22

Multilevel Models

22.1 INTRODUCTION

In Parts I through IV the major focus has been on the analysis of longitudinal data. A distinctive feature of longitudinal data is that they are *clustered.* In longitudinal studies the cluster is composed of the repeated measurements obtained from a single individual at different occasions. There are, however, many studies in the health sciences that are not longitudinal but that give rise to data that are clustered or cluster-correlated. For example, clustered data commonly arise when intact groups are randomized to health interventions in so-called cluster-randomized trials or when naturally occurring groups in the population are randomly sampled. In addition there can be more than a single level of clustering in the data. The term *multilevel data* (or *hierarchical data*) encompasses all of these cases. The distinctive feature of multilevel data is that measurements on units within a cluster are more similar than measurements on units in different clusters. The clustering can be expressed in terms of correlation among the measurements on units within the same cluster and this correlation must be appropriately accounted for in the analysis.

Because longitudinal data are a special case of multilevel data, with only a single level of clustering and a natural ordering of the measurements within a cluster, this chapter provides a description of regression models for multilevel data, more broadly defined. One of our goals is to demonstrate that many of the methods for the analysis of longitudinal data considered in earlier chapters are, more or less, special cases of more general regression methods for multilevel data. The overview of multilevel models presented in this chaper provides a basic introduction to a general methodology

Applied Longitudinal Analysis, Second Edition. By Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware.

for analyzing the wide range of clustered data that commonly arise in studies in the biomedical and health sciences.

22.2 MULTILEVEL DATA

Multilevel data arise when there is a hierarchical or clustered structure to the data. Data of this kind frequently arise in the health sciences, since individuals can be grouped in so many different ways. For example, in studies of health services and outcomes, assessments of quality of care are often obtained from patients who are nested within different clinics. Such data can be regarded as multilevel, with patients referred to as the level 1 units and clinics the level 2 units. In this example there are two levels in the data hierarchy and, by convention, the lowest level of the hierarchy is referred to as level 1. The term "level," as used in this context, signifies the position of a unit of observation within a hierarchy.

Broadly speaking, the clustering in multilevel data can be a consequence of the study design or due to a naturally occurring hierarchy in the target population, or sometimes due to both. An example of a naturally occurring two-level data hierarchy arises in developmental toxicity studies. In a typical developmental toxicity experiment, pregnant mice or rats (dams) are assigned to increasing doses of a chemical or a test substance over the period of major organogenesis (when organ systems are developing in a growing fetus). Following sacrifice, each fetus in the litter is weighed (a continuous response) and examined for evidence of malformations (a binary response, present or absent). Data collected in developmental toxicity experiments are clustered (i.e., the litter is the cluster), with observations on the fetuses (level 1 units) nested within dams/litters (level 2 units). Two-level data also arise in family studies designed to assess the association or "aggregation" of disease (or markers of disease development) among relatives. In family studies the goal is to determine whether the presence of disease in a family member is associated with increased risk of disease for relatives. The associations among siblings and between parent-child pairs are of main interest because their relative magnitudes can be used to provide indirect evidence of genetic risk due to the sharing of the same genes. Data from studies of nuclear families are clustered, with observations on the mother, father, and children (level 1 units) nested within families (level 2 units).

Other common examples of naturally occurring clusters in the population are households, hospital wards, clinics, medical practices, neighborhoods, and schools. Furthermore naturally occurring hierarchical data structures can have more than two levels. For example, observations may be obtained on patients nested within clinics, which in turn are nested within different geographical regions of the country. Another example of a naturally occurring data hierarchy is when observations are obtained on children nested within classrooms, nested within schools. In both of these examples there are three levels in the data hierarchy. In principle, there can be many levels in the data hierarchy.

