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 2 types of parameters or effects:

- **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 gained prominence with 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, 2000a), and we elaborate on this issue 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, 25, and 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 blood pressure dataset (bp—described in more detail in Chapter 31) is used to illustrate the methods numerically. It contains data from a multicentre trial to compare treatments for hypertension (Hall *et al*, 1991). Within each centre, patients were randomly allocated to one of 3 treatments, of which one (Carvedilol) was a new drug, whereas the other 2 (Nifedipine and Atenolol) were standard drugs. Pre-treatment blood pressure measurements were taken at the first visit to the centre, treatments were administered at the second visit and follow-up measurements were obtained at visits 3–6 (at weeks 3, 5, 7, and 9), and thus constitute **repeated measures** per patient (individual). The data structure may be considered as 3-level hierarchical: 1,092 records within 288 patients within 29 centres. Our outcome of interest is the diastolic blood pressure (-dbp-). In this chapter, we include only a single measurement per individual—obtained at the first post-treatment visit (week 3). The data have been analysed extensively in Brown and Prescott (2006). For a 2-level structure, these authors considered week 9 data for each patient, with the last measurement carried forward for patients who dropped out of the study (one simple method of dealing with missing data, see Section 15.5). Records were available at week 3 for all patients but one, so patient dropout is not a serious concern here. Obviously, any inferences about treatment effects should not be based on analysis of a subdataset such as the present one (the full dataset is considered in Chapter 23). Table 21.1 lists the variables of the bp data used in the examples.

21.2 LINEAR MIXED MODEL

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

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

In our example, we will take as our outcome Y—the blood pressure—and as our regressors $X_1,\ldots,X_k$—the continuous and dummy variables necessary to represent the chosen predictors. Further, the errors $\varepsilon_1,\ldots,\varepsilon_n$ are assumed independent and $\sim N(0,\sigma^2)$. This equation (and its assumptions) would be meaningful if we considered one visit per individual and there was no clustering in centres (*eg* we might have data from only one centre).

Table 21.1 Selected variables from the dataset bp

Variable	Level of measurement	Description
centre	3:centre	centre identification
patient	2:patient	patient identification
visit	1:visit	visit number: 3,4,5,6
tx	2:patient	drug: 1 (Carvedilol), 2 (Nifedipine), 3 (Atenolol)
dbp	1:visit	diastolic blood pressure (mm Hg)
dbp1	2:patient	diastolic blood pressure before treatment (at visit 1) (mm Hg)

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 (29) centres, and we would like the centres to enter our model as well, because we know that there might be some variation of blood pressures across centres. Previously, we have discussed including groups (here, centres) in the model by adding a set of (29-1=28) indicator variables and estimating a separate β for each of them. A **mixed model** with a **random group effect** is written:

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

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, $u_{\text{group}(i)}$ refers to the group containing the i^{th} individual (*eg* u_7 for individuals in group (centre) 7). As there are 29 groups, u would have one of 29 values: u_j, j=1,...,29. An alternative notation uses multiple indices such as $u_j + \varepsilon_{ij}$ where j refers to the group and i to the i^{th} individual in the j^{th} group.

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 $u_{\text{group}(i)}$, a random group effect for the group of the i^{th} individual. 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 group as a random term for now, and first look at the assumptions for u and ε:

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

where all u_j and ε_i are independent.

Thus, we assume the impact of each group to be a random fluctuation with mean zero (and consequently centred at the mean determined by the fixed effects) and standard deviation σ_g. Therefore, the parameter σ_g^2 can be interpreted as the random variation in blood pressures between groups (centres). Furthermore, we could calculate:

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

In effect, we have decomposed the total variance to a sum of the variance between groups and the error variance (or the variance within groups). 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 groups can be motivated in different ways. Strictly speaking, it corresponds to effects (groups) 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 groups are generally representative of the population even if they were not randomly selected. In our example, the 29 centres were presumably not randomly selected from a larger set of centres, and some of the centres originally intended for inclusion in the study did not recruit patients after all and were hence omitted. Consequently, the population these centres could be representative of is probably not well-defined. With random effects, the focus shifts from the individual group to the variability between groups in the population σ_g^2 . In a study with only a few groups of particular interest (possibly because they were individually selected for the study), one might prefer to model groups 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 individuals in groups in regions would lead to random effects for both groups and regions, and we then split the variation into 3 terms: $\text{var}(Y_i) = \sigma_r^2 + \sigma_g^2 + \sigma^2$. 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 may 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* centre or region) is not a random sample of a larger population and does not have a large

Example 21.1 Variance components and random effects
data = bp

When restricting analysis to records from visits at week 3, this dataset contains one observation from each of 287 individuals visiting one of 29 centres. In a 2-level random effects model for -dbp- with no fixed effects (a 'null' or 'empty' model'), the variance components were estimated at:

$$\sigma_g^2 = 3.43 \qquad \text{and} \qquad \sigma^2 = 81.40$$

Thus, the total (unexplained) variance was 3.43+81.40=84.83. It is often useful to compute the fractions at the different levels; here we have 3.43/84.83=4.0% of the variance between groups (centres) and 96.0% within groups. We can also give a direct interpretation of σ_g^2: 95% of the group effects should be within an interval of $\pm 1.96\ \sigma_g = \pm 3.63$. As the overall mean ($\beta_0$) was 93.41, this means that most group mean -dbp- values in the population lie between 89.8 and 97.0.

