22

MIXED MODELS FOR DISCRETE DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand the differences between linear mixed models (continuous data) and generalised linear mixed models (GLMMs) (discrete and continuous data) and the role of the link function in the latter.

2. Fit random effects logistic and Poisson models.

3. Understand the differences between population-averaged and subject-specific modelling.

4. Use a latent variable approach to compute the intra-cluster correlation coefficient for binary outcomes.

5. Use either quasi-likelihood or maximum likelihood methods for fitting GLMMs.

6. Assess the statistical significance of both fixed and random effects in GLMMs.

7. Evaluate residuals to assess the adequacy of a GLMM that you have fit.

22.1 INTRODUCTION

In both theory and practice, it has proven more difficult than one might have anticipated to generalise the mixed-models approach from continuous to discrete data. One effect of these difficulties is the existence of a wide variety of generalisations of mixed models to discrete data, some of them only for a particular type of discrete data (mostly binary) and some of them within wider frameworks. In this chapter, we review the model class most analogous to linear mixed models: the **generalised linear mixed models** (GLMM). In order to fully appreciate this analogy, the reader is encouraged to review linear mixed models first (Chapter 21).

Our main focus here will be on binary data (logistic regression with random effects, Section 22.2) and on count data (Poisson regression with random effects, Section 22.3), but the random effects extension applies to a flexible class of discrete models which, for example, includes multinomial regression. As in Chapter 21, our mixed models will reflect a hierarchical structure but it is also possible to build models for other data structures. However—and this goes generally for mixed models for discrete data—the statistical analysis is more difficult than for continuous data, and requires more care and choices by the researcher (of which the choice of software is an important one). The field is still growing and advancing but we attempt to give the applied researcher a snapshot of its present state.

We will use the Brazilian data on the incidence of diarrhea already introduced in Chapter 2 to illustrate the modelling in this chapter. We consider both the full 4-level hierarchical data with individual binary outcomes nested within families, communities, and municipalities, as well as versions of the data aggregated to the family level with either a binary outcome (whether any family member reported diarrhea), a count outcome (the number of family members that reported diarrhea) or an ordinal outcome (the severity of diarrhea experienced in the family). The family-aggregated data do not allow accounting for predictors for individuals such as gender and age. Due to this loss of information, they are not recommended for a definitive analysis of the data but are used here for illustrative purposes. The data form a subset of a more detailed, complete dataset for which a 4-level logistic regression analysis has been reported (Marcynuk *et al*, 2012). Table 22.1 lists the variables used in the examples.

Table 22.1 Selected variables from the dataset brazil_smpl

Variable	Level of measurement	Description
mun	4:municipality	municipality identification (1-21)
comm	3:community	community identification (1-159)
fam	2:family	family identification (1-717)
id	1:individual	individual identification (1-3399)
diarr	1:individual	diarrhea within the last month (0/1)
di_cnt	1:individual	number of cases of diarrhea in the last month
age	1:individual	age in years
cistern	2:family	presence of a cistern (0/1)
water_tx	2:family	treatment of water with sodium chloride (0/1)
m_pop_families	4:municipality	number of families in municipality

22.2 LOGISTIC REGRESSION WITH RANDOM EFFECTS

Let us consider the example of disease observed for individuals in several groups (*eg* the families in the diarrhea data). The logistic regression analogue of Eq 21.2 for the probability p_i of the i^{th} individual being diseased is:

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_{\text{group}(i)} \qquad \textit{Eq 22.1}$$

where $u_{\text{group}(i)}$ is the random effect of the group containing individual i, assumed to be $u_{\text{group}(i)} \sim N(0,\sigma_g^2)$, the X_is are the predictor values for the i^{th} individual, and the relationship between the probability p_i and the binary outcome Y_i is unchanged: $p(Y_i=1)=p_i$. The only change from the ordinary logistic regression model is the group random-effects term. To keep things simple at our first acquaintance with the logistic random effects model, we first consider the diarrhea data aggregated to the family level and with families clustered within communities in Example 22.1.

22.2.1 Analogies and differences to a linear mixed model

We have seen that a mixed logistic regression model adds the random effects to the fixed effects, both on a logistic scale. So bearing the logistic scale in mind, we build the models in a similar way to linear mixed models and they might include multiple random effects and possibly random slopes as well. The statistical analysis also has strong similarities in the way confidence intervals and tests are computed.

The 2-level model (Eq 22.1) induces correlations between the observations in a similar way as its linear mixed model analogue: equal correlations between individuals within the same group and independence between groups. However, we have to be careful here: the correlations within a group are the same only for individuals with the same fixed effects. In our example, all -cistern- positive families within a community are equally correlated, and the same for all families without a cistern. This difference between individuals with different predictor values may seem strange and is usually small in practice (unless the predictor has a very strong effect). It is one of the many consequences of modelling the fixed and random effects on the logit scale. Nevertheless, the model is perfectly valid as a method to account for correlation (or clustering) between individuals in the same group.

Strictly speaking, the model in Eq 22.1 has a 2-step interpretation which is perhaps best understood by imagining how data would be simulated by the model. For an individual i in the j^{th} group, we would first select the group random effect (u_j) according to its $N(0, \sigma_g^2)$ distribution and compute p_i from the fixed effects and the selected u_j-value. We would then select the outcome Y_i as positive with probability p_i or negative with probability $1-p_i$. A common shorthand for this 2-step interpretation is that Eq 22.1 is 'conditional on' the random effects. In linear mixed models we modelled the outcome directly, so there was no need for a conditional interpretation.

In the next 2 sections, we describe how the interpretation of fixed and random effects parameters in the mixed logistic model changes from both the logistic regression model and the linear mixed model.

Example 22.1 Random effects logistic model for 2-level diarrhea data
data = brazil_smpl

The association between the presence of cisterns in the families' water supply, and the occurrence of diarrhea with at least one family member (-famdiarr-) was explored in the Brazilian diarrhea data. Families were sampled within 159 communities which in turn belonged to 21 municipalities, but for a 2-level hierarchy, we focus on families within communities because the presence of cisterns varied mostly between communities. A total of 436 (60.8%) out of the 717 families did not report diarrhea.

The unconditional association between -famdiarr- and -cistern- was:

		cistern		
		1	**0**	**Total**
famdiarr	1	115	166	281
	0	245	191	436
Total		360	357	717

Odds ratio=0.540
95% CI=(0.394,0.740)
Chi-square=15.9
P-value=0.0001

These statistics indicate a moderate but clearly significant association between -famdiarr- and -cistern-: absence of a cistern is associated with an odds ratio of 1/0.540=1.85. However, we have ignored the fact that the families came from 159 communities, and the prevalence of families reporting (at least one case of) diarrhea actually varied from 0% to 100% across communities. Consequently, it appears that we should be concerned about community effects. The logistic regression with random effects (Eq 22.1) gave the estimates:

	Coef	**SE**	**Z**	**P**	**95% CI**	
cistern	-0.703	0.182	-3.87	0.0001	-1.059	-0.347
constant	-0.154	0.138	-1.11	0.267	-0.425	-0.118

In addition, the estimated variance of community random effects (with SE) was:

$$\sigma_g^2 = 0.592\,(0.247)$$

We shall later see how to compute the significance of the random effect (it is highly significant). The regression coefficient for -cistern- should be compared with the log of the simple odds ratio (ln(0.540)=-0.616). Accounting for the community effects apparently increased the association slightly (we will discover this to be an artifact of the random effects model) but hardly changed the assessment of its significance. In other words, the communities neither had a substantial clustering nor a confounding effect.

22.2.2 Interpretation of fixed effects parameters

The interpretation of fixed effects in a linear mixed model was essentially unaffected by the added random effects. Again, the modelling on the logit scale complicates the interpretation of models such as Eq 22.1. The conditional interpretation of the model means that when exponentiating a regression coefficient (for absence of a cistern in the example) to obtain the odds ratio (*OR*) (*ie* OR=exp(0.703)=2.02), this odds ratio refers to comparing families with and without cisterns **in a particular community** (corresponding to a selected community random effect, no matter the actual u_j-value). Frequently this is called a **subject-specific** (SS) or **cluster-specific** (in our example, a community-specific) estimate, as opposed to a **population-averaged** (PA) or marginal estimate, which would refer to the *OR* for comparing families with and without cisterns **from any communities** in the population of communities (*ie* the 2 families

can be from different communities). Therefore, if we think of the *OR* as the answer to questions such as 'how much has the risk increased?' (in our example, the risk of family diarrhea for a family without a cistern versus a family with a cistern), the SS estimate answers the community-level answer (how much would the risk change within the community, *eg* if the family did not move but installed a cistern) and the PA estimate answers the question at a regional (or national) level (*ie* if families in different communities were compared, or a family moved between communities). That these 2 questions have different answers challenges our intuition, but is an incontestable fact.

Two alternatives exist to the SS odds ratio. One is to convert from SS to PA parameters (on the logit scale) using the following formula.

$$\beta^{PA} \approx \beta^{SS} / \sqrt{1 + 0.346\,\sigma_g^2} \qquad \textit{Eq 22.2}$$

Example 22.2 illustrates the procedure (see also Section 22.4 for further discussion of SS and PA estimates). The second alternative to the SS odds ratio is a reinterpretation of this value as a median odds ratio (*MOR*) across the population of clusters (communities in our case). The rationale behind this idea, introduced by Larsen *et al* (2000), is that when comparing 2 families, with and without cisterns, from different communities, the *OR* is really a random quantity because its value depends on the community effects for the 2 selected communities. Just as any other random variable, it has a distribution, and hence it makes sense to look for a central value in this distribution. The mean in this distribution is the PA odds ratio, and the median in this distribution (*MOR)* is equal to the SS odds ratio (computed at 2.02 above). We can now say that when comparing families with and without cisterns from the population, the odds ratio will, with probability 0.5, take values above and below 2.02. An associated range within which the *OR* will lie with a given probability, *eg* 80%, can also be computed; the details are shown in Example 22.2.