Alternatively, the hierarchical data structure can be a consequence of the study design. For example, the U.S. National Health and Nutrition Examination Survey

(NHANES) uses a multi-stage sampling design to produce information on nutrition and health status. The target population is the total U.S. civilian non-institutionalized population, 2 months of age or over. Because it is not practical to obtain a simple random sample of the U.S. population, complex sampling methods are commonly used. For example, NHANES III, conducted in 1988 to 1994, used the following multi-stage sampling design (National Center for Health Statistics, 1992, 1994). In the first stage, so-called primary sampling units (PSUs) were defined based on counties or combined counties in the United States. A first-stage random sample of PSUs was selected from these geographical regions. In the second-stage sampling, within each of the selected PSUs, a random sample of area segments consisting of census blocks was selected. In the third stage, within each of the selected area segments, a random sample of households was selected. Finally, in the fourth stage, eligible persons were randomly selected within households. The resulting data can be regarded as hierarchical, with individuals being the level 1 units, households the level 2 units, area segments the level 3 units, and counties the level 4 units.

Additional examples of study designs that produce multilevel data structures include cluster-randomized clinical trials, repeated measures experiments, and longitudinal studies. In a cluster-randomized trial, groups of individuals, rather than the individual subjects, are randomized to different treatments or health interventions. For example, the Promotion of Breastfeeding Intervention Trial (PROBIT) was designed to determine whether efforts to promote breastfeeding have any impact on the duration and exclusivity of breastfeeding (Kramer et al., 2001). In this trial maternity clinics, rather than the mothers, were randomized to either the intervention or control (standard care). The mothers were followed-up for one year after the birth of their infants and the effectiveness of the health intervention was assessed by the responses of mothers in each treatment group. When regarded as multilevel data, the level 1 units are the mothers and the level 2 units are the maternity clinics. Of note, the main covariate of interest, denoting the assignment to intervention or control, is defined at level 2. Longitudinal studies are another common example where the study design produces data with a two-level structure. In a longitudinal study the clusters are composed of the repeated measurements obtained from a single individual at different occasions. When longitudinal data are regarded as multilevel data, the level 1 units are the repeated occasions of measurement and the level 2 units are the subjects.

Finally, the clustering in multilevel data can be due to both the design of the study and naturally occurring hierarchies in the target population. For example, clinical trials are often conducted in many centers to ensure sufficient numbers of patients and/or to assess the effectiveness of the treatment in different settings. These studies are referred to as multi-center trials. Observations from a multi-center longitudinal clinical trial can be regarded as multilevel data having 3 levels, with repeated measurement occasions (level 1 units) nested within subjects (level 2 units) nested within clinics (level 3 units).

Although we have distinguished between clustering that occurs naturally and clustering due to study design, the consequence of clustering at different levels is the same: units that are grouped at any level are likely to respond more similarly. For example, two patients selected at random from the same clinic are expected to respond more

similarly than two patients randomly selected from different clinics. In general, the degree of clustering can be expressed in terms of correlation among the observations on units within the same level. Statistical models for multilevel data must account for the intra-cluster correlation at each level; failure to do so can result in misleading inferences.

22.3 MULTILEVEL LINEAR MODELS

In this section we discuss linear models for multilevel data. The dominant approaches to multilevel modeling have the same basis: clustering in the data is accounted for via the introduction of random effects at different levels in the hierarchy. Multilevel linear models can be regarded as extensions of the linear mixed effects models described in Chapter 8, which allow random effects to be incorporated at more than one level. In addition to accounting for clustering in the data, multilevel models permit estimation of the effects of covariates, measured at any of the levels of the hierarchy, on the outcome.

In a multilevel model the response is obtained on the lowest level (or level 1) units, but covariate information can by measured at any level. Combining covariates measured at different levels of the hierarchy within a single regression model is central to multilevel modeling. For example, multilevel models can determine and disentangle the relative importance of patient-level, clinic-level, and regional-level factors on quality of care. In general, multilevel models can be used to make inferences about the population of units at any level of the hierarchy and to discern how variation in the outcome at different levels depends on covariates. In this section we present an overview of multilevel linear models for a continuous outcome. We begin with a discussion of models for two-level data. The models generalize in a natural way when there is additional clustering in the data (e.g., three-level and higher-level data). The major focus of this section is on the specification of multilevel models; estimation is mentioned but not emphasized.