Example 21.2 Mixed model estimates for 2-level blood pressure data
data = bp

A linear mixed model with tx and initial blood pressure (-dbp1-) was fit to the 29-centre, 2-level diastolic blood pressure data at visit 3. The initial blood pressure was centred by subtracting the median (102). The reference level for -tx- was the new drug, Carvedilol.

	Coef	SE	Z	P	95% CI	
dbp1 centred	0.558	0.107	5.21	<0.001	0.348	0.768
tx = Nifedipine	-1.35	1.24	-1.08	0.279	-3.78	1.09
tx = Atenolol	-3.42	1.24	-2.77	0.006	-5.85	-1.00
constant	94.39	0.937	-	-	92.56	96.23

Note that, because of the random group (centre) effects, the constant refers to the blood pressure in an average group, not to the value of an average individual across the population of groups. As groups differ in size, these means are not necessarily the same. For example, if the lowest blood pressures were obtained at the largest centres (some indication of this exists in the data), then the patient average would typically be lower than the centre average. The individual and group 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:

$$\sigma_g^2 = 2.22\,(2.65) \quad \text{and} \quad \sigma^2 = 73.86\,(6.45)$$

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. Here, by comparison with Example 21.1, both variance components are reduced, and the proportion of variance at the group level has dropped to 3%, despite the fact that none of the added predictors reside at the group level. Some group-level differences in initial blood pressure are present, though (we discuss this further in Example 21.6). In general, it is not unusual that adding fixed effects to hierarchical models redistributes the variation across the levels and may increase 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 and Bosker (1999) for details and ways of defining measures of the variance explained by fixed effects.

number of elements. Some 'final' remarks on fixed vs random effects have been collected in Section 21.5.7.

21.2.1 Intra-class correlation coefficient

The model assumptions allow us to examine the dependence or correlation between observations from the same group. 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, centre) is described by the intra-class 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:

$$\rho = \frac{\sigma_g^2}{\sigma_g^2 + \sigma^2} = \frac{3.43}{3.43 + 81.40} = 0.040 \qquad \textit{Eq 21.4}$$

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 (*ie* groups were clustered within regions), the correlation between individuals in the same group (and hence within a region) would be:

$$\rho(\text{individuals in same group}) = \frac{\sigma_r^2 + \sigma_g^2}{\sigma_r^2 + \sigma_g^2 + \sigma^2} \qquad \textit{Eq 21.5}$$

Similarly, the correlation between individuals in different groups in the same region would be:

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

Example 21.3 shows similar computations for a 4-level hierarchical structure in hospital data. The correlation in Eq 21.6 referred to individuals in different groups but an intuitively more appealing value might be the correlation **between groups**—more precisely, **between group means**. The correlation between means of 2 groups of size m is

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

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

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

Example 21.3 Intra-class correlations in a 4-level mixed model

Krogstad *et al* (2005) used 4-level mixed models to investigate how the variation of nurse evaluations of aspects of hospital work was distributed across individuals (2,606), wards (124), departments (36), and hospitals (15). Answers were combined across multiple questions using exploratory factor analysis into indices, scaled into 0–100 'satisfaction scores', at a total of 9 domains. The multilevel models did not include fixed effects, so the model for a score Y_i computed from answers by nurse i could be written:

$$Y_i = \beta_0 + u_{\text{ward}\,(i)} + v_{\text{department}\,(i)} + w_{\text{hospital}\,(i)} + \varepsilon_i$$

The variance components reported for the domain 'workload' were:

$$\text{hospital}: \sigma_h^2 = 0, \quad \text{department}: \sigma_d^2 = 18.6, \quad \text{ward}: \sigma_w^2 = 84.6, \quad \text{individual}: \sigma^2 = 351.5$$

The authors reported variances as zero (*eg* the between-hospital variance above) when the data showed no statistical evidence of variation at the corresponding level. For each domain, proportions of variance at the different levels as well as the ICC for observations within wards were computed and used to discuss the implications of similarities and differences across the hierarchical levels in the responses for different domains. In risk factor studies, the amount of unexplained variation at a certain hierarchical level might indicate the potential for improvement by interventions at that level.

From the listed estimates, we could compute a total variance of 454.7 and the following correlations between observations (scores by individuals):

individuals within the same ward:	ρ=(0+18.6+84.6)/454.7=0.227
individuals in different wards of the same department:	ρ=(0+18.6)/454.7=0.041
individuals in different departments of the same hospital	ρ=0/454.7=0

The last calculation was only included for demonstration purposes, but some domains showed appreciable correlation within hospitals, most notably the physical layout of workplace (ρ=0.176).