22.2.3 Interpretation of variance parameter(s)

In Eq 22.1, the group random effect variance σ_g^2 has no direct interpretation in terms of the probabilities of disease. The equation shows that it refers to the variation between groups of the disease probabilities on a logit scale. We can still interpret it qualitatively: a value of zero means no variation between groups (and therefore no clustering) and a large positive value means a high degree of clustering. However, the (correct) statement that the logits of probabilities vary within $\pm 1.96\sigma_g$ across groups with a probability of 95%, is not very intuitive. The variance σ_g^2 can, without too much extra work, be interpreted in terms of either variance components or median odds ratios; we discuss these in turn.

In linear mixed models, the variance parameters could be interpreted as **variance components**, but in models of discrete data, we have a problem with this interpretation. If we compare Eq 22.1 with the linear mixed model (Eq 21.2), the error term (ε_i) is missing in the logistic equation. This is because the distribution assumption is on the original scale—in our example $Y_i \sim$ bin(1, p_i), so that the errors in the model stem from the binomial (binary) distribution instead of a normal distribution. Recall that in this binary distribution the variance equals $p_i(1\text{-}p_i)$. Now the total variance in the data, var(Y_i), is no longer just the sum of the error variance and the random effects variance, as they refer to different scales. Even worse, the total variance is not constant because the binomial variance varies with p, so a single decomposition of the variance does not exist. Some years ago, several papers reviewed the computation of

variance components and intra-class correlation coefficients (*ICC*s, see Section 20.3.3; sometimes also denoted as variance partition coefficients (*VPC*s) in acknowledgement of the non-constant variances and correlations, as explained above) in random effects logistic regression, and a number of different methods were suggested (Browne *et al*, 2005; Goldstein *et al*, 2002; Rodriguez and Elo, 2003). We confine ourselves to explaining a simple approximation method based on **latent response variables** to represent the logistic model as a threshold model (latent variables were introduced in Chapter 17); see also Snijders and Bosker (1999), Section 14.3, and Rabe-Hesketh and Skrondal (2012), Chapter 10.

The simplest approach to getting both the individual and group variances on the same (logistic) scale is to associate with every individual *i* a latent continuous measure, Z_i, which represents the 'degree' of sickness (when the outcome measures disease). The observed binary outcome Y_i is then obtained simply as whether the degree of sickness exceeds a certain threshold (*ie* is sufficiently severe to be detected). In formulae, if we denote the threshold by *t*, then Y_i=1 if Z_i>*t*, and Y_i=0 when $Z_i \leq t$. Sometimes this may seem a plausible theoretical construct, and sometimes less so. Some diseases may be detected or reported in individuals if their clinical or quantitative signs exceed a certain level, but other events such as conception are truly binary in nature. Mathematically speaking, any model for Z_i is then translated into a model for the binary outcomes. In particular, Eq 22.1 is obtained exactly when *t*=0 and

$$Z_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_{\text{group}(i)} + \varepsilon_i \qquad \textbf{Eq 22.3}$$

where the fixed effects and the group effects are exactly as before, and where the error terms ε_i are assumed to follow a logistic distribution with mean zero and variance $\pi^2/3$=3.29. (The logistic distribution is similar in shape to the normal distribution, and for most practical purposes, it is equivalent to assume either of these distributions.) Eq 22.3 is a linear mixed model for Z_i. Therefore, computation of variance components and *ICC*s for Z_i-variables follows the rules of Chapter 21:

$$\text{var}(Z_i) = \text{var}(u_{\text{group}(i)}) + \text{var}(\varepsilon_i) = \sigma_g^2 + \pi^2/3$$
$$\rho = \sigma_g^2 / (\sigma_g^2 + \pi^2/3) \qquad \textbf{Eq 22.4}$$

We demonstrate the procedure in Example 22.2. To summarise, the latent variable approach allows interpretation in terms of variance components and *ICC*s by fixing the error variance at $\pi^2/3$. We should keep in mind that the strict interpretation is for the latent variables, and the values are only approximate for the binary outcomes. In particular, as noted, the variances and correlations are not constant for the binary outcomes but depend on the predictors; this dependence has disappeared for the latent variables. Experience with different methods for computing *ICC*s indicates that the latent variable *ICC* tends to be somewhat larger than the true *ICC* for the binary outcome (see the above-cited papers).

The variance σ_g^2 can also be interpreted in terms of an odds-ratio between the risk in 2 randomly selected clusters, where the individuals and groups compared should have the same fixed effects. The *OR* between the larger and smaller of the 2 risks is ≥1, and the median in its distribution (cluster-median odds-ratio, MOR_c) can be calculated as

$$MOR_c = \exp(0.954 * \sigma_g) \qquad \textbf{Eq 22.5}$$

(where 0.954 is a constant). The interpretation of the MOR_c is that when comparing (identical) subjects from 2 randomly selected clusters (groups), the odds ratio will, with probabilities 0.5,

Example 22.2 Interpretation of fixed and random parameters for 2-level diarrhea data
data = brazil_smpl

Based on the model presented in Example 22.1, we calculate an odds-ratio for the absence of a cistern as exp(0.703)=2.02. It can be interpreted either as a SS value (valid when comparing families with and without a cistern from the same community) or as the median odds ratio (*MOR*) across the population of communities. For the PA odds ratio, we first convert the parameter to marginal scale (using Eq 22.2):

$$\beta^{PA} = 0.703 / \sqrt{1 + 0.346 * 0.592} = 0.640,$$

and then compute the odds ratio the usual way as exp(0.640)=1.90. This is the mean odds ratio across the population of communities and is quite close to the unconditional (marginal) estimate of 0.540 seen in Ex. 22.1. The difference between the 2 *OR*s is modest here, due to the moderate (not very large) between-community variation. An 80% range for the *OR* of families selected from randomly selected communities is computed as:

$$80\%\ \text{range}:\ \exp(0.703 \pm 1.282\sqrt{2 * \sigma_g^2}) = \exp(0.703 \pm 1.395) = (0.50, 8.15),$$

where 1.282 is the Z_α for α=0.2 from Chapter 2 (the 90% percentile of *Z*). The range is wide and spans well across 1, essentially stating that the impact of the between-community variation is appreciable compared to that of the predictor (cistern).

Turning next to the variance parameter, we calculate by the latent variable approach a total variation of 0.592+3.290=3.882, and an *ICC* (and proportion of variance at the community level) of

$$\rho = 0.592 / 3.882 = 0.15.$$

Finally, the cluster-median odds-ratio for the random effect is calculated as:

$$MOR_c = \exp(0.954\sqrt{0.592}) = 2.08.$$

The median odds ratio for 2 comparable families (same fixed effects) from 2 randomly chosen communities is 2.08, which is about the same magnitude as the odds ratio comparing families from the same community with and without cisterns (2.02). This shows that even a moderate between-group variation has an appreciable impact on the individual risk.

take values above and below the MOR_c. One advantage of the MOR_c is that the heterogeneity between clusters is now comparable to the impact of the fixed effects (Larsen *et al*, 2000). Example 22.2 illustrates the procedure for the 2-level diarrhea data.

22.3 POISSON REGRESSION WITH RANDOM EFFECTS

A Poisson random intercept regression model with exposure *n* and group random effect *u* can be written:

$$\ln(\lambda_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_{\text{group}(i)}$$
$$Y_i \sim \text{Poisson}(n_i \lambda_i) \qquad \textit{Eq 22.6}$$

with the assumption: $u_{\text{group}(i)} \sim N(0,\ \sigma_g^2)$. Thus, the random effect is added to the fixed effects in a similar way as for logistic regression. The differences described in the previous section between models with and without random effects to a large extent carry over to Poisson regression. We briefly discuss the interpretation of fixed and random effects parameters from Eq 22.6, and illustrate by analysing the family counts of diarrhea cases (see Example 22.3).

Example 22.3 Random effects Poisson model for 2-level diarrhea data
data = brazil_smpl

In Example 22.1, we analysed the presence of family diarrhea cases in a logistic model accounting for clustering within communities. A more informative family-level outcome is the number of diarrhea cases with the number of family members included as the exposure, or population at risk. We extend the fixed effects part of the model by including also water treatment (-water_tx-). The 2-level model from Eq 22.6 is fit to the data for demonstration purposes; we discuss several relevant model extensions in a subsequent example (22.5).

log likelihood = -812.08

	Coef	SE	Z	P	95% CI	
cistern	-0.588	0.111	-5.32	<0.0001	-0.805	-0.371
water_tx	-0.309	0.111	-2.77	0.006	-0.527	-0.091
constant	-1.653	0.124	-	-	-1.897	-1.409

In addition, the estimated variance of community random effects was

$$\sigma_g^2 = 0.459\,(0.113).$$

Both water treatment and the presence of cisterns were associated with lower incidence of diarrhea, the count ratios being 0.73 and 0.56, respectively. The community-level variance was 0.459 (on the logarithmic scale) and thus fairly modest. Fig. 22.1 shows the estimated intraclass correlation (*ICC*) for 2 families within the same community, of the same size, ranging from 2–14, and both with either the highest or lowest incidence (with or without a cistern and water treatment, respectively). Despite the moderate between-group variance the *ICC* depends quite strongly on the predictors and on the population at risk (family size) because they both impact on the mean number of diarrhea cases in a family, and the *ICC* is increasing as a function of the mean. In datasets with strong variations in the mean (due to predictors or the population at risk), the *ICC* therefore, does not seem particularly useful to illustrate the group-level clustering.