22.3.1 Two-Level Linear Models

Before describing models for two-level data we need to introduce some notation. For a two-level data structure, let i index level 1 units and j index level 2 units. We assume that there are n_2 units at level 2 in the sample. Each of these clusters (for $j = 1, ..., n_2$) is composed of n_{1j} level 1 units. For example, consider a multi-center clinical trial comparing two treatments (active drug versus placebo) conducted in 20 medical clinics. Patients are enrolled from each clinic and randomly assigned to one of the two treatment conditions. In this example, clinics are the level 2 units ($j = 1, ..., 20$) and patients are the level 1 units ($i = 1, ..., n_{1j}$), where n_{1j} is the number of patients enrolled in the study from the j^{th} clinic (and $n_2 = 20$ is the number of clinics). Alternatively, consider two-level data arising from a longitudinal study where 150 subjects are measured at four occasions. In this example, subjects

are the level 2 units ($j = 1, ..., 150$), and measurement occasions are the level 1 units ($i = 1, ..., 4$), with $n_2 = 150$ level 2 units, and $n_{1j} = 4$ level 1 units (within each level 2 unit). For the latter example the alert reader will have noticed that the indices i and j have now been reversed from their use in earlier chapters; here we have adopted the usual convention in much of the multilevel modeling literature of letting i denote level 1 units, j denote level 2 units, and so on. We must caution the reader that some of the literature on multilevel modeling reverses this notation (and/or occasionally reverses the ordering of the levels).

Let Y_{ij} denote the response on the i^{th} level 1 unit within the j^{th} level 2 cluster. For example, Y_{ij} might denote the primary outcome for the i^{th} patient in the j^{th} clinic. Associated with each Y_{ij} is a $1 \times p$ (row) vector of covariates, X_{ij}. These can include covariates defined at each of the two levels and can also include "compositional" covariates, so called because they are formed by aggregating values over lower level units. For example, severity of disease defines a patient-level (or level 1) covariate. However, a "compositional" covariate at the clinic level can be formed by taking the average disease severity for all patients within each clinic.

Consider the following linear model relating the mean response to the covariates:

$$E(Y_{ij}|X_{ij}) = X_{ij}\beta. \tag{22.1}$$

For example, in a multi-center clinical trial, a simple model for the mean response is given by

$$E(Y_{ij}) = \beta_1 + \beta_2 \, \text{Group}_{ij},$$

where Group_{ij} denotes the treatment assignment for the i^{th} patient in the j^{th} clinic, with $\text{Group}_{ij} = 1$ for active drug and $\text{Group}_{ij} = 0$ for placebo. The model given by (22.1) specifies how the mean response depends on covariates, where the covariates can be defined at level 2 and/or level 1. A multilevel model accounts for the variability in Y_{ij}, around its mean, by allowing for random variation across both level 1 and level 2 units. In particular, a multilevel model for Y_{ij} assumes there is random variation across level 1 units and random variation in a subset of the regression parameters across level 2 units. The two-level linear model for Y_{ij} is given by

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_j + \epsilon_{ij}, \tag{22.2}$$

where Z_{ij} is a design matrix for the random effects at level 2, formed from a subset of the appropriate columns of X_{ij}. The random effects, b_j, vary across level 2 units but, for a given level 2 unit, are constant for all level 1 units. These random effects are assumed to be independent across level 2 units, with mean zero and covariance, $\text{Cov}(b_j) = G$. The level 1 random components, ϵ_{ij}, are also assumed to be independent across level 1 units, with mean zero and variance, $\text{Var}(\epsilon_{ij}) = \sigma^2$. In addition the ϵ_{ij}'s are assumed to be independent of the b_j's, with $\text{Cov}(\epsilon_{ij}, b_j) = 0$. That is, the level 1 units are assumed to be conditionally independent given the level 2 random effects (and the covariates).

The regression parameters, β, are the fixed effects and describe the effects of covariates on the mean response

$$E(Y_{ij}) = X_{ij}\beta,$$

where the mean response is averaged over both level 1 and level 2 units. The two-level model given by (22.2) also describes the effects of covariates on the conditional mean response

$$E(Y_{ij}|b_j) = X_{ij}\beta + Z_{ij}b_j,$$

where the response is averaged over level 1 units only.