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* doctors and nurses in hospital data. We denote the 2 groupings by A and B. If both doctors and nurses 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 and Voss, 2000) (and more commonly written in a 2-index notation with i and j representing the units (or categories) of 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:

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

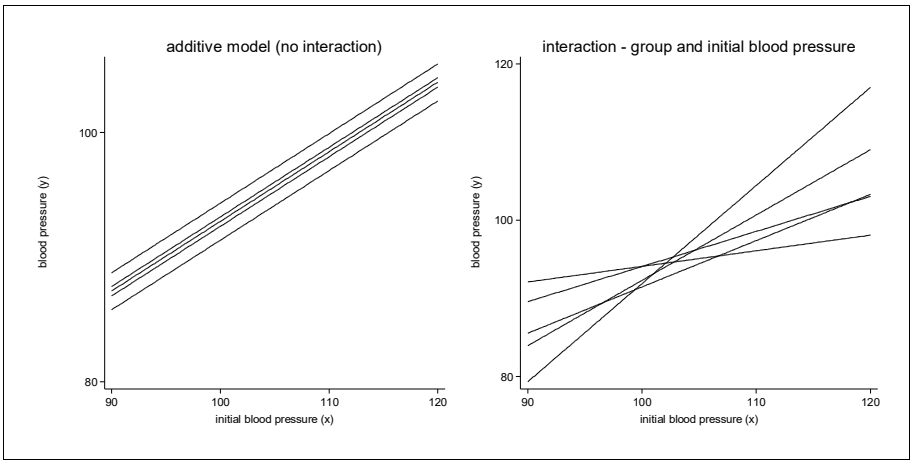

Fig. 21.1 Schematic graphs of additive and non-additive modelling of a continuous predictor for a continuous outcome

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, such as the initial blood pressure in Example 21.2. 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 (treatment). Then the predicted blood pressures from the model for different treatments 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 individuals in a specific treatment group (assuming here for the sake of illustration that more than 3 treatment groups exist). If an interaction term between treatment and initial blood pressure was added, this would produce non-parallel lines (for different treatment groups), as outlined on the right.

Exactly the same interpretation is valid for individuals in different groups: in an additive model (Eq 21.2) the regression lines corresponding to different centres are parallel, and the random group effects can be read as the vertical distances between the lines. This is because Eq 21.2 assumes the impact on the blood pressure of a difference in initial blood pressure (*eg* 1-unit increase) to be the same for individuals in all groups (parallel lines). Fig. 21.1 was indeed drawn based on estimates for 5 centres in the models of Examples 21.2 and 21.4.

21.3.2 Random slopes as non-additive group effects

An assumption of additive group effects (parallel lines) might not be biologically obvious because other factors related to the groups, such as group management factors or heterogeneity in individuals between groups (inherent in the group effects), could influence the relationship. Adding an interaction between groups and X_1 means that slopes vary among groups. If group was included in the model as a set of fixed effects, the interaction term would result in a

specific effect being estimated for each group. With group as a random effect, the slopes are assumed to vary according to some distribution (in addition to the intercepts varying between groups). A model with random slopes for a single fixed effect (X_1) is written as:

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

where in addition to the previous assumptions, we assume for the random slopes that the $b_{\text{group}} \sim N(0,\sigma_1^2)$. The parameter σ_1^2 is interpreted as the variation in slopes (for predictor X_1) among groups. The regression parameter β_1 is now the overall or average slope for X_1, which is then subject to random fluctuations between groups. As a rough rule, with probability 95%, the slope in a given group 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 groups are modelled as random, any slopes varying between groups should also be random. (**Note** The random group effect, u_{group}, and its variance, σ_g^2, now represent the variation between groups 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_{group}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 group level (u_{group} and b_{group}) correspond to the intercept and slope for the regression on X_1 at the group 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 group intercept and slope. It is useful to display the 3 parameters:σ_g^2, σ_1^2 and the covariance σ_{g1}, in a 2X2 matrix as follows:

$$\begin{pmatrix} \sigma_g^2 & \sigma_{gl} \\ \sigma_{gl} & \sigma_1^2 \end{pmatrix}$$

and the correlation between the group intercepts and slopes is computed as $\sigma_{gl}/(\sigma_g\sigma_1)$. Example 21.4 shows the effect of adding a random slope to the model for the blood pressure 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:

$$\begin{aligned} \text{var}(Y_i) &= \text{var}(u_{\text{group}(i)}) + \text{var}(b_{\text{group}(i)} X_{1i}) + 2\,\text{cov}(u_{\text{group}(i)}, b_{\text{group}(i)} X_{1i}) + \text{var}(\varepsilon_i) \\ &= \sigma_g^2 + X_{1i}^2 \sigma_1^2 + 2 X_{1i} \sigma_{g1} + \sigma^2 \end{aligned} \qquad \textit{Eq 21.11}$$

Example 21.4 Random slopes of -dbp1c- for blood pressure data
data = bp

Adding a random slope of -dbp1c- (the centred -dbp1-) to the model of Example 21.2 gave almost the same regression coefficient (0.537) but with a substantially increased SE (0.156), and the random effect parameters (with SEs) were:

$$\begin{pmatrix} \sigma_g^2 & \sigma_{g1} \\ \sigma_{g1} & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} 1.773\,(2.429) & -0.138\,(0.460) \\ -0.138\,(0.460) & 0.285\,(0.171) \end{pmatrix} \quad \text{and} \quad \sigma^2 = 67.73\,(6.13)$$

The value of σ_1^2 suggests that 95% of the slopes for -dbp1c- lie roughly within 0.54±1.05=-0.51,1.58. The correlation between intercepts and slopes is fairly small $\left(-0.138/\sqrt{1.773*0.285} = -0.19\right)$ so the centring of -dbp1c- largely removed the correlation. The value of σ_1^2 is less than twice as large as its SE and σ_{g1} 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 significant). Note finally that a model with random slopes for any predictors at the group level (we don't have such predictors in the bp data, but we could imagine centre descriptors or demographics related to centre location) 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_{g1}, 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 bp data. The dependence of the total variance on X_1 is weak, because the major portion of the variance is at the individual level; nevertheless, it may be biologically plausible that more extreme values of initial blood pressure are associated with larger variability.

