for clustering in herds (say), particularly when:

i. there are no herd-level predictors,
ii the number of herds is reasonably small, and
iii there is more interest in the specific herds than assuming they represent a population.

A more technical comparison of fixed and random effects modelling can be found in (Rabe-Hesketh & Skrondal, 2008, Section 3.8).

21.5.8 Robustness against model misspecification

In addition to endogeneity (discussed above), the most obvious violations of the assumptions of (standard) linear mixed models are heteroscedasticity and non-normality of random effects (including the error term). Recent research has examined the robustness of estimation procedures based on (standard) linear mixed models to such model misspecifications. One obvious idea is to adjust linear mixed model estimates by robust variance estimation (Section 20.5.4). Now the purpose is not to account for clustering (the mixed model already does that), but to achieve robustness against heteroscedasticity and non-normal error distribution (Hox, 2002, Section 11.2). It is known that estimates of regression coefficients are robust to misspecification of the random effects distribution (McCullagh *et al*, 2008, Section 12.3), so variance adjustment may be all that is needed. Although robust variance estimation cannot guarantee against strong violations of model assumptions, they may constitute a substantial improvement, in particular for SEs of variance parameters (Verbeke & Lesaffre, 1997) and also may be used as a diagnostic tool (*ie* large differences between robust and model-based SEs are taken to indicate problems with model specification (Maas & Hox, 2004)). The robust standard errors are usually implemented to follow the hierarchical structure (clustered at the highest level), so their efficiency depends on a reasonable number of clusters (sample sizes are discussed in the next section). Adding robust standard errors to the linear mixed model analysis of Example 21.2 leads to moderate increases in SEs (1%-21% for fixed effects, 11%-33% for variances; results not shown); the robust standard errors will give a more cautious analysis, at the cost of some loss of power. Non-parametric and semiparametric specifications of the random effects distribution have been studied but are not readily available in standard software and also have their drawbacks (McCullagh *et al*, 2008, Section 12.4). Bayesian modelling can incorporate other random effects distributions than the normal, *eg* a *t*-distribution (Chapter 24).

One of the strong points of linear mixed models is that they allow heteroscedasticity to be built directly into the model. We have already seen that random slopes models are heteroscedastic (*ie* the variance depends on the predictors). Such modelling may be preferable to adjustments by robust standard errors because it provides extra information about the data that perhaps can lead to better understanding of the causal mechanisms and can also be used to obtain better predictions (Fitzmaurice *et al*, 2004, Section 11.3). Direct modelling of heterogeneity of the lowest level variance is also possible within the multilevel framework and supported by many software implementations. It is recommended to compute descriptive statistics for the

standardised residuals across the levels of all categorical predictors and to plot standardised residuals against quantitative predictors as part of routine model-checking. If some differences in variation appear, a heteroscedastic model may be explored. Example 21.11 illustrates the procedure in our SCC example and the predictor -heifer-.

21.5.9 Sample size

A frequently asked question is: how many units are needed at each hierarchical level for multilevel analysis? A simulation study on the impact of the number of units at the highest (second) level on the parameter estimates (Maas & Hox, 2004) provided the following rule of thumb: "If one is only interested in the fixed effects, 10 groups can lead to good estimates. If one is also interested in contextual effects, 30 groups are needed. If one also wants correct estimates of the standard errors, at least 50 groups are needed." For the cluster size, Rabe-Hesketh & Skrondal (2008), Section 3.8 stated that a cluster size of 2 suffices if there are many clusters.

Calculation of the required sample size to achieve a desired accuracy or a desired power for a hypothesis test is a difficult problem for multilevel models because of the complexity involved

Example 21.11 Heterogeneous variances for somatic cell count data
data = scc40_2level

We noted in Example 21.5 that the random slopes model for -heifer- indicated larger variance for older cows than heifers. The standard deviation among the standardised cow-level residuals from the random intercept model (Example 21.9) was 0.93 for heifers and 1.035 for older cows. This motivates fitting a model that allows for different error term variances for heifers and older cows. The table below gives estimates for -heifer- and the variance parameters for several models, fitted by ML estimation (so estimates are not identical to those of previous examples). The standard errors of the heterogeneous variance parameters were computed by the delta method (Weisberg, 2005, Section 6.1.2).