Let us return to the multi-center clinical trial example introduced earlier. A simple two-level model for the data is given by

$$Y_{ij} = \beta_1 + \beta_2 \,\mathrm{Group}_{ij} + b_{1j} + \epsilon_{ij},$$

where b_{1j} is a random clinic effect. The random effect b_{1j} varies across clinics but, for a given clinic, is constant and shared by all patients belonging to that clinic. The inclusion of b_{1j} accounts for the clustering of patients within clinics, due perhaps to similarities in severity of illness and/or quality of care. The model explicitly accounts for the fact that some clinics have patients that respond higher (or lower) than patients in other clinics. However, the model assumes that the effect of treatment is the same across all clinics. This assumption can be relaxed by allowing the effect of treatment to vary among clinics

$$\begin{aligned} Y_{ij} &= \beta_1 + \beta_2 \,\mathrm{Group}_{ij} + b_{1j} + b_{2j}\,\mathrm{Group}_{ij} + \epsilon_{ij} \\ &= (\beta_1 + b_{1j}) + (\beta_2 + b_{2j})\,\mathrm{Group}_{ij} + \epsilon_{ij}. \end{aligned}$$

In this model the magnitude of the effect of treatment varies randomly across the different clinics. The average effect of treatment, when averaged over the population of clinics (and not simply those included in the trial), is β_2.

The example just presented involves randomizing patients (level 1) within clinics (level 2). In the language of experimental design, patients (level 1) are *nested* within clinics (level 2), but treatment is *crossed* with clinics because patients within each clinic are randomized to each treatment. Another very different type of design is one where patients (level 1) are nested within a clinic (level 2), but clinics are randomized to treatments, so that all patients from any given clinic receive the same treatment. In this case clinics are nested within treatment and not crossed. Formally, the same model just presented can be used for the analyses of these data, except that the effect of treatment can no longer vary randomly across the different clinics, since each clinic is assigned to only one treatment group. Note also that the treatment group variable, Group_{ij}, does not vary over i for fixed j, and hence can be replaced by Group_j. However, it is important to note that the nesting of clinics within treatment has a negative impact on efficiency of the treatment effect estimate, relative to a design with no nesting. This general principle will be illustrated later with analyses of the *Television, School and Family Smoking Prevention and Cessation Project*. In this study schools were randomized to treatments, and in the analysis both classroom and student variability were accounted for. The first design, where clinics are crossed with treatment, is generally more efficient than a design which does not stratify on clinic. The principle behind this is the same as that of a longitudinal study, where we can generally measure change more efficiently by using repeated measures on the same subject than using a cross-sectional design.

So far our discussion of two-level models has very closely paralleled the description of the linear mixed effects model given in Chapter 8. In Chapter 8 we focused on models for two-level data where measurement occasions are the level 1 units and subjects are the level 2 units. However, it should be recognized that longitudinal data are simply a special case of two-level data and the linear mixed effects model given by (22.2) can be applied more broadly.

Finally, the two-level model given by (22.2) can also be written in terms of two models, one for each level of the hierarchy, using the two-stage formulation described in Section 8.4. That is, the two-level model can be expressed in terms of a level 1 model,

$$Y_{ij} = Z_{ij}\beta_j + \epsilon_{ij},$$

where ϵ_{ij} are assumed to be independent across level 1 units, with mean zero and variance, $\text{Var}(\epsilon_{ij}) = \sigma^2$, and a level 2 model,

$$\beta_j = A_j\beta + b_j,$$

where b_j are assumed to vary independently across level 2 units, with mean zero and covariance, $\text{Cov}(b_j) = G$. Substituting the second model equation into the first yields (22.2)

$$\begin{aligned} Y_{ij} &= Z_{ij}(A_j\beta + b_j) + \epsilon_{ij} \\ &= (Z_{ij}A_j)\beta + Z_{ij}b_j + \epsilon_{ij} \\ &= X_{ij}\beta + Z_{ij}b_j + \epsilon_{ij}, \end{aligned}$$

where $X_{ij} = Z_{ij}A_j$. An advantage of specifying a multilevel model in terms of a series of models for each level of the hierarchy, rather than as a combined model, is that it becomes more transparent which covariates are operating at which level of the model. However, this does introduce some unnecessary restrictions on the model of the kind discussed in Section 8.4.

In summary, the two-level linear model given by (22.2) accounts for the clustering of the level 1 units by incorporating random effects at level 2. The model explicitly distinguishes two main sources of variation in the response: variation across level 2 units and variation across level 1 units (within level 2 units). The relative magnitude of these two sources of variability determines the degree of clustering in the data. The larger the variance of the level 2 random effects, relative to the level 1 (within level 2) variability, the greater is the degree of clustering. Next we describe how the linear mixed effects model can be generalized to three-level data structures; the extensions to four or more levels follows directly.