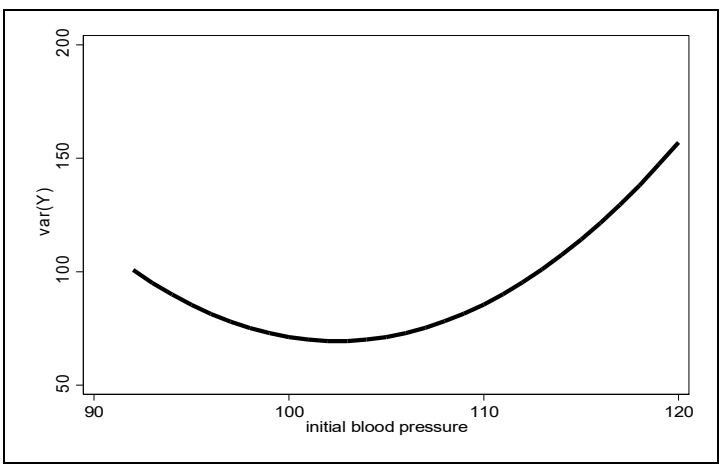

Fig. 21.2 Variance function for model which included a random slope for initial blood pressure

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 groups, and one might want to allow it to vary between groups. 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 effects of adding random slopes for the treatment in the bp data. We also discuss the alternative approach of modelling the treatment by group combinations as an additional hierarchical level in the model, nested within groups. This leads to independent random terms ('deviations') for each treatment in every group, and the variance associated with the hierarchical level measures the magnitude of these deviations from the average treatment effects.

Example 21.5 Random slopes of -tx- for blood pressure data
data = bp

Adding a random slope (of -tx-, represented by 2 indicator variables relative to Carvedilol) to the model from Example 21.2 produces treatment effects of -1.23(1.52) and -3.11(1.35) (for Nifedipine and Atenolol, respectively), and the variance parameters (with SEs):

$$\begin{pmatrix} \sigma_g^2 & \sigma_{g1} & \sigma_{g2} \\ \sigma_{g1} & \sigma_1^2 & \sigma_{12} \\ \sigma_{g2} & \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 3.96(6.64) & -4.76(8.48) & -3.18(7.75) \\ -4.76(8.48) & 14.16(14.73) & 6.99(11.00) \\ -3.18(7.75) & 6.99(11.00) & 6.58(12.41) \end{pmatrix} \quad \text{and} \quad \sigma^2 = 70.79(6.63)$$

The interpretation of the random slope for the treatment contrast between Nifedipine and Carvedilol is that the former drug is associated with a reduction in mean blood pressure of 1.23 units, but that this reduction varies across centres with a standard deviation of √(14.16)=3.76, and hence is very variable across centres. The 3 variance contributions at the group (centre) level of this model are 3.96 for Carvedilol, 3.96+14.16+2*(-4.76)=8.60 for Nifedipine, and 3.96+6.58+2*(-3.18)=4.18 for Atenolol. 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 larger variance for the Nifedipine drug. As we will see later, and one might expect from the large SEs for the variance parameters, the random slopes in this case do not offer any significant model improvement.

Instead of modelling -tx- effects as random slopes, we can incorporate differences in -tx- effects across groups by introducing an intermediate hierarchical level at treatment by group combinations (effectively, a random interaction between the treatment factor and groups). This model extension only involves one additional parameter, the treatment by group variance σ^2_{g*tx}, which can be interpreted as the amount of variation in -tx- effects across groups beyond the overall fixed -tx- effect. Extending our model from Example 21.2 in this way yields the estimates (with SEs):

$$\sigma_g^2 = 0.96(3.05) \quad \sigma^2_{g*tx} = 4.60(4.83) \quad \sigma^2 = 70.77(6.67)$$

The estimates indicate that the majority of the unexplained group level variation is related to differences in -tx- effects across groups; again, the improvement over the 2-level model in reality is modest (and non-significant).

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 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 (Raudenbusch and Bryk, 2002). We will outline the rationale behind the modelling approach by slightly rewriting the random slopes model of Eq 21.10 as:

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

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_{group} is often termed a random intercept (at the group 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 slope) models in epidemiology (Greenland, 2000a) 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 and 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 and Bosker, 1999). We describe first a contextual effect of a predictor in a random intercept model (Eq 21.2), and then consider the extension to a random slopes model. The predictor X_1 is said to have a contextual effect if the following conditions are both satisfied:

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

Two situations where condition i. is **not** satisfied are: when X_1 is a group-level predictor, and when the group averages ($\overline{X}_{1\,\text{group}}$) are constant between groups (*eg* in the blood pressure clinical trial with treatment groups almost equally represented within each centre). For condition ii., the within-group regression of Y on X_1 refers to a regression equation