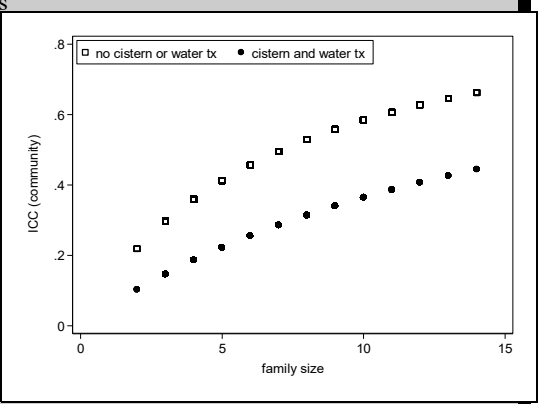

Fig. 22.1 Intra-class correlations for counts of family diarrhea cases within communities, across a range of family sizes and two predictor groups

22.3.1 Interpretation of fixed effects parameters

The distinction between SS and PA parameters largely vanishes in Poisson regression because it is only the intercept among the fixed effect parameters that takes different values in its SS and PA versions ((Diggle *et al*, 2002), Section 7.4). This perhaps somewhat surprising fact is related to the log link function and holds true regardless of the form of the random effects, *ie* also if they include multiple hierarchical levels or random slopes.

22.3.2 Interpretation of variance parameters

In a linear mixed model (Chapter 21), the *ICC* (for observations within the same cluster) could be computed by decomposing the total unexplained variance into terms for each level of the

hierarchy (see Examples 21.1 and 21.3). For a random effects logistic regression, a similar (approximate) calculation was enabled by the latent variable approach. For Poisson regression, exact formulae exist (Stryhn *et al*, 2006), but the resulting variance decomposition and *ICC* depend on the predictor values, so there is no longer a simple and unique *ICC* across the entire dataset (except for the 'null' model with no predictors). To simplify the notation, denote by βX a set of predictor values, including the logarithmic offset, of interest:

$$\beta X = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \ln(n).$$

Then the variances at level 1 (lowest) and level 2 (highest) as well as the *ICC* (and proportion of variance at the highest level) are given by:

$$\begin{aligned} &\text{level 1:} && \sigma(1) = \exp(\beta X + \sigma_g^2/2) \\ &\text{level 2:} && \sigma(2) = \exp(2\beta X + 2\sigma_g^2) - \exp(2\beta X + \sigma_g^2) \\ &ICC: && ICC = \sigma(2)/(\sigma(2) + \sigma(1)) \end{aligned} \qquad \textit{Eq 22.7}$$

It is recommended to calculate the *ICC* across a range of βX values of interest. Example 22.3 illustrates the procedure.

22.4 GENERALISED LINEAR MIXED MODEL

The examples of mixed models in the first 2 sections extend to a larger class of models called generalised linear mixed models. These models are constructed by adding the desired random effects on the transformed scale specified by the link function in the same way as we did in the logistic and Poisson regressions. The random effects are, as a practical rule, assumed normally distributed with mean zero but possibly involving some non-zero correlations (between random effects at the same level). The model assumptions listed in Section 16.5 are still valid, although the distributional form and the equation for the linear predictor are now conditional on the values of the random effects (Section 22.2.1). Also, the general discussion of correlation structure and interpretation of fixed effects and variance parameters from Section 22.2 carries over to GLMMs, but some of the specific procedures for binary/binomial data do not, *eg* the latent variable approach for computing variance components and *ICC*s. To keep things simple, we will confine ourselves throughout to random intercept models (*ie* no random slopes are included), but adding random slopes and contextual effects to GLMMs is possible (and relevant) in the same way as in linear mixed models (Sections 21.3 and 21.4).

In this section we set out by discussing in general terms the PA and SS interpretations of parameters in GLMMs (Section 22.4.1), and then we move on to specific models for binary, count, and categorical data (Sections 22.4.2–22.4.5).

22.4.1 Population-averaged versus cluster-specific parameters

The distinction between **population-averaged** (PA) and **cluster-specific** (subject-specific; SS) modelling for clustered data was introduced in Sections 22.2.1 and 22.3.1. Here we give more details and examples (largely following Diggle *et al* (2002)). First a note on the term 'subject-specific'. It originates from repeated measures data consisting of several observations (*eg* over time) on different subjects (Chapter 23); in this case, measurements are 'clustered' within subjects. In the context of our usual hierarchical clustering, we might instead have our subjects (*eg* individuals) clustered in groups. To avoid any confusion of this double use of 'subjects', we

shall refer to the upper level of the structure as clusters or groups (instead of subjects) but will remain with SS as the acronym in keeping with convention. Our next observation is that the PA and SS interpretations of regression coefficients are equivalent for linear mixed models. This is not due to their usual normal distribution assumption but to the fact that the linear predictor is modelled on the original scale; in the terminology of GLMMs, the link function is the identity function and there is no shift of scale. Therefore, the proper reference for our discussion is a GLMM with non-identity link, and we also assume a 2-level structure.

The difference between the PA and SS approaches is in the way the clustering or grouping of the data is dealt with. As previously seen, SS models include a random effect for each cluster in the linear predictor of the model. The assumptions for the random effects (*ie* their distribution and correlation) imply a particular form of the distribution of the set of observations within a cluster, including their correlation structure. Population-averaged or marginal models involve only the **marginal** means, *ie* the expected values for a particular set of predictors **averaged** across the **population of clusters**, and do not include specific effects for each cluster. To show the difference between the parameters involved in the 2 types of model in simple formulae, denote by Y our observations and by u the random effects for the clusters (in an SS model). Then, using the vector-matrix notation introduced in Section 21.2.2, GLM(M)s of SS and PA types are based on the equations:

$$\begin{aligned} &\text{cluster-specific:} && \text{link}[E(Y\,|\,u)] = X\beta^{SS} + Zu \\ &\text{population-averaged:} && \text{link}[E(Y)] = X\beta^{PA} \end{aligned} \qquad \textit{Eq 22.8}$$

where, as before (Eq 21.8), X and Z are our shorthand for the fixed and random part predictors of the model, and $E(Y|u)$ is the mean of Y conditional on the value of u (as discussed in Section 22.2.1). As indicated by the notation in Eq 22.8, the SS and PA regression parameters are not identical (unless the link is the identity link or there is no clustering). Generally the PA parameters are closer to the null ('attenuated') than their SS counterparts; we already noted this attenuation to be absent for identity and log links (except the intercept). The difference depends on the amount of clustering and is often small relative to estimation error. Formulae for specific models are given in Sections 22.4.2 and 22.4.3.

The selection of the most appropriate model type (SS or PA) depends on the predictor(s) being examined. Consider, as an example, a clinical trial on the effect of a treatment for a non-life-threatening condition such as the common cold (in this case the treatment will be a prophylactic aimed at preventing colds). The trial is carried out in multiple families; family members are randomised to receive either the treatment or a placebo, and after a certain period of time a dichotomous indicator of success (absence of a cold) is established for each family member. In addition to the treatment, the final logistic regression model with family random effects also may include the individual's age and the family's socioeconomic status. The β^{SS} for treatment in a SS model estimates the effect of the treatment in a **particular family** on probability of success (compared with the probability for a comparable family member). This makes sense for a person or family facing a decision on whether to try out the product. On the other hand, the β^{PA} gives the effect of this product being adopted on a large scale **across all families**. Thus, the interpretation has shifted from the specific family to effects across families. Regression parameters for age are interpreted similarly, now with β^{SS} referring to a comparison of family members of different ages, but age profiles of success in the entire population (*ie* across families, hence a PA interpretation) may be of interest as well. Note that an SS interpretation would become almost meaningless for family-level predictors that are inherent in the family (*eg* the socioeconomic status that would seem difficult to change). This problem with an SS

interpretation of a predictor that is unchangeable for the cluster is more common in repeated measures data where clusters are individuals, for example, with predictors such as sex or ethnicity.

A final note on recommendations for the (not uncommon) situation that a dataset/model contains multiple predictors, some of which have a desired SS interpretation and others with a desired PA interpretation. If a conversion formula such as Eq 22.2 exists for the model used, one should convert (or "marginalise", in the terminology of Hedeker and Gibbons (2006)) the parameters with the desired PA interpretation. It is generally more difficult to convert from PA to SS estimates because most PA estimation procedures do not have information about the variances (see Section 23.5).

22.4.2 GLMMs for binary data

Random effects logistic regression was introduced in Section 22.2. Here we add some comments about alternative link functions and give additional formulae for the conversion between SS and PA estimates, and for the latent variable *ICC*. Example 22.4 shows the results of fitting a multilevel logistic regression with predictors at multiple levels to the full diarrhea data; we will use this model to illustrate statistical methods for binary GLMMs.

For a random intercept logistic regression model, the approximation was presented in Eq 22.2 and is repeated here for convenience:

$$\beta^{PA} \approx \beta^{SS} / \sqrt{1 + 0.346\,\sigma_g^2}$$

where σ_g^2 is the group or cluster variance. By this formula, for PA parameters to be more than 10% lower than SS parameters, we need $\sigma_g^2 \geq 0.68$. For a model with multiple random intercepts (*eg* 3+ levels in the data hierarchy), Eq 22.2 still applies after replacing σ_g^2 by the sum of all variance components. When the model includes random slopes, the variance associated with the random effects is no longer constant across the data, so the conversion depends on the values of random-effects predictors (Z). The general formula for subject i can be written (see also Hedeker and Gibbons (2006), Section 9.7):

$$\beta_i^{PA} \approx \beta^{SS} / \sqrt{1 + 0.346 * \mathrm{var}\left((Zu)_i\right)} \qquad \textit{Eq 22.9}$$

where the variance computation follows similar lines as Eq 21.11 (depending on the details of the model).

For binary/binomial data, 2 occasionally encountered alternatives to the logit function are the so-called **probit** function (inverse cumulative probability for the standard normal) and the **complementary log-log** function. The choice of link function is largely dictated by the same considerations as for GLMs (Section 16.11); in practice, the difference in model fit and in the resulting statistical inference between the links is often minimal (see Example 22.4). For probit regression, the formula (22.9) becomes exact when the constant 0.346 is replaced by 1. The latent variable calculation of *ICC*s is valid for probit regression as well, by replacing the constant $\pi^2/3$ by 1.