22.3.2 Three-Level Linear Models

As mentioned earlier, there can be many levels in the data hierarchy. The extension of the two-level model given by (22.2) to three or more levels is very natural. With

three-level data, there is clustering in the data that is assumed to be due to variation in the response across level 1, level 2, and level 3 units. Although the extension from 2 to 3 levels is conceptually straightforward, the description of the three-level model does require the introduction of additional notation that often obfuscates the salient features of the model. The basis of a three-level model is that variability in the response is accounted for by the introduction of random effects at all higher levels in the hierarchy (e.g., by allowing random variation in a subset of the regression parameters at both levels 2 and 3). The model explicitly distinguishes three sources of variation in the response: (1) variation across level 3 units, (2) variation across level 2 units (within level 3 units), and (3) variation across level 1 units (within level 2 units nested within level 3 units).

For a three-level data structure, let i index level 1 units, j index level 2 units, and k index level 3 units. We assume that there are n_3 units at level 3 in the sample. Each of these clusters (for $k = 1, ..., n_3$) is composed of n_{2k} level 2 clusters, and each of these, in turn, is composed of n_{1jk} level 1 units. For example, consider a multi-center *longitudinal* clinical trial comparing two treatments (active drug versus placebo) conducted in 20 different centers or clinics. Patients in each clinic are measured at baseline and at three post-treatment occasions. In this example clinics are the level 3 units ($k = 1, ..., 20$), patients are the level 2 units ($j = 1, ..., n_{2k}$), and measurement occasions are the level 1 units ($i = 1, ..., 4$), where n_{2k} is the number of patients in the k^{th} clinic ($n_3 = 20$ and $n_{1jk} = 4$, for all j and k).

Let Y_{ijk} denote the response of the i^{th} level 1 unit within the j^{th} level 2 cluster within the k^{th} level 3 cluster. For example, in a multi-center longitudinal clinical trial, Y_{ijk} denotes the outcome at the i^{th} occasion for the j^{th} patient in the k^{th} clinic. Associated with each Y_{ijk} is a $1 \times p$ (row) vector of covariates, X_{ijk}, with the covariates defined at different levels. A three-level model for Y_{ijk} is given by

$$Y_{ijk} = X_{ijk}\beta + Z^{(3)}_{ijk} b^{(3)}_k + Z^{(2)}_{ijk} b^{(2)}_{jk} + \epsilon_{ijk}, \tag{22.3}$$

where $Z^{(3)}_{ijk}$ is a design matrix for the random effects at level 3, formed from a subset of the appropriate columns of X_{ijk}, and $Z^{(2)}_{ijk}$ is a design matrix for the random effects at level 2 (also formed from a subset of the appropriate columns of X_{ijk}). In this notation the superscripts attached to $b^{(3)}_k$ and $b^{(2)}_{jk}$ denote the levels at which the random effects vary. In general, the design matrices for the random effects contain covariates that vary at lower levels than that of the corresponding random effects. That is, $Z^{(3)}_{ijk}$, the design matrix for $b^{(3)}_k$, will, in general, contain covariates that vary across level 2 and level 1 units.

To fix ideas, consider the example of a multi-center longitudinal clinical trial introduced earlier. A three-level model for the outcome is given by

$$Y_{ijk} = \beta_1 + \beta_2 t_{ijk} + \beta_3 (\text{Group}_{ij} \times t_{ijk}) + b^{(3)}_k + b^{(2)}_{jk} + \epsilon_{ijk},$$

where t_{ijk} denotes the time since baseline for the i^{th} observation on the j^{th} patient in the k^{th} clinic. In this model, $b^{(3)}_k$ is a random clinic effect and $b^{(2)}_{jk}$ is a random

patient effect. The inclusion of the former accounts for the clustering of patients within clinics, while the inclusion of the latter accounts for the positive correlation among the repeated measures on the same patient. Additional random effects, at both levels 2 and 3, can easily be incorporated in the model (e.g., random slopes for time, and possibly random effects for the $\text{Group}_{ij} \times t_{ijk}$ interaction).