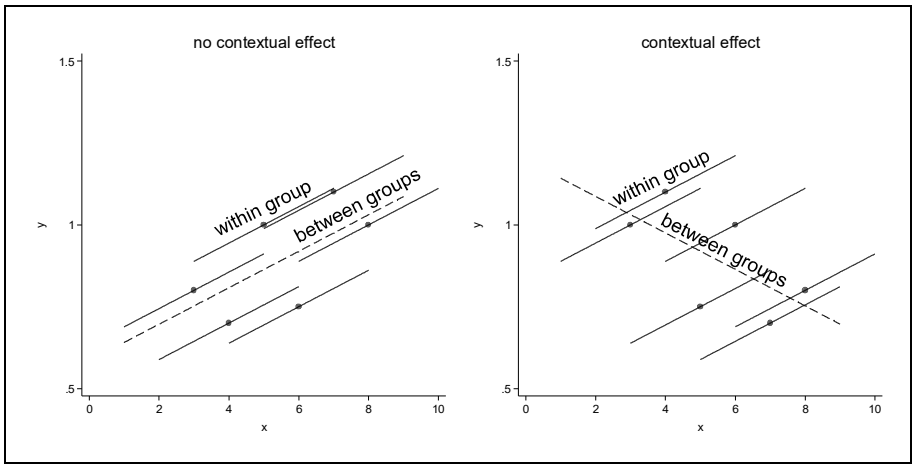

Fig. 21.3 Schematic graphs showing no contextual effect (left) and a strong contextual effect (right) of the predictor X_1

corresponding to different individuals within a single group. Furthermore, the between-group regression is a regression of group mean outcomes $(\overline{Y}_{\text{group}})$ on group predictor means $(\overline{X}_{1\,\text{group}})$. Fig. 21.3 illustrates situations where the between-group and within-group regressions of Y on a continuous predictor X_1 coincide (left-hand panel) and are completely different (right-hand panel). The within-group regressions are indicated by solid lines (without showing individual data points), and the between-group regression is obtained by fitting a straight line to the dotted points (group means of Y and X_1) and is represented by the dashed line. In the right-hand panel, the within-group slope is positive whereas the between-group regression would have a negative slope.

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

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

where $(\overline{X}_{1\,\text{group}(i)})$ is the X_1 mean for the group 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-group centred version, $Z_{1i} = X_{1i} - \overline{X}_{1\,\text{group}(i)}$, as follows:

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

Eqs 21.12 and 21.13 represent the **same** model, and the coefficients for X_1 and Z_1 are identical (β_1), whereas $\tilde{\beta}_2 = \beta_1 + \beta_2$. The parameter $\tilde{\beta}_2$ is the slope of the between-group regression of Y on X_1 (*ie* between the corresponding group means, as explained above; dashed lines in Fig. 21.2) and the parameter β_1 in models 21.12 or 21.13 is the slope of the within-group regression of Y on X_1 (solid lines in Fig. 21.2). Example 21.6 shows how these models can be fit to the blood pressure data.

As demonstrated in the example, contextual effects may also be incorporated into random slopes models, by adding the group-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 parameterisations

Example 21.6 Contextual effects for blood pressure data
data = bp

We illustrate the modelling of contextual effects of the initial blood pressure (-dbp1c-) in the blood pressure data although they fail to provide a convincing demonstration of the utility of such effects. Even though our previous analyses demonstrated significant effects of -dbp1c-, both overall and as a random slope, and some differences in the -dbp1c- values between centres do exist, it is hardly obvious why such differences would have predictive power for between-centre differences. In addition, the group sizes are very variable, ranging from 1 to 39, leaving group means with variable precision and hence a between-group regression difficult to establish. Stryhn *et al* (2006) presented results for contextual effects of 2 predictors at the herd level in veterinary observations on cows in herds.

Model	Random intercept		Contextual		Random slopes		Contextual + random slopes	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
β(dbp1c)	0.558	0.107	0.520	0.115	0.537	0.156	0.500	0.165
β(gdbp1c)	-	-	0.297	0.320	-	-	0.229	0.339
σ^2(centre)	2.218	2.652	2.146	2.628	1.773	2.429	1.747	2.446
σ^2(dbp1c)	-	-	-	-	0.285	0.171	0.282	0.170
σ^2(patient)	73.86	6.45	73.94	6.46	67.73	6.13	67.94	6.15

In the table, the estimates for the intercept and -tx- effects have been omitted, and the variable -gdbp1c- contains the group means of -dbp1c-. The lack of significance of the contextual effects in both the random intercept and random slopes models is indicated by the estimates of -gdbp1c- being smaller than their SEs. In the random intercept model with contextual effects, the within- and between-group regression slopes are estimated at 0.520 and 0.817 (computed as 0.520+0.297), respectively, but the SE of the between-group slope is large (0.298; computed by fitting Eq 21.13). The added contextual effect did not substantially change the within-group regression coefficient. In models with a significant contextual effect, however, the within-group slope may change dramatically and even become non-significant, in which case the apparent individual-level regression in the random intercept model changes in interpretation to a group level regression. It is also possible to have a significant between-group regression without a strong impact on the within-group regression; see Stryhn *et al* (2006) for examples of such scenarios.

above lead to different models, depending on whether X_1 is included as X_1 or Z_1 (for both fixed effects and random slopes). 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 and 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 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 early 2012, the main software options (with corresponding texts providing theory, examples and code) were (in unstructured order): Stata (Rabe-Hesketh and Skrondal, 2012), S-Plus/R (Gelman and Hill, 2006; Pinheiro and 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 and 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 smaller 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 and Voss (2000), Chapter 17). It is simpler to implement and is offered by more software packages. By and large, this approach is obsolete by today's standard,s 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* correct standard errors are not readily available.