22.4.3 GLMMs for count data

Random-effects Poisson regression was introduced in Section 22.2. Compared to mixed models

Example 22.4 Generalised linear mixed models (random effects logistic and probit regression) for 4-level diarrhea data

data = brazil_smpl

For analysis of the full dataset described in Section 22.1, we selected predictors at different levels. Our models include age (dichotomised as an indicator for young children of at most 5 years, -age5-) at the individual level, water treatment (-water_tx-) and the presence of a cistern (cistern) at the family level, and a municipality size recorded as the number of families in thousands of families and centred at 1000 families (-smsize-). We first discuss the estimates from a 4-level logistic regression model.

logL=-1219.23

	Coef	SE	Z	P	95% CI	
age5 (age<=5)	0.850	0.125	6.79	<0.0001	0.605	1.096
cistern	-0.861	0.180	-4.77	<0.0001	-1.215	-0.507
water_tx	-0.582	0.202	-2.88	0.004	-0.978	-0.186
mun. size (-smsize-)	0.178	0.114	1.56	0.119	-0.046	0.402
constant	-2.057	0.268	-	-	-2.581	-1.532

In addition, the estimated variances of the municipality, community, and family random effects were:

$$\sigma^2_m = 0.590\,(0.281),\quad \sigma^2_c = 0.057\,(0.168)\quad \text{and}\quad \sigma^2_f = 2.197\,(0.371)$$

We see that young children are at much higher risk of diarrhea, that both water treatment and the presence of a cistern reduce risk (as in previous examples) and that larger municipalities are possibly associated with increased risk. The between-family variance is very large, and there is some variation between municipalities and hardly any between communities. The *ICC*s between 2 observations from the same family (ρ_f) and between 2 observations from different families in the same community (ρ_c) can be estimated using the latent variable approach (Sections 21.2.1 and 22.2.2):

$$\rho_f\,(\text{individuals in the same family}) = \frac{0.590 + 0.057 + 2.197}{0.590 + 0.057 + 2.197 + \pi^2/3} = 0.464$$

and

$$\rho_c\,(\text{individuals in different families in the same community}) = \frac{0.590 + 0.057}{0.590 + 0.057 + 2.197 + \pi^2/3} = 0.105$$

The *ICC* for municipality will be almost the same as ρ_c. The corresponding probit random effects model achieved virtually the same model fit (as measured by the log-likelihood) and gave the estimates:

logL=-1219.63

	Coef	SE	Z	P	95% CI	
age5 (age<=5)	0.472	0.070	6.74	<0.0001	0.334	0.609
cistern	-0.475	0.099	-4.80	<0.0001	-0.669	-0.281
water_tx	-0.325	0.112	-2.91	0.004	-0.544	-0.106
mun. size (-smsize-)	0.100	0.062	1.59	0.111	-0.023	0.222
constant	-1.165	0.146	-	-	-1.452	-0.879

The estimated variances of the municipality, community and family random effects were, respectively:

$$\sigma^2_m = 0.174\,(0.082),\quad \sigma^2_c = 0.017\,(0.051)\quad \text{and}\quad \sigma^2_f = 0.691\,(0.115)$$

The probit regression estimates and SEs are, as expected, scaled towards zero by a factor close to 1.8 (*eg* 0.850/0.472=1.80); in practice, the scaling is often in the range 1.6–1.8 and thus, slightly less than the 'theoretical' scaling factor of $\pi/\sqrt{3}$=1.81 (Hedeker and Gibbons, 2006, Section 9.4). The variances are scaled by the square of this factor, and the latent variable *ICC*s are about the same (*eg* ρ_f =0.468).

for binary data, the choice of mixed models for count data is considerably more diverse and confused (from an applied point of view). One reason for this is the larger selection of models for count data, including several versions of negative binomial models and zero-inflated models (Chapter 18), all of which could be extended with random effects. Another reason is that both the Poisson model and its extensions can incorporate random effects of different types and with different distributions. In this section, we briefly indicate some of the models and demonstrate their fit to the diarrhea data of Example 22.3.

An alternative version of the random effects Poisson regression model in Eq 22.6 assumes a log-gamma distribution instead of a normal distribution for the random effects $u_{group(i)}$. What this really means is that the group random effects are $v_{group(i)}=\exp(u_{group(i)})$, and these are gamma-distributed variables which act as multiplicative random effects: $Y_i \sim \text{Poisson}(n_i \lambda_i v_{group(i)})$. One technical advantage of this model is that its likelihood function is easier to compute which facilitates likelihood-based inference (*eg* maximum likelihood estimation).

The negative binomial distribution (in its standard form) is parameterised by the mean λ and an added dispersion parameter α (Chapter 18). Clustering of the data (in groups) may manifest itself as similarity of the means within groups (while the dispersion is constant) or conversely as similarity of the dispersion within groups (while the means are constant). These 2 scenarios would be modelled by incorporating random effects in the means or in the dispersion parameters, respectively. Perhaps the most intuitive extension of the Poisson regression model has normally distributed random effects on the log-scale for the means (in a similar fashion as in Eq 22.6), but it may be numerically difficult to estimate. Access to specific negative binomial random effects models varies between statistical software and may involve manual programming of the model. Example 22.5 illustrates 3 alternative Poisson and negative binomial models fit to the family-level diarrhea data.

An additional question arises when it comes to extension of various forms of models for zero-inflated counts (Chapter 16). As these models have different modelling equations for the zero and non-zero portions of the data, there is choice between inserting random effects in any one of these equations or in both of them. In the latter case, the model will contain 2 random effects per cluster and these should probably be correlated. Zero-inflation models with random effects are, at the current stand of statistical software, beginning to become available, and further research and applications are likely to be appear in this field.

22.4.4 GLMMs for categorical data

The multinomial models of Chapter 17 can also be extended with random effects. Most attention in the literature has been paid to extensions of the proportional odds model (Section 17.5), and we illustrate the simplest of these in Example 22.6. The simple multinomial logistic regression model for nominal (Section 17.3) data can be extended with separate random effects in each model relative to the reference category ((Hedeker and Gibbons, 2006), Chapter 11), and similar extensions can be proposed for other multinomial models, although such models are not generally available in statistical software.

The proportional odds model is simpler to extend by random effects than the multinomial models because the fixed effects are expressed in a single equation. Adding random effects to this equation corresponds to adding random effects to the latent (unobserved) variables S_i in Eq 17.7. The subject-specific interpretation of estimates and the latent variable method for computing *ICC*s are virtually unchanged from the logistic regression models because of the

Example 22.5 Random effects count models for diarrhea data
data = brazil_smpl

The table gives maximum likelihood estimates (SE) and the log-likelihood value for Poisson regression models with normally and log-gamma distributed municipality random effects, a negative binomial regression model with normally distributed municipality random effects for the linear predictor of the mean (log scale), and a 3-level Poisson model with both community and municipality random effects.

Outcome dist.	Poisson	Poisson	Neg. binomial	Poisson
Random effects dist.	normal	log-gamma	normal	normal
cistern	-0.566 (0.093)	-0.567 (0.093)	-0.543 (0.115)	-0.586 (0.105)
water_tx	-0.313 (0.102)	-0.320 (0.101)	-0.340 (0.129)	-0.273 (0.110)
constant	-1.573 (0.166)	-1.401 (0.156)	-1.560 (0.174)	-1.670 (0.170)
σ^2(municipality)	0.394 (0.157)	0.345 (0.124)	0.343 (0.149)	0.333 (0.151)
σ^2(community)				0.180 (0.076)
log-likelihood	-806.50	-805.53	-782.74	-800.36

While the table still only covers a subset of the models, one might want to try for the data, we can extract some first conclusions. A comparison with the community-level random effects Poisson model in Example 22.3 shows that the models with municipality-level variance fit better (larger log-likelihood), but the 3-level model is the best of the Poisson models. The 2-level negative binomial model had a dispersion parameter of α=0.65 (0.14) and much improved fit, thus indicating substantial overdispersion related to the Poisson distribution. A 3-level negative binomial model would seem the obvious next model to try but ML estimation is not easily accessible in standard software for this model. Alternative distributions for random effects, such as the shown log-gamma distribution in a Poisson model and a beta distribution in a negative binomial model (not shown, see Cameron and Trivedie (1998) for details), did not offer substantial improvements. It is seen that the fixed effect estimates and their SE were relatively stable across the different modelling approaches. In such situations, model choice is often guided by ease (or meaningfulness) of the interpretation of model parameters.

similarity of the modelling equation. Several extensions have been proposed to account for possible lack of fit, including the possibility of allowing the scale of the linear predictor to vary between suitably chosen predictor groups. For example, it could be of interest to allow the scale to vary between raters in datasets with items graded by multiple raters ((Rabe-Hesketh and Skrondal, 2012), Chapter 11) or groups of subjects with particular characteristics related to smoking experience ((Hedeker and Gibbons, 2006), Chapter 10). In Example 22.6, we illustrate the proportional odds model with random effects by the diarrhea data with a trichotomous family-level outcome.

22.4.5 Other random effects models

Mixed models with the random effects on an original scale (instead of the transformed scale as in a GLMM) do exist, and we briefly mention 2 of them here.

The **beta-binomial** model has been used in many different ways in medicine and public health (*eg* Bandyopadhyay *et al* (2011); Gakidou and King (2002)). As indicated by the name, it is a model for binomial data incorporating beta-distributed random effects for probabilities. If the 2

Example 22.6 Random effects proportional odds model for diarrhea data
data = brazil_smpl

In this example, we consider again the family-level diarrhea data from Examples 22.1 and 22.2 where the presence and count of diarrhea cases in each family were analysed by logistic and Poisson regression models with community random effects, respectively. We seek to create an ordinal outcome for each family to distinguish 'mild' and 'severe' family cases among the 281 families with at least one individual affected. For the purpose of the example we quite arbitrarily define a mild case as one where the average number of diarrhea cases in the last month did not exceed 0.5 (per family member included in the data), and a severe case by this average exceeding 0.5. This definition designates 163 (22.7%) cases as mild and 118 cases (16.5%) as severe. The results of fitting a 3-level proportional odds model with the predictors -cistern- and -water_tx- were as follows.