The three-level model makes the following two assumptions about the different sources of variability:

1. The random effects $b_k^{(3)}$ are assumed to be independent across level 3 units, with mean zero and covariance, $\text{Cov}(b_k^{(3)}) = G^{(3)}$; similarly the random effects $b_{jk}^{(2)}$ are assumed to be independent across level 2 units, with mean zero and covariance, $\text{Cov}(b_{jk}^{(2)}) = G^{(2)}$. Random effects may be correlated within a given level, but not between levels.

2. The level 1 random components, ϵ_{ijk}, are assumed to be independent across level 1 units, with mean zero and variance, $\text{Var}(\epsilon_{ijk}) = \sigma^2$. In addition the ϵ_{ijk}'s are assumed to be independent of the random effects, $b_k^{(3)}$ and $b_{jk}^{(2)}$.

That is, the random effects at the same level are, in general, correlated within units at that level but not between units; random effects at different levels are assumed to be independent of each other and of the level 1 random components, ϵ_{ijk}. In principle, we can replace ϵ_{ijk} in (22.3) with $Z_{ijk}^{(1)} b_{ijk}^{(1)}$, where $b_{ijk}^{(1)}$ has mean zero and $\text{Cov}(b_{ijk}^{(1)}) = G^{(1)}$. This would allow for heterogeneity in the level 1 variability, with possible dependence of the level 1 variance on certain covariates. However, for the remainder of this discussion, we assume the simpler variance structure for the ϵ_{ijk}, with $\text{Var}(\epsilon_{ijk}) = \sigma^2$ (i.e., we assume $Z_{ijk}^{(1)} = 1$ for all i, j, and k).

In model (22.3) the regression parameters, β, are the fixed effects and describe the effects of covariates on the mean response (averaged over level 1, level 2, and level 3 units),

$$E\left(Y_{ijk}\right) = X_{ijk}\beta.$$

The three-level model given by (22.3) also describes the effects of covariates on the conditional mean response

$$E\left(Y_{ijk} \mid b_k^{(3)}\right) = X_{ijk}\beta + Z_{ijk}^{(3)} b_k^{(3)},$$

where the response is averaged over level 1 and level 2 units only, and the effects of covariates on the conditional mean response

$$E\left(Y_{ijk} \mid b_k^{(3)}, b_{jk}^{(2)}\right) = X_{ijk}\beta + Z_{ijk}^{(3)} b_k^{(3)} + Z_{ijk}^{(2)} b_{jk}^{(2)},$$

where the response is averaged over level 1 units only.

22.3.3 Estimation

The general specification of two- and three-level models that we have just described can be readily extended to more levels. The parameters of multilevel models are the fixed effects regression parameters, β, and the covariance (or variance) of the random effects at each level. Given estimates of the latter, predictions (empirical BLUPs) of the random effects at any level can also be obtained. For multilevel linear models, it is common to assume that the random components have multivariate normal distributions. For example, in the three-level model it is usually assumed that $b_k^{(3)} \sim N(0, G^{(3)})$, $b_{jk}^{(2)} \sim N(0, G^{(2)})$, and $\epsilon_{ijk} \sim N(0, \sigma^2)$. Given these distributional assumptions, maximum likelihood (ML) estimation of the multilevel model parameters is relatively straightforward.

The ML estimate of β is obtained from the generalized least squares (GLS) estimator. For the two-level model, the GLS estimator has the same form as that given in Chapter 4 (albeit, with the indices i and j reversed). For the three-level model, the GLS estimator of β also has a closed-form expression and is given by

$$\widehat{\beta} = \left\{ \sum_{k=1}^{n_3} \left(X_k' V_k^{-1} X_k \right) \right\}^{-1} \sum_{k=1}^{n_3} \left(X_k' V_k^{-1} Y_k \right),$$

where Y_k is a column vector, of length $\sum_{j=1}^{n_{2k}} n_{1jk}$, formed by stacking the responses for all second- and first-level units within the k^{th} cluster. Similarly, X_k is an $(\sum_{j=1}^{n_{2k}} n_{1jk}) \times p$ matrix formed by stacking the covariates for all second- and first-level units within the k^{th} cluster. Finally, V_k is the covariance among observations on first- and second-level units within the k^{th} cluster and has a random effects covariance structure, expressed as a function of $G^{(3)}, G^{(2)}$, and σ^2 (and the corresponding design matrices for the random effects).