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

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

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 bp data analysed in previous examples, the above formula yields ρ=0.034; Eq 21.4 gives a value of 0.040.

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 is 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

Example 21.7 Inference for fixed effects for blood pressure data
data = bp

A multiple Wald test for the combined effect of -tx- gives $\chi^2(2)=7.75$ and a P-value of 0.02; thus, there are moderately significant differences between treatments (in this subdataset). Analysis by SAS Proc Mixed or R (nlme library) yields finite sample reference *t*-distributions with about 260 degrees of freedom for the coefficients for -tx- and -dbp1c- which corresponds roughly to the residual degrees of freedom at the individual level. With such large degrees of freedom there is no difference between *t* and *z* distribution inference. Only one pairwise comparison between drugs is not already included in the listing of Example 21.2: the contrast between Nifedipine and Atenolol is estimated at 2.08 (1.26) with a P-value of 0.10. Thus only the difference between Carvedilol and Atenolol attains statistical significance.

The finite sample reference distribution for the group means -gdbp1c- in Example 21.6 is $t(63)$, reflecting that it is a group-level predictor but that most of the unexplained variance resides at the individual level. This is because the Satterthwaite approximation involves a weighted degrees of freedom from the 2 estimated variance components. In this case, change to a *t* reference distribution has only minimal impact on the inference and does not affect our conclusions. In situations with a fairly small number of groups and a large proportion of unexplained variance at the group level, the difference between *z* and *t* distribution inferences can be appreciable.

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 and 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 blood pressure data.

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: $\sigma_g=0$ is calculated as $G^2=-2(\ln L_{\text{full}}-\ln L_{\text{red}})$ where the full and reduced models refer to the models with and without the group 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 and 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: $\sigma_g>0$). 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 and 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 blood pressure 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

Example 21.8 Inference for random effects for blood pressure data
data = bp

The table below gives values for twice the log likelihood function (based on REML) for various models in this chapter and likelihood-ratio test statistics for model comparisons (comparing all models with the random intercept model presented in Example 21.2). 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 group random effect	-2050.73	2060.73	1.08	1	0.150
random intercept (Ex 21.2)	-2049.65	2061.65	-	-	-
random slope of -dbp1c- (Ex. 21.4)	-2040.41	2056.41	9.24	2	0.005
random slope of -tx- (Ex. 21.5)	-2047.74	2069.74	1.91	5	0.431
random -tx- by group interaction (Ex. 21.5)	-2048.35	2062.35	1.30	1	0.127

The table shows no formal statistical evidence against the hypothesis of no (random) variation between groups (centres), but as the group variance component was non-zero there is no pressing need to adopt a model without group effects. The table also shows a strong random slope for -dbp1c-, no indication of random slopes for -tx- effects, but some gain from introducing a treatment by group level interaction. The different conclusions about the 2 versions of -tx- related random effects may be due to the more parsimonious modelling, involving only one additional variance parameter, by the group by treatment interaction. Based on these results, one might want to explore a model with both a random slope of -dbp1c- and the random -tx- by group interaction, but we stop here.

The 95% confidence interval for σ_g^2 for the model of Example 21.2 was (0.213,23.1). It is asymmetric around the estimate (2.218) because variance is estimated on a 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 0.5) belongs to the confidence interval, estimate the model with σ_g^2 fixed at 0.5, obtain this model's 2lnL value (-2050.23) and check whether it is within 3.84 of the full model's value (-2049.65). In this case it is, therefore the value 0.5 belongs to the 95% confidence interval. Repeat the process with other values than 0.5 to determine the interval. The profile-likelihood CI obtained by this method was (0,10.6), which is remarkably different from the above interval. The strict lower bound of any CI for a variance parameter is 0, and this value should be the lower limit of a likelihood-based 95% CI whenever the variance parameter is non-significant (based on a LRT).

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 on 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 blood pressure data.

21.5.4 Prediction

Even though the random effects in a mixed model are not parameters in the usual sense, it is 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, or **empirical Bayes estimates** (Greenland, 2000b), 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 (centres, 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 and 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 and 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 $\sigma_g^2/(\sigma_g^2+\sigma^2/m)$, where m is the group size. It is seen that this formula qualitatively has the 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 blood pressure dataset has only 29 group-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 group 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 and Lewis, 1998) recommended to first inspect the residuals at the highest hierarchical level, and then gradually work downwards. Thus, before looking at individual patients being influential, or not fitted well by the model, we examine the same questions for the groups (centres). This is because several of the individuals being flagged could stem from the same group, so the 'problem' might be with the group rather than with the individual. Example 21.9 presents group-level residuals and diagnostics for the blood pressure data.