	Coef	SE	Z	P	95% CI	
cistern	-0.691	0.163	-4.25	<0.0001	-1.010	-0.372
water_tx	-0.501	0.179	-2.79	0.005	-0.852	-0.149
cutpoint 1	-0.186	0.224				
cutpoint 2	1.140	0.232				

In addition, the estimated variances components for municipalities and communities were

$$\sigma_m^2 = 0.439\,(0.201) \quad \text{and} \quad \sigma_c^2 = 0.080\,(0.137).$$

The coefficient for -cistern- corresponds to the effect of cistern presence on the log(odds) for a 'mild' case compared with absence of diarrhea (as well as the log(odds) for a severe case compared with a mild case) and is hence, comparable to the coefficient of the logistic model in Example 22.1. The two estimates are indeed similar (with similar SEs), despite that the present model also includes -water_tx- (which does not strongly affect the coefficient for -cistern-) and includes an additional hierarchical level. Assessments of the goodness-of-fit of the proportional odds assumption did not reveal problems with this assumption; *eg* parameter estimates were quite similar for a logistic regression across the mild-severe threshold. The majority of the unexplained variance above the family level resides at the municipality level, and this is also true for the logistic regression model although we did not discuss this in Example 22.1, and the 3-level logistic regression model has quite similar variance estimates (not shown). The estimates here therefore agree well with the analysis of the dichotomous outcome, and the direction and relative magnitude of estimates also agree with the Poisson regression model (Examples 22.2 and 22.5), although no direct translation between estimates from these two model types exist.

parameters (α_1, α_2) of the beta-distribution are expressed in terms of the mean (μ) and the *ICC* (ρ), the model can be used as a regression model by incorporating predictors into a linear predictor on logit (or probit) scale just as in a GLM. The expressions for this reparametrisation are:

$$\alpha_1 = \mu\,(1-\rho)/\rho \quad \text{and} \quad \alpha_2 = (1-\mu)(1-\rho)/\rho\,,$$

In this model, the regression parameters will have a PA interpretation. It has been recommended as one of the best models for estimating the *ICC* (Ridout *et al*, 1999). One major advantage of the beta-binomial model is that the likelihood function is given by a relatively simple and explicit formula (which is not the case for GLMMs), and therefore the model is numerically simpler to compute than GLMMs (Andreasen and Stryhn, 2008). As one of its drawbacks, it does not, in a natural way, allow for predictors at the lowest level, nor does it have any easy extension to several hierarchical levels; it is essentially a model for grouped or replicated binary data. Recall the assumed relation for the variance of the grouped (binomial)

outcome (Eq 20.4); this assumption is different but not necessarily worse than other variance assumptions (*eg* the one implicit in a logistic random effects model); the fit of the beta-binomial model may be compared with that of other models by the log-likelihood or AIC statistics.

The **negative binomial** distribution was introduced in Chapter 18 as an extension of a **Poisson** distribution with overdispersion. Overdispersion could be understood as random variation in the mean (λ) of a Poisson-distributed variable (Y). Such variation may be attributed to 'inter-subject variability'—a heterogeneity between subjects not accounted for by the Poisson model. If λ has a gamma distribution with shape parameter $1/\alpha$ and scale parameter $\alpha\mu$ (equivalently: mean μ and variance $\alpha\mu^2$), then Y is negative binomial distributed with mean μ and variance $\mu+\alpha\mu^2$, as shown in Eq 18.9. This distribution may also be called a compound or mixture Poisson model. Note that these random effects cannot be used for modelling of a hierarchical structure, because they are already incorporated into the negative binomial distribution and because they are at the lowest (subject) level. But the negative binomial distribution can also be extended with random effects, as illustrated in Example 22.5; a recent book on negative binomial models gives a full theoretical treatment (Hilbe, 2011).

22.5 Statistical analysis of GLMMs

Despite the apparent simplicity of models such as Eq 22.1 and Eq 22.6, analysis of GLMMs is not straightforward, even in the logistic and Poisson regression settings. In contrast to most other models in the book, even the estimation of parameters is not clear-cut. Different methods exist, and they may give appreciably different results. No definitive answer exists at this point as to which method is generally preferable. The maximum likelihood method has been used throughout most parts of the book and is considered the standard choice here as well if it does not pose unmanageable computational challenges. Advances in computing power and software implementations over the last years has made maximum likelihood estimation feasible for moderate to large datasets and models. The implementation of GLMMs is still an active research area, and one is advised to investigate the options in different software before deciding on an approach (see also Section 22.6 for notes on current software). We outline briefly the methods available and indicate where they are discussed in this text.

1. Maximum likelihood estimation (Section 22.5.1): the likelihood function involves an integral over each random effect, which must be approximated by a summation and therefore makes ML estimation computationally demanding for large models.
2. Quasi-likelihood or iterative weighted least squares estimation (Section 22.5.2): algorithms for linear mixed models and GLMs are combined to produce multiple slightly different variants of an algorithm, which is fast and computationally simpler than ML estimation.
3. Bayesian MCMC (Markov chain Monte Carlo) estimation (Chapter 24): based on a different statistical approach (Bayesian statistics) and a simulation-based estimation that is computationally intensive.

All results shown so far in this chapter have been from ML estimation. But how does one determine which method is best, in general, for one's own data? One standard answer is to use simulation, *ie* generate artificial data from a model with known values of all parameters and then compare the results of different methods with those known values. Such simulation studies are regularly published in statistical journals (*eg* Browne and Draper (2006)), and you could also carry out your own simulation study for the data structure at hand (Masaoud and Stryhn, 2010).

Even if we as researchers are committed to always using the best possible method to analyse our data, it is useful to have a sense of when major differences between approaches might appear (see also the discussion of biases by quasi-likelihood procedures in Section 22.5.2). Estimation in GLMMs is most difficult if variances are large and/or the information contained in the data is limited, *eg* if replication is sparse. It is generally true that binary data are more difficult than count data, and that we should avoid fitting too ambitious models to even moderately sized binary datasets. Another possible cause of problems is if multiple clustering units should have 'extreme' predicted values, *eg* if in a logistic model all individuals in a group are negative (or all are positive). It almost goes without saying that whenever a dataset or model shows signs of being 'difficult' to estimate, one should be particularly careful with the analysis, and in such cases comparison of results from analyses using different procedures is often a fruitful approach.

22.5.1 Maximum likelihood estimation

Maximum likelihood (ML) estimation in GLMMs would, at first sight, seem to be our first choice, because of the overall strengths of the method (good statistical properties of the ML estimates) and the access to likelihood-based inference (*eg* likelihood ratio tests). However, ML estimation has, until recent years, had the reputation of being unfeasible for any GLMM beyond the simplest 2-level models, due to the massive and difficult computations required. Recent advances in computer power and software have changed this judgement, although the options currently available vary considerably among statistical software. It seems plausible that, within a few years, ML estimation will become the standard estimation approach for all but huge GLMMs. Even if the method's numerical side now looks promising, we outline why computation of the likelihood function is so difficult and give some cautions (complex procedures always have pitfalls, even if the complexities are hidden in the software).

For simplicity, consider the 2-level logistic regression model (Eq 22.1) and let us begin by focusing on a single group—group 1. Given the value of u_1 (the random effect of group 1), the conditional likelihood of the observations from that group is binomial,

$$L_1(\beta|u_1) = \prod_{i:\,\text{group}(i)=1} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

and the full (sometimes denoted marginal) likelihood for those individuals is obtained by integration over the distribution of the random effect u_1:

$$L_1(\beta) = \int L_1(\beta|u_1)(2\pi\sigma_g)^{-1/2}\exp(-\tfrac{1}{2}u_1^2/\sigma_g^2)\,du_1 \qquad \textit{Eq 22.10}$$

The integration weights the possible values of u_1 according to their likelihood in a normal distribution with mean zero and standard deviation σ_g. Integrals such as Eq 22.10 cannot be solved analytically, and therefore a numerical integration or **quadrature** becomes necessary. By this procedure, the integral is approximated by a weighted sum of values of the integrand (*ie* the function being integrated) at a number of selected quadrature points. Specific weighting schemes for integrals that involve exponential terms of a squared argument, as in Eq 22.10, are called Gauss-Hermite quadrature. In such schemes, you need to decide on the number of quadrature points and the way they are selected. Generally, increasing their number improves both accuracy and calculation time. Also, it is generally recommended that an **adaptive** approximation method be used, where the quadrature points (and their weights) are successively adapted to the integrand.

So far, we have dealt only with observations from one group. Observations from different groups are independent, so the full likelihood function for the entire dataset is obtained as a product of terms such as Eq 22.10 over the total set of groups. We trust it is not necessary to write out the equation to make the point that, not only computing, but also maximising, a quadrature approximation to such a multiple integral with respect to the fixed and random effects parameters of the model can be a formidable task. Extension to multiple levels and/or multiple random effects at the same level rapidly increases the complexity of the problem.

To summarise, a few recommendations and cautions for the use of ML estimation for GLMMs:

- ML estimation might be computationally unstable or the approximation of the likelihood function may be insufficient; it is highly recommended, therefore, that the stability of results be checked by trying different starting values of the algorithm and/or different variants of the numerical integration procedure, such as a different number of quadrature points as well as adaptive procedures,
- ML estimation could be compared with other approaches (either quasi-likelihood estimation or other approaches for clustered data), and caution should be exercised if major differences appear; this is in particular recommended if the estimation problem is 'difficult' (as discussed above),
- ML estimation in GLMMs may be impractical for model selection (because of computational demands); it is then considered legitimate to use computationally simpler methods for (part of) the model selection and then confirm the results by running selected models by ML estimation.

In Example 22.7, we examine the stability of the quadrature behind the ML estimates.