Model	Random intercept		Heterogenous variance		Random slopes		Heterogeneous + random slopes	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β(heifer)	-0.737	0.055	-0.739	0.054	-0.734	0.066	-0.736	0.065
σ^2(herd)	0.139	0.040	0.133	0.039	0.191	0.058	0.185	0.057
σ^2(heifer)	-	-	-	-	0.047	0.037	0.049	0.037
σ(covar)	-	-	-	-	-0.073	0.040	-0.068	0.040
σ^2(cow)	1.554	0.048	-	-	1.543	0.048	-	-
- heifer	-	-	1.359	0.065	-	-	1.354	0.065
- older cow	-	-	1.695	0.069	-	-	1.679	0.068
2lnL	-7209.36		-7197.23		-7203.69		-7192.07	

The values of the log-likelihood show that the heterogeneous variances model are a substantial improvement in both the random intercept and the random slopes models: the cow-level variance is about 25% larger for older cows than heifers. The inference for the difference between heifers and older cows is, however, virtually unaffected.

in the effects at multiple levels. The variance inflation inherent in the design effect (Section 20.3.3) only applies to a group-level predictor. For a 2-level setting, the PinT shareware program (Snijders & Bosker, 1993) has been a standard reference in multilevel analysis for years. Recently, the simulation-based approach to power calculation (Section 2.11.8) has been extended to complex multilevel designs, including cross-classification, by William Browne and co-workers (MLPowSim, available at the Multilevel Modelling website).

REFERENCES

Berkhof J, Snijders T. Variance component testing in multilevel models Journal of Educational and Behavioral Stat. 2001; 26: 133-52.

De Vliegher S, Laevens H, Barkema HW, Dohoo IR, Stryhn H, Opsomer G, de Kruif A. Management practices and heifer characteristics associated with early lactation somatic cell count of Belgian dairy heifers J Dairy Sci. 2004; 87: 937-47.

Dean A, Voss D. Design and Analysis of Experiments. Springer; New York. 2000.

Diez-Roux AV. Multilevel analysis in public health research Annu Rev Public Health. 2000; 21: 171-92.

Dohoo IR, Tillard E, Stryhn H, Faye B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island Prev Vet Med. 2001; 50: 127-44.

Fitzmaurice G, Laird N, Ware J. Applied Longitudinal Analysis. Wiley; Hoboken, New Jersey. 2004.

Gelman A, Hill J. Data Analysis using Regression and Multilevel/Hierarchical Models. Cambridge University Press; Cambridge. 2006.

Goldstein H. Multilevel Statistical Models, 3rd Ed. Arnold; London. 2003.

Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion) J R Stat Soc A. 1996; 385-443.

Greenland S. Principles of multilevel modelling Int J Epidemiol. 2000a; 29: 158-67.

Greenland S. When should epidemiologic regressions use random coefficients? Biometrics. 2000b; 56: 915-21.

Gustafson P, Greenland S. The performance of random coefficient regression in accounting for residual confounding Biometrics. 2006; 62: 760-8.

Hox J. Multilevel Analysis: Techniques and Applications. Lawrence Erlbaum; Mahwah, NJ. 2002.

Langford I, Lewis T. Outliers in multilevel models (with Discussion) J R Stat Soc A. 1998; 161: 121-60.

Littell R, Milliken G, Stroup W, Wolfinger R, Schabenberger O. SAS for Mixed Models, 2nd ED. SAS Publishing; Cary, NC. 2006.

Longford N. Standard errors in multilevel analysis. Multilevel Newsletter 1999; 11:10-13.

Maas C, Hox J. Robustness issues in multilevel regression analysis Statistica Neerlandica. 2004; 58: 127-37.

McCullagh C, Searle S, Neuhaus J. Generalized, Linear, and Mixed Models, 2nd Ed. Wiley; New York. 2008.

Pinheiro J, Bates D. Mixed-effects Models in S and S-Plus. Springer; New York. 2000.

Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling using Stata, 2nd ED. Stata Press; College Stn. Tx. 2008.

Raudenbush S, Bryk A. Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ED. Sage; Thousand Oaks CA. 2002.

Schaalje GB, McBride JB, Fellingham GW. Adequacy of approximations to distributions of test statistics in complex mixed linear models J Agr, Biol, and Env Stat. 2002; 7: 512-24.

Skrondal A, Rabe-Hesketh S. Prediction in multilevel generalized linear models J R Stat Soc A. 2009; 172: 659-87.

Snijders T, Bosker R. Standard errors and sample size for two-level research Journal of Educational Statistics. 1993; 18: 237-59.

Snijders T, Bosker R. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling. Sage Publications; London. 1999.

Stryhn H, De Vliegher S, Barkema H. Contextual multilevel models: effects and correlations at multiple levels. In: Proceedings of the XIth International Symposium on Veterinary Epidemiology and Economics. 2006.

Verbeke G, Lesaffre E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data Computational Statistics and Data Analysis. 1997; 23: 541-56.

Weisberg S. Applied Linear Regression, 3rd ED. Wiley; New York. 2005.