The restricted maximum likelihood (REML) estimates of $G^{(3)}, G^{(2)}$, and σ^2 are obtained by maximizing the restricted log-likelihood with respect to $G^{(3)}, G^{(2)}$, and σ^2. In general, it is not possible to write down simple, closed-form expressions for the REML estimators of $G^{(3)}, G^{(2)}$, and σ^2; instead, estimates must be obtained using iterative techniques. Once the REML estimates of $G^{(3)}, G^{(2)}$, and σ^2 have been obtained, the estimate of $V_k(G^{(3)}, G^{(2)}, \sigma^2)$, say $V_k(\widehat{G}^{(3)}, \widehat{G}^{(2)}, \widehat{\sigma}^2)$, is substituted into the generalized least squares estimator of β to obtain the REML estimate of β. REML estimation for multilevel linear models has been implemented in many major statistical software packages (e.g., PROC MIXED in SAS, the `lme` function in the `nlme` package in R and S-Plus, and the `xtmixed` command in Stata) and in stand-alone programs that have been specifically tailored for multilevel modeling (e.g., MLwiN and HLM).

22.3.4 Case Studies

Next we illustrate the main ideas by conducting analyses of two- and three-level data. The first example analyzes two-level data on fetal weight from a developmental toxicity study of laboratory mice exposed to ethylene glycol (EG). The data on

Table 22.1 Descriptive statistics on fetal weight from the ethylene glycol (EG) experiment.

Dose				Weight (gm)	
(mg/kg)	$\sqrt{\text{Dose}/750}$	Dams	Fetuses	Mean	St. Deviation[a]
0	0	25	297	0.972	0.098
750	1	24	276	0.877	0.104
1500	1.4	22	229	0.764	0.107
3000	2	23	226	0.704	0.124

[a]Calculated ignoring clustering.

the weights of live fetuses, nested within litters, are from an experiment conducted through the National Toxicology Program (NTP) (Price et al., 1985). The second example analyzes three-level data from a cluster-randomized trial to determine the efficacy of school-based interventions to prevent tobacco use. The data on seventh grade children, nested within classrooms, nested within schools are from the *Television, School, and Family Smoking Prevention and Cessation Project* (TVSFP) (Flay et al., 1995, Hedeker et al., 1994).

Developmental Toxicity Study of Ethylene Glycol

Developmental toxicity studies of laboratory animals play a crucial role in the testing and regulation of chemicals and pharmaceutical compounds. Exposure to developmental toxicants typically causes a variety of adverse effects, such as fetal malformations and reduced fetal weight at term. In a typical developmental toxicity experiment, laboratory animals are assigned to increasing doses of a chemical or test substance. In this section we describe an analysis of data from a development toxicity study of ethylene glycol (EG). Ethylene glycol is a high-volume industrial chemical used in many applications. It is used as an antifreeze, as a solvent in the paint and plastics industries, and in the formulation of various types of inks. In a study of laboratory mice conducted through the National Toxicology Program (NTP), EG was administered at doses of 0, 750, 1500, or 3000 mg/kg/day to 94 pregnant mice (dams) over the period of major organogenesis, beginning just after implantation. (See Price et al., 1985, for additional details concerning the study design.) Following sacrifice, fetal weight and evidence of malformations were recorded for each live fetus. In our analysis of the data, we focus on the effects of dose on fetal weight; in Section 22.4 we present a complementary analysis that examines the effect of dose on fetal malformations.

Summary statistics (ignoring clustering in the data) for fetal weight for the 94 litters (composed of a total of 1028 live fetuses) are presented in Table 22.1. Fetal weight decreases monotonically with increasing dose, with the average weight ranging from

Table 22.2 Fixed and random effects estimates for the fetal weight data from the ethylene glycol (EG) experiment.