22.5.2 Quasi-likelihood estimation

A quasi-likelihood function could be thought of as a substitute for a (real) likelihood function whenever the latter does not exist or is too difficult to compute. In the early 1990s, when computers were much less powerful, several algorithms employing an iterative weighted least squares scheme were developed to maximise quasi-likelihood functions for GLMMs. These algorithms are referred to by many different acronyms, typically containing the letters QL (for quasi-likelihood), PL (for pseudo-likelihood) or ILS (for iterative and least squares), and often in combination with a G for generalised or a W for weighted or an R for reweighted or restricted. The main idea of the iterative weighted least squares methods is to compute an 'adjusted' variate on the scale given by the link function (*eg* logistic scale) in each step of the iteration. Technically, the adjusted variate is obtained by a Taylor expansion of Y around the current estimated mean, but one may think of it as a continuous version of the discrete outcome. Estimation for this adjusted variate is carried out using estimation procedures for linear mixed models (weighted REML or ML estimation). The procedure continues until convergence of the parameter estimates. Again, for the technically interested reader, some common options in the procedure are mentioned below:

- first- or second-order Taylor expansion, the latter being considered more accurate whenever the procedure converges,
- ML or REML estimation for the adjusted variate, the latter more commonly used,
- MQL or PQL form of the adjusted variate (M=marginal, P=predictive or penalised), the former being computationally more robust by omitting estimates of random effects in the linear predictor, and yields estimates with a PA interpretation (Breslow and Clayton, 1993), contrary to the other procedures for estimation in a GLMM.

Example 22.7 Checking maximum likelihood estimation of a GLMM
data = brazil_smpl

In Example 22.4, ML estimation was used to fit a random effects logistic regression model to the 4-level diarrhea data. This model was refit using a range of number of quadrature points (at the municipality, community, and family levels) in the estimation procedure (the estimates in Example 22.4 were obtained with 12 quadrature points at each level). The fixed and random effects estimates from each estimation were:

	Number of quadrature points at (munic.,comm.,family) level			
	(1,1,1)	**(3,3,3)**	**(7,7,7)**	**(12,12,12)**
age5 (age<=5)	0.836	0.839	0.850	0.850
cistern	-0.857	-0.851	-0.861	-0.861
water_tx	-0.583	-0.559	-0.583	-0.582
smsize	0.184	0.172	0.178	0.178
constant	-2.022	-2.012	-2.057	-2.057
σ^2(municipality)	0.384	0.558	0.590	0.590
σ^2(community)	0.055	0.121	0.054	0.057
σ^2(family)	1.920	1.847	2.205	2.197
log likelihood	-1226.911	-1223.604	-1219.189	-1219.226

The default in the software used for the estimation is 7 quadrature points. Using a single (1) quadrature point is sometimes referred to as a (lowest order) Laplace approximation (Section 22.5.2). With 20 quadrature points, exactly the same estimates were obtained as with 12 points. It is seen that low order quadrature may affect in particular the variance estimates quite substantially. From 7 quadrature points upwards the estimates were quite stable; thus the default number of quadrature points seems adequate.

These 3 options can be combined arbitrarily (depending on the facilities of the software package used). Example 22.8 shows results from some of these algorithms.

Many (statistical) papers have discussed the different versions of algorithms and their implementation in software packages (*eg* Browne and Draper (2006); Zhou *et al* (1999)). For well-behaved data, the different variants of the algorithms give very similar results (taking into account the standard errors of the estimates). One should whenever possible use the 'best' possible of the above options (second-order, REML, PQL). More importantly, any 'strange-looking' estimates or standard errors should cause the model to be examined carefully and the results to be confirmed with other models or estimation methods.

Early simulation studies showed that estimates from some of the iterative least squares algorithms for GLMMs could be markedly biased towards the null. The bias might affect both fixed and random-effect parameters, but the latter are particularly sensitive. The general consensus seems to be that particular caution should be exercised if:

- the number of replications at a hierarchical level is 'small' (*eg* less than 5),
- the corresponding random effect is 'large' (*eg* the variance exceeds 0.5).

In our Example 22.8, the number of family-level replications was moderate, with an average of 4.7 observations per family. The fact that differences in the regression coefficients were still appreciable is due to the fairly large variance components, in particular at the family level.

Example 22.8 Quasi-likelihood estimation of a GLMM
data = brazil_smpl

Three quasi-likelihood estimation procedures were applied to the 4-level logistic model of Example 22.4. The estimates were obtained using MLwiN software.

	ML (Ex 22.4)	1st order MQL	1st order PQL	2nd order PQL
	Coef (SE)	Coef (SE)	Coef (SE)	Coef (SE)
age5 (age<=5)	0.850 (0.125)	0.628 (0.104)	0.716 (0.112)	0.868 (0.132)
cistern	-0.861 (0.180)	-0.649 (0.133)	-0.697 (0.143)	-0.866 (0.202)
water_tx	-0.582 (0.202)	-0.415 (0.151)	-0.454 (0.160)	-0.580 (0.222)
smsize	0.178 (0.114)	0.129 (0.194)	0.139 (0.096)	0.195 (0.127)
constant	-2.057 (0.268)	-1.424 (0.194)	-1.649 (0.217)	-2.109 (0.290)
σ^2(municipality)	0.597 (0.281)	0.304 (0.128)	0.437 (0.177)	0.734 (0.313)
σ^2(community)	0.059 (0.168)	0.031 (0.080)	0.041 (0.091)	0.060 0.173)
σ^2(family)	2.197 (0.371)	1.190 (0.170)	1.140 (0.185)	2.472 (0.347)

The estimates of both first order quasi-likelihood procedures are generally closer to zero than the ML estimates in Example 22.4. The largest disagreement is seen for the family-level variance which is about half of the ML estimate. With underestimated variances the regression parameters in random effects models tend to be attenuated as well. The MQL estimates for the regression coefficients have a PA interpretation (and should therefore be closer to zero than the SS counterparts), but the scaling inherent in Eq 22.9 usually implies a similar bias towards zero for SS estimates when the variance is underestimated. Conversely, the second order PQL variance estimates were somewhat larger than the ML estimates, so the regression coefficients tended to be larger than their ML counterparts. This example illustrates how crucial unbiased variance estimation is for SS estimation; we discuss this further in Section 22.6. Note also that in this example, the Laplace approximation of Example 22.7 gives estimates closer to the ML estimates than the quasi-likelihood procedures.

We finally also mention a related type of approximation method that goes under the name Laplace approximation. The lowest order Laplace approximation corresponds to using a single quadrature point, but Laplace methods have been continually developed in the last decade, involving high order (*ie* more accurate) approximations and now also providing useful approximations to the log-likelihood function (so as to enable likelihood-based inference). A recent paper concluded that Laplace approximation methods were fully acceptable for Poisson regression (Pinheiro and Chao, 2006), and a further simulation study explored the accuracy of the method for both count and binary data (Joe, 2008).

22.5.3 Confidence intervals and tests

Statistical inference in GLMMs is generally only approximate (asymptotically correct when the number of observations at all hierarchical levels is large). Fixed effects parameters are usually assessed by Wald-type confidence intervals and tests, however likelihood-based inference (profile likelihood CIs and likelihood ratio tests, see Section 21.5) may be preferable, in particular when the parameters are highly correlated or not well-determined. However, likelihood-based inference is only feasible when ML estimation is used. As with GLMs, Wald-type statistics are useless for parameters that are 'out of bounds', *eg* in logistic regression when

one category of a predictor has no cases. Such situations often signal separation issues (Heinze and Schemper, 2002), which would typically also affect estimation of the random effects (therefore, also ML estimates may be affected).

Reference distributions are most commonly 'asymptotic', *ie* the standard normal or χ^2-distributions. The resulting inference may be too liberal if replication is sparse at the level of the parameter of interest, and some software packages give the option of using similar approximations as for linear mixed models using *t*- and *F*-distributions. In general, no clear guidelines can be given about the accuracy of approximate inference in GLMMs. As in linear mixed models, Wald-type statistics are inappropriate for variance parameters, which should therefore be assessed by likelihood-based inference (Example 22.9) or alternative procedures such as bootstrapping.

22.5.4 Prediction

The random effects in GLMMs can be predicted along similar lines as in linear mixed models (Section 21.5.4) and also exhibit a shrinkage towards the mean. However, some new issues arise for predictions of observations or their means, because the fixed and random effects reside on a different scale (*eg* logit scale). This is related to the distinction between SS and PA parameters. Following Skrondal & Rabe-Hesketh (2009), we describe 3 different ways of computing predicted probabilities in a random effects logistic regression model in Example 22.10. To keep things simple we consider the initial 2-level logistic regression from Example 22.1, but the ideas apply also to prediction in models with multiple levels (where one needs to consider the possible roles of the units at all hierarchies in the prediction).

22.5.5 Residuals and diagnostics

The standard tools for model-checking—residuals and diagnostics—are even less developed and accessible for GLMMs than for linear mixed models (Section 21.5.5). The distinction between different types of standard error still holds, but calculations are more difficult, and in practice one may need to accept whatever is offered by the statistical software (Skrondal and

Example 22.9 Statistical inference in a GLMM
data = brazil_smpl

The tests and confidence intervals given for the fixed effects in Example 22.4 are 'asymptotic'; for example, the 95% CIs are computed as $\beta \pm 1.96 * SE(\beta)$. Predictors at the family and community levels have ample replication for the asymptotic inference to be acceptable. With only 21 municipalities in the data, the inference for the municipality-level predictor -smsize- might be too liberal, but as this predictor did not achieve statistical significance, there is no great need for concern.

To compute tests for the random effects of the model, we note the log-likelihood value of the fitted model (-1219.226) and refit the model without the random effect of interest. The models without family, community or municipality random effects had log-likelihood values of (-1284.944), (1219.288) and (-1228.827), respectively, so that the corresponding χ^2-statistics with 1 df would be 131.4, 0.13, and 19.2, and thus totally non-significant for the community level and strongly significant for the other two. Recall from Section 21.5.3 that P-values should be computed as half the tail probability from the $\chi^2(1)$-distribution to account for the one-sided alternative hypothesis. To illustrate likelihood-ratio tests for fixed effects, the likelihood-ratio test for -smsize- gave $\chi^2(1)=2.18$ and a P-value of 0.14, close to the result of the Wald test in Example 22.4.