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 in 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),

Example 21.9 Residuals and diagnostics for blood pressure data
data = bp

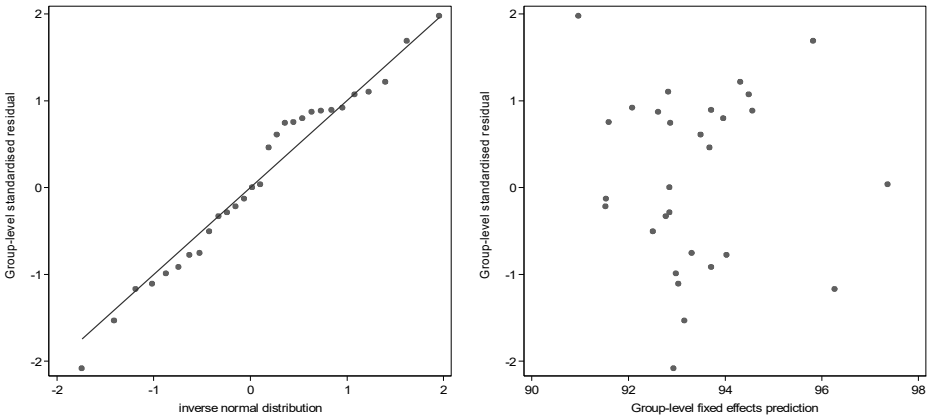

Fig. 21.4 Quantile plot (left) and residual plot (right) for group-level residuals

We present here group-level residual plots and a listing of the residuals and diagnostics for the 7 most 'interesting' groups (centres): with the 2 most extreme negative and positive residuals, and with the 3 largest values of Cook's distance. The analysis of individual-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). In the table, group size is the number of individuals included per group.

centre number	group size	raw residual	standardised residual	Cook's distance	leverage	DFITS
36	22	-1.892	-2.085	1.42	0.392	0.343
2	10	-1.077	-1.534	0.27	0.529	0.282
7	18	-0.854	-0.992	6.93	0.426	0.160
31	36	-0.784	-0.755	3.33	0.312	0.104
1	39	-0.020	-0.003	1.63	0.302	0.002
3	8	1.075	1.684	1.13	0.571	0.325
24	1	0.497	1.971	0.13	0.830	0.470

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. The residuals and diagnostics for individual groups point to different groups. The highest values of leverage were attained by very small groups; centre 24 with only a single patient also had the largest positive standardised residual and the largest value of DFITS, but analysis without this group led, as expected from the small group size, to only minor changes in the parameter estimates (details not shown). The 2 largest groups, with 36 and 39 patients, had relatively small residuals (and hence also small values of DFITS), but large values of Cook's distance, and in particular centre 1 had an appreciable impact on parameter estimates (one -tx- effect dropped by about 40%). The largest impact on parameter estimates was from centre 7, the omission of which would be associated with a decline in the slope for -dbp1c- by 20% and the contrast between Carvedilol and Nifedipine dropping down to almost zero. Such large impacts of single groups are noteworthy and relevant for the discussions of the internal and external validity of a study.

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 blood pressure 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

Example 21.10 Box-Cox analysis for blood pressure data
data = bp

The data contain n=287 observations and the mean (natural) logarithmic blood pressure (-dbp-) 4.5285. The following table and graph (which includes additional λ-values) give a Box-Cox analysis:

λ	2	1	0.5	0	-0.5	-1
ln(L) for $Y^{(\lambda)}$	-2329.30	-1026.06	-376.90	270.61	916.45	1560.64
pl(λ) from Eq 21.15	-1029.62	-1026.06	-1026.74	-1029.08	-1033.07	-1038.72

The table and figure indicate the optimal value of λ to be very close to 1, indicating that no power transformation will improve the compliance with the distributional assumptions. The 95% CI for λ is wide and would include both 0 and 2. Our discussion in Example 21.9 indicated the group-level residuals to be acceptable, and the individual-level residuals also looked reasonably good (not shown), so there does indeed not seem to be any need for a transformation.

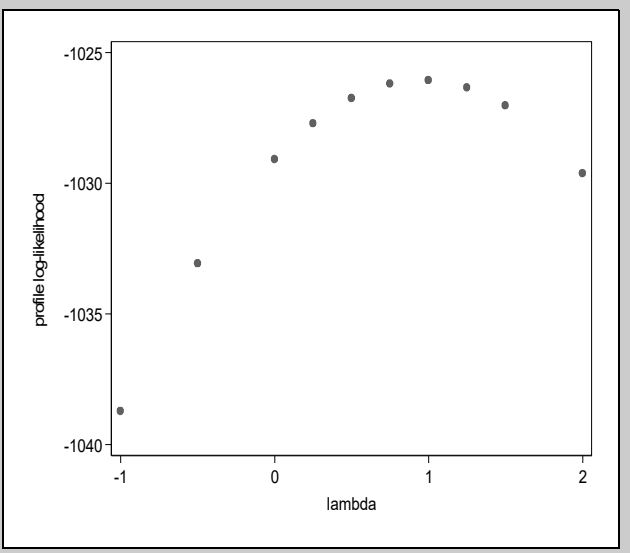

Fig. 21.5 Profile-likelihood function for Box-Cox analysis of blood pressure data

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'. The Hausmann 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 unnecessary 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 and Skrondal (2012), Section 3.7). 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 $\chi^2(3)=0.63$, which is absolutely non-significant in a χ^2 distribution with 3 df. We already established in Example 21.6 that -dbp1c- does not have a contextual effect, and nor does -tx- because of its randomisation within groups.