Example 22.10 Prediction in a 2-level random effects logistic regression
data = brazil_smpl

We consider again the simple family-level logistic regression model for diarrhea among 717 families in 159 communities with the presence of cisterns as the sole predictor and community random effects (Example 22.1). The ML estimates for the intercept, the coefficient for X (-cistern-) and the between-community variance were, respectively:

$$\beta_0 = -0.154\,,\quad \beta_1 = -0.703\,,\quad \sigma_g^2 = 0.592$$

We wish to predict the probability of diarrhea in families with and without cisterns. Three possible interpretations exist for such probabilities, when taking into account the community:

1. **Probability for families in a hypothetical community** For any given (hypothetical) community random effect u, we can compute a community-specific (conditional) probability as: $p(1)=\text{logit}^{-1}(\beta_0+\beta_1 X+u)$. Using $u=0$ gives the median probability across the population of communities. We can also insert $u=\pm 1.96\sigma_g$ to get a 95% range across this population.

2. **Mean probability for families from any community** The approximation formula Eq 22.2 gives this PA probability as: $p(2)\approx\text{logit}^{-1}((\beta_0+\beta_1 X)/\sqrt{(1+0.346\ \sigma_g^2)})$. In the table below, we used a more exact approximation for $p(2)$ based on quadrature (see Skrondal & Rabe-Hesketh (2009) for details).

3. **Probability for families in a specific community included in the study** When predicting for a specific community (as in the table below), we need to incorporate the information we have about its random effect. Due to the non-linearity of the logit function, simply inserting the predicted community random effect ($p(1)$ with u(comm)) does not work exactly; instead, we need to compute the mean probability averaged across the posterior distribution (in Bayesian terminology, see Chapter 24) of the random effect. Some statistical software will provide this calculation ($p(3)$).

The table below gives the 3 probabilities for the 2 categories of the predictor and two of the most extreme communities.

cistern	comm #	u(comm)	p(1) with u=0	p(2)	p(1) with u(comm)	p(3)
no	51	1.010	0.462	0.466	0.709	0.692
	64	-1.058			0.538	0.536
yes	51	1.010	0.298	0.319	0.229	0.244
	64	-1.058			0.128	0.142

As the intercept is close to zero, the first 2 predicted probabilities are close to 0.5 for families without a cistern. Some population-averaging towards 0.5 is seen for $p(2)$ when compared with $p(1)$ in families with a cistern. The predicted probabilities in study communities with actual estimated random effects are quite different (but these were extreme communities). The calculation with the estimated random effect inserted gives only slightly different values than the correctly calculated probability $p(3)$.

Rabe-Hesketh, 2009). The main new point for GLMMs (compared with linear mixed models) is that, because the model has no normally distributed error terms at the lowest level, the corresponding residuals and diagnostics at that level are difficult to assess. As an extreme example, in a binary model all the lowest-level residuals are dichotomous and cannot be expected to conform to a normal distribution. In this case, the residuals at the lowest level are not very informative. Unfortunately, the problems with the lowest-level residuals could penetrate to the higher levels if there is limited replication. Reference distributions and points

for residuals and diagnostics are therefore difficult to use rigorously, and one is advised instead to look for data points that are extreme in some way relative to the rest of the data. Example 22.11 presents residuals from our full 4-level analysis of the diarrhea data.

Example 22.11 Residuals from a 4-level GLMM
data = brazil_smpl

The 4-level logistic regression the full diarrhea data of Example 22.4 has residuals at all 4 hierarchical levels but the lowest-level residuals are of little use in this case so we disregard them completely. A normal (Q-Q) plot for the 717 family-level standardised residuals is given in Fig. 22.2. The plot shows a curious pattern, far from a straight line but instead with 2 separate, almost straight, lines. One must realise that with typically only 3–6 observations per family, these residuals are too discrete to realistically be expected to look like a normal distribution sample. For example, the lower part of the plot corresponds to families without any cases of diarrhea (so negative residuals) and the upper part of the plot to families with multiple cases. It seems difficult to assess from the plot whether there are problems with the normal distribution assumption at the family level.

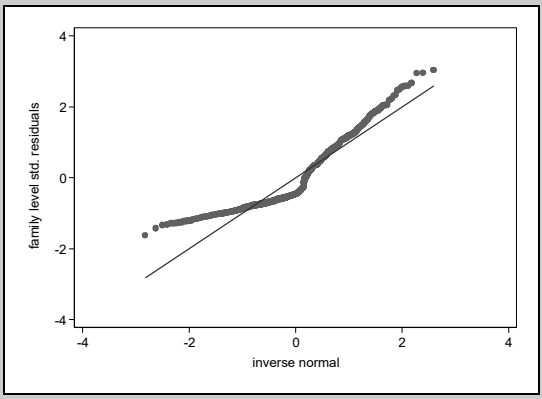

Fig. 22.2 Normal plot for family-level residuals of 4-level model for diarrhea data

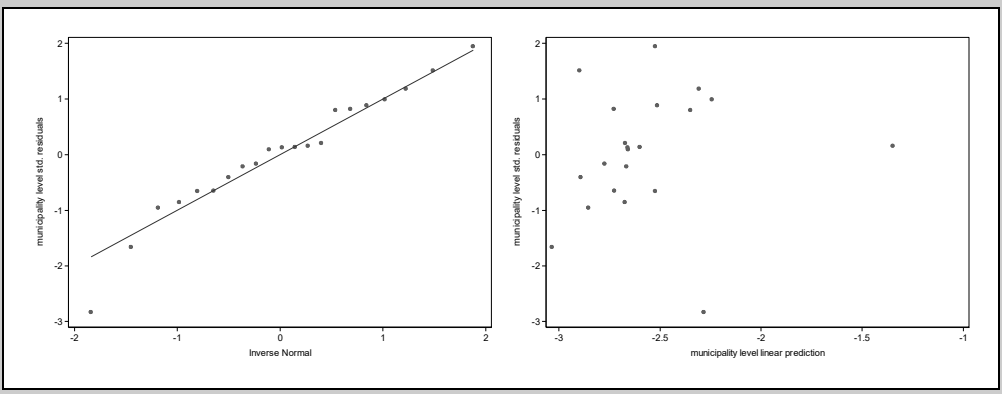

Fig. 22.3 Normal plot (left) and plot against predicted values (right) for municipality-level residuals of 4-level model for diarrhea data

As the community-level random effects do not explain any major part of the variation, we focus instead on the top (municipality) level. Fig. 22.3 shows the 21 municipality-level residuals depicted in a normal plot and plotted against the municipality-level predicted values based on all predictors. The normal plot is fairly straight but reveals one negatively outlying municipality, and the residual plot reveals another municipality with a much larger predicted values than the rest. Both of these municipalities warrant further inspection, in particular would one want to assess how influential they are for the results.

GLMM analogues of some of the special statistics for discrete data, such as the Hosmer-Lemeshow test for goodness of fit in a logistic-regression model, are not available. A simulation-based goodness-of-fit test for GLMMs has been described, but this procedure does not seem to be available in standard software (Waagepetersen, 2006).

22.5.6 Robustness against model misspecification

Much of the discussion of robustness of linear mixed model analysis to model specification in Section 21.5.7 carries over to GLMMs. One notable difference is that the use of robust standard errors is less obvious with non-normal data. A substantial body of research has been undertaken in the last decade on misspecification of GLMMs, in particular by McCulloch and Neuhaus whose work is summarised in McCulloch *et al* (2008), Chapter 12. Additional work includes (Heagerty and Kurland, 2001), and (Litiere *et al*, 2008). One conclusion that seems to be common for this work is that misspecification of the random effects distribution may not be terribly serious (McCulloch and Neuhaus, 2009).

22.5.7 Over- and underdispersion in GLMMs

Non-distributional dispersion in GLMs was discussed in Section 20.5.3, and it was explained how an extra-binomial dispersion parameter could be added to a binomial model within the GLM framework. A similar multiplicative dispersion parameter φ can be added to a Poisson model by the specification that $\text{var}(Y_i)=\varphi\lambda_i$, where $\varphi=1$ corresponds to the usual Poisson distribution, and also to other models for count and categorical data. We discussed in Section 20.5.3 how the extra-binomial parameter could account for clustering, although this method was not as attractive as other modelling approaches such as mixed models. In Chapter 16, we also discussed other ways that an apparent over- or underdispersion could arise. The question we address here is the utility of allowing for extra-distributional dispersion in mixed models, where the random effects should account for the hierarchical structure.

Our first observation is that extra-distributional parameters only exist within the GLM framework where models are incompletely specified and estimation is based on quasi-likelihood-type functions (Section 22.5.2). The fact that no data-generating mechanism exists for these models has been put forward as a major disadvantage of the approach (Skrondal and Rabe-Hesketh, 2007). It does indeed seem awkward to recommend the use of an approach that is only available in a subset of less attractive estimation procedures for GLMMs. Nevertheless, it is conceivable that a dataset could contain a dispersion that does not match the 'natural' distribution, even after the fixed and random effects have been incorporated into the model. In this sense, inclusion of an extra-distributional parameter may serve as a diagnostic tool. Values of φ substantially different from 1 would then lead us to either explore different distributions (where feasible), adopt the scaling of standard errors implicit in the quasi-likelihood estimation procedures, or perhaps ignore the finding.

If underdispersion is indicated, one should look for any reasons for negative correlations between observations, the standard example being competition in a group of individuals for a limited resource. If no such explanation can be found, as underdispersion means a better fit than expected to the data of our model, we often tend not to worry much about it (maybe it was just 'good luck'). By ignoring an appreciable underdispersion and pretending the dispersion to be as predicted by our model (when it is in reality smaller), our statistical inference becomes conservative—which may be considered the appropriate approach for 'a case of good luck'.