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 groups, particularly when:

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

A more technical comparison of fixed and random effects modelling can be found in Rabe-Hesketh and Skrondal (2012), 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 and 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 and 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 to strong changes in SEs (at most 15% changes in SE for fixed effects, except for a 54% increase for -dbp1c-; a doubling of the SE of the error variance but no impact on the group-level variance; results not shown). The strong increase for -dbp1c- may indicate model misspecification (previous examples demonstrated clear improvements in fit by adding random slopes for -dbp1c-). Generally speaking, 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 blood pressure example with the -tx- and -dbp1c- predictors.

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 and Hox, 2004) provided the following guide: “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 and Skrondal (2012), 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 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 and Bosker, 1999) 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).

Example 21.11 Heterogeneous variances for blood pressure data
data = bp

The random slopes model explored in Examples 21.5 and 21.6 implied that the group-level variance was non-constant and depended on predictor values. In this example we explore whether allowing the individual-level variance to depend on predictor values improves model fit. As implemented in the gllamm library for Stata, the dependence is on log standard deviation scale, and we (separately) fit a categorical effect for -tx- and a quadratic regression for -dbp1c- (similar to Eq 21.11). The table below gives estimates for the variance parameters for several models, fitted by ML estimation. The standard errors of the heterogeneous variance parameters for -tx- were computed by the delta method (Weisberg, 2005, Section 6.1.2).

Model	Random intercept model		Heterogeneous variance for -tx-		Heterogeneous variance for -dbp1c-	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
σ^2(group)	1.82	2.46	1.88	2.48	2.02	2.51
σ^2(individual)	73.09	6.36				
- Carvedilol			74.27	10.83		
- Nifedipine			68.88	10.60		
- Atenolol			75.86	11.35		
log σ(intercept)					2.0742	0.0517
log σ(slope -dbp1c-)					-0.0132	0.0129
log σ(slope -dbp1c_sq-)					0.0030	0.0014
2lnL	-2052.06		-2051.89		-2046.57	

The differences in variance between treatments are minimal and far from statistical significance. The quadratic regression for the log standard deviation on -dbp1c- offers a significant improvement of model fit. The estimate for the intercept can be converted into an individual variance at dbp1=102 (because -dbp1c- was centred at this value); the resulting variance estimate is 63.32 (6.55), and the estimated variance function has a similar qualitative form as in Fig. 21.2 with increasing variance at both extremes of -dbp1-. However, the random slope model of Example 21.4 (with a 2lnL-value of -2043.43) offers a better fit and thus would seem preferable. Adding heterogeneous variances to the random slopes model does not further improve the fit (2lnL=-2041.77).

REFERENCES

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

Brown H, Prescott RI. Applied Mixed Models in Medicine 2nd Ed. Chichester: Wiley; 2006.

Dean AM, Voss D. Design and Analysis of Experiments: Springer; 2000.

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

Fitzmaurice GM, Laird NM, JHWare. Applied Longitudinal Analysis. New York: Wiley; 2004.

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

Goldstein H. Multilevel Statistical Models, 3rd Ed. London: Arnold; 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. When should epidemiologic regressions use random coefficients? Biometrics. 2000a Sep;56(3):915-21.

Greenland S. Principles of multilevel modelling. Int J Epidemiol. 2000b Feb;29(1):158-67.

Gustafson P, Greenland S. The performance of random coefficient regression in accounting for residual confounding. Biometrics. 2006 Sep;62(3):760-8.

Hall S, Prescott RI, Hallman RJ, Dixon S, Harvey RE, Ball SG. A comparative study of Carvedilol, slow-release Nifedipine, and Atenolol in the management of essential hypertension. J Cardiovasc Pharmacol. 1991;18 Suppl 4:S35-8.

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

Krogstad U, Hofoss D, Veenstra M, Gulbrandsen P, Hjortdahl P. Hospital quality improvement in context: a multilevel analysis of staff job evaluations. Qual Saf Health Care. 2005 Dec;14(6):438-42.

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

Littell RC, Milliken GA, Stroup WS, Wolfinger RD, Schabenberger O. SAS for Mixed Models, 2nd Ed: SAS Publishing; 2006.

Longford N. Standard errors in multilevel analysis. Multilevel Newsletter. 1999;11:10-3.

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

McCullagh CE, Searle SR, Neuhaus JM. Generalized, Linear, and Mixed Models, 2nd Ed: Wiley; 2008.

Pinheiro JC, Bates DM. Mixed-effects Models in S and S-Plus: Springer; 2000.

Rabe-Hesketh S, Skrondal A. Multilevel and Longitudinal Modeling using Stata, 3rd Ed: Stata Press; 2012.

Raudenbush SW, Bryk AS. Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Ed: Sage; 2002.

Schaalje GB, McBride JB, Fellingham GW. Adequacy of approximations to distributions of test statistics in complex mixed linear models. J Agric Biol and Environ Stat. 2002;7(4):512-24.

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

Snijders TAB, Bosker RJ. Multilevel Analysis: An Introduction to Basic and Advanced

Multilevel Modelling. London: Sage Publications; 1999.

Stryhn H, De Vliegher S, Barkema HW, editors. Contextual multilevel models: Effects and correlations at multiple levels. In Proc XIth Int Symp Vet Epidem and Econ. Cairns, Aust. 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; 2005.