Underdispersion (as well as very small values of one-sided test statistics) may, however, also indicate something strange to be going on in the data, so one should inspect the data critically (once more). To scale down the standard errors by an underdispersion factor is a serious decision because it may lead to spurious significance, and should probably only be done when there is a biological explanation of the phenomenon. It might be useful to also try robust standard errors (Section 20.5.4) to see if they point in the same direction.

Overdispersion may be easier to understand intuitively and it may be considered less serious to inflate the standard errors; again, a comparison with robust standard errors might be useful. In some special cases, specific advice can be given on the modelling. First, if overdispersion is encountered in a Poisson model, it seems natural to try instead a negative binomial distribution (Chapter 18). Second, if overdispersion is encountered in a (mixed) model for grouped binary data (*ie* a binomial model with denominator >1), one may introduce a random effect at the group level which could then effectively remove the overdispersion (Browne *et al*, 2005; Skrondal and Rabe-Hesketh, 2007). Third, if the outcome is binary, extra-binomial dispersion cannot exist (Skrondal and Rabe-Hesketh, 2007); this is the same situation as in an ordinary GLM (Section 16.12). The same consideration exists for single categorical observations in a multinomial model. Notwithstanding this fact, many quasi-likelihood estimation procedures allow estimation of an 'extra-binomial' parameter for binary data, and many examples exist of such models fitted and published (Skrondal and Rabe-Hesketh, 2007). It is not clear what these estimation procedures actually estimate in the data, and interpretation of the estimated value of φ beyond a nondescript 'diagnostic' is hard to give. Skrondal and Rabe-Hesketh argue that the extra-dispersion parameter should be avoided in these instances.

22.6 SUMMARY REMARKS ON ANALYSIS OF DISCRETE CLUSTERED DATA

Throughout this chapter we have emphasised the distinction between cluster-specific (SS) and marginal (PA) modelling and interpretation of effects. We have also noted that SS parameters reside on a different scale than PA parameters, and that the difference between the 2 scales depends on the magnitude of the variance components (*eg* Eq 22.2). This scaling of SS parameters (relative to PA parameters) by a factor depending on the variances has the perhaps undesired consequence that SS parameters become difficult to compare between different datasets and analyses. As the estimates of variance parameters are particularly sensitive to the choice of estimation procedure, the fixed effects will, whenever the variances are large, be equally sensitive. This was the main reason behind our recommendation in Section 22.5 to exercise particular caution with the analysis when variances are large. In situations where the interest is in PA parameters, it seems awkward to start the process by obtaining estimates on another scale that may be difficult to establish firmly (when variances are large) before converting back to the scale of interest. Such reasoning has spurred the development of **marginalized models**, in which fixed effects are modelled on PA scale while a random structure is retained on SS scale ((Diggle *et al*, 2002), Chapter 11). Although first results with this new class of models were promising (Heagerty and Zeger, 2000), these models have not gained much popularity because they are not available in standard statistical software.

Several topics for mixed models covered in Chapter 21 (*eg* sample size) have not received a special treatment in this chapter because the coverage in Chapter 21 largely carries over or gives the relevant pointers also for GLMMs. The literature on GLMMs is huge and still rapidly expanding, including in recent years many excellent textbooks (often also covering linear

mixed models, see the brief overview in Section 21.5, and/or repeated measures data, see Chapter 23).

Let us at this point attempt a brief summary of the current status of statistical software for GLMMs. The field is more diverse and confusing than for linear mixed models, due to the existence of different estimation procedures and the continuing emergence of algorithms improved in speed and flexibility. For **maximum likelihood estimation** by numerical integration, Stata is arguably the most versatile statistical software package, because it offers both standard multilevel routines for binomial and count data with no restrictions on the number of hierarchical levels and cross-classification, and the powerful Generalised Linear Latent And Mixed Models (-gllamm-) macro for multilevel modelling implemented in Stata (Rabe-Hesketh and Skrondal, 2008). The -gllamm- software also implements a wide range of models involving latent variables (Skrondal and Rabe-Hesketh, 2004). Implementations of ML estimation in other packages is more limited, but updates are likely to occur rapidly. For **quasi-likelihood estimation,** many different implementations exist, both in general-purpose statistical software (*eg* SAS, R/S-Plus) and in specialised multilevel packages (MLwiN, HLM), with variable accuracy and flexibility of the algorithms. Another specialised package for mixed models (AD Model Builder) offers high order Laplace approximations.

A variety of approaches for dealing with clustered data has been presented in this and previous chapters, and 2 more are to come in Chapters 23 and 24. We conclude by revisiting the family level diarrhea data in which we present a comparative table of estimates and a brief discussion (Example 22.12).

Example 22.12 Summary of analyses for family-level diarrhea data
data = brazil_smpl

In order to enable a sensible comparison between approaches to account for a 2-level structure, only the clustering of families in the 21 municipalities was considered. Previous analyses in Examples 22.1 and 22.10 accounted (by random effects) for clustering at the community level, but just as in the full 4-level hierarchy the clustering is stronger at the municipality than the community level. The preferred approach is still to account for all hierarchical levels; see Section 23.5 for a comparison between GEE and GLMM in a 3-level hierarchy. The outcome evaluated was any diarrhea within the family and the models fit were:

- logistic - ordinary logistic model ignoring clustering
- robust - ordinary logistic regression with robust SEs
- fixed - ordinary logistic regression with municipalities fit as fixed effects
- stratification - Mantel-Haenszel estimation (stratified by municipality) (Chapter 13)
- GLMM - generalised linear mixed model
- GEE - generalised estimating equation (Chapter 23) using an exchangeable working correlation structure
- Bayesian (Chapter 24) - analysis reported the posterior median and SD from a mixed model with standard flat priors.

The grouping of estimates into SS (both GLMMs) and PA (logistic and GEE) was expected (see Example 20.4), whereby the fixed effects and stratified estimates fell somewhat outside the pattern. Due to the small number of clusters, robust SEs (also for GEE) may have been overestimated.

Predictor	Model	β	SE
cistern	logistic	-0.616	0.155
	robust variance	-0.616	0.219
	fixed effects	-0.684	0.166
	stratification	-0.648	0.161
	GLMM	-0.668	0.162
	GEE	-0.612	0.221
	Bayesian GLMM	-0.668	0.163

REFERENCES

Andreasen C, Stryhn H. Increasing weed flora in Danish arable fields and its importance for biodiversity. Weed Research (Oxford). 2008;48(1):1-9.

Bandyopadhyay D, Reich BJ, Slate EH. A spatial beta-binomial model for clustered count data on dental caries. Stat Methods Med Res. 2011 Apr;20(2):85-102.

Breslow NE, Clayton DG. Approximate inference in generalized linear models. Journal of the American Statistical Association. 1993;88:9-25.

Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis. 2006;3:473-514.

Browne WJ, Subramanian SV, Jones K, Goldstein H. Variance partitioning in multilevel logistic models that exhibit over-dispersion. J R Stat Soc A. 2005;168:599-613.

Cameron AC, Trivedie PK. Regression Analysis of Count Data. Cambridge: Cambridge University Press; 1998.

Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of Longitudinal Data, 2nd Ed. Oxford: Oxford University Press; 2002.

Gakidou E, King G. Measuring total health inequality: adding individual variation to group-level differences. Int J Equity Health. 2002 Aug 12;1(1):3.

Goldstein H, Browne WJ, Rasbash J. Partitioning variation in multilevel models. Understanding Statistics. 2002;1:223-32.

Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. Statistical Science. 2000;15:1-26.

Heagerty PJ, Kurland FK. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika. 2001;88:973-85.

Hedeker D, Gibbons RD. Longitudinal Data Analysis: Wiley; 2006.

Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002 Aug 30;21(16):2409-19.

Hilbe JM. Negative Binomial Regression - 2nd Ed. Cambridge: Cambridge University Press; 2011.

Joe H. Accuracy of Laplace approximation for discrete response mixed models. Computational Statistics & Data Analysis. 2008;52:5066-74.

Larsen K, Petersen JH, Budtz-Jorgensen E, Endahl L. Interpreting parameters in the logistic regression model with random effects. Biometrics. 2000 Sep;56(3):909-14.

Litière S, Alonso A, Molenberghs G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Stat Med. 2008 Jul 20;27(16):3125-44.

Marcynuk P, Flint J, Sargeant J, Jones A, Brito A, Luna C, et al. Comparison of the burden of diarrhoeal illness among individuals with and without household cisterns in northeast

Brazil: a cross-sectional study. ??? 2012;submitted.

Masaoud E, Stryhn H. A simulation study to assess statistical methods for binary repeated measures data. Prev Vet Med. 2010 Feb 1;93(2-3):81-97.

McCulloch C, Neuhaus J. Misspecification of the Distribution and Prediction of Random Effects in Generalized Linear Mixed Models. San Francisco: Division of Biostatistics, University of California; 2009.

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

Pinheiro JC, Chao EC. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. J Comput Graph Stat. 2006;15:58-81.

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

Ridout MS, Demétrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics. 1999 Mar;55(1):137-48.

Rodriguez G, Elo I. Intra-class correlation in random-effects models for binary data. The Stata Journal. 2003;3:32-46.

Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modelling: Chapman & Hall/CRC; 2004.

Skrondal A, Rabe-Hesketh S. Redundant overdispersion parameters in multilevel models for categorical responses. Journal of Educational and Behavioral Statistics. 2007;32:419-30.

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, Sanchez J, Morley P, Booker C, Dohoo IR. Interpretation of variance parameters in multilevel Poisson regressions models. Proc Int Symp on Vet Epidemiol and Econ; 2006; Cairns, Australia.

Waagepetersen R. A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. Scandinavian Journal of Statistics. 2006;33:721-31.

Zhou XH, Perkins AJ, Hui SL. Comparisons of software packages for generalized linear multilevel models. American Stat. 1999;53:282-90.