Variable	Estimate	SE	Z
Intercept	0.984	0.016	61.32
$\sqrt{\text{Dose}/750}$	−0.134	0.012	−10.85
Level 2 Variance:			
$\sigma_2^2(\times 100)$	0.726	0.119	
Level 1 Variance:			
$\sigma_1^2(\times 100)$	0.556	0.026	

0.972 (gm) in the control group to 0.704 (gm) in the group administered the highest dose. The decrease in fetal weight is not linear in increasing dose, but is approximately linear in increasing $\sqrt{\text{dose}}$.

The data on fetal weight from this experiment are clustered, with observations on the fetuses (level 1 units) nested within dams/litters (level 2 units). The litter sizes range from 1 to 16. Letting Y_{ij} denote the fetal weight of the i^{th} live fetus from the j^{th} litter, we considered the following model relating the fetal weight outcome to dose:

$$Y_{ij} = \beta_1 + \beta_2 \, \mathrm{d}_j + b_j + \epsilon_{ij},$$

where $\mathrm{d}_j = \sqrt{\text{Dose}_j/750}$ is the square-root transformed dose administered to the j^{th} dam. The random effect b_j is assumed to vary independently across litters, with $b_j \sim N(0, \sigma_2^2)$. The errors, ϵ_{ij}, are assumed to vary independently across fetuses (within a litter), with $\epsilon_{ij} \sim N(0, \sigma_1^2)$. Note that in a slight departure from the notation introduced previously, the first- and second-level variances are denoted by σ_1^2 and σ_2^2, respectively. This model assumes that fetuses within a cluster are exchangeable and the positive correlation among the fetal weights is accounted for by their sharing a common random effect, b_j. The degree of clustering in the data can be expressed in terms of the intra-cluster (or intra-litter) correlation

$$\rho = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

In Table 22.2 the results of fitting the model to the fetal weight data are presented. The REML estimate of the regression parameter for (transformed) dose indicates that the mean fetal weight decreases with increasing dose. The estimated decrease in weight, comparing the highest dose group to the control group, is 0.27 (or 2×-0.134, with 95% confidence interval: -0.316 to -0.220). Of note, we calculated both model-based and empirical (or "sandwich") standard errors and they were very similar,

suggesting that the simple random effect structure for the clustering of fetal weights is adequate. The estimate of the intra-cluster correlation, $\hat{\rho} = 0.57$, indicates that there are moderate litter effects.

Finally, to assess the adequacy of the linear dose–response trend, we considered a model that included a quadratic effect of (transformed) dose. Both Wald and likelihood ratio tests of the quadratic effect of dose indicated that the linear trend is adequate for these data (Wald $W^2 = 1.38$, with 1 df, $p > 0.20$; likelihood ratio $G^2 = 1.37$, with 1 df, $p > 0.20$).

Television School and Family Smoking Prevention and Cessation Project

Although smoking prevalence has declined among adults in recent decades, substantial numbers of young people begin to smoke and become addicted to tobacco. The *Television, School and Family Smoking Prevention and Cessation Project* (TVSFP) was a study designed to determine the efficacy of a school-based smoking prevention curriculum in conjunction with a television-based prevention program, in terms of preventing smoking onset and increasing smoking cessation (Flay et al., 1995). The study used a 2×2 factorial design, with four intervention conditions determined by the cross-classification of a school-based social-resistance curriculum (CC: coded 1 = yes, 0 = no) with a television-based prevention program (TV: coded 1 = yes. 0 = no). Randomization to one of the four intervention conditions was at the school level, while much of the intervention was delivered at the classroom level. The TVSFP study is described in greater detail in Flay et al. (1995).

The original study involved 6695 students in 47 schools in Southern California. Our analysis focuses on a subset of 1600 seventh-grade students from 135 classes in 28 schools in Los Angeles. The response variable, a tobacco and health knowledge scale (THKS), was administered before and after randomization of schools to one of the four intervention conditions. The scale assessed a student's knowledge of tobacco and health.

We considered a linear model for the post-intervention THKS score, with the baseline or pre-intervention THKS score as a covariate. This model for the adjusted change in THKS scores included the main effects of CC and TV and the CC $\times$ TV interaction. School and classroom effects were modeled by incorporating random effects at levels 3 and 2, respectively. Letting Y_{ijk} denote the post-intervention THKS score of the i^{th} student within the j^{th} classroom within the k^{th} school, our model is given by

$$Y_{ijk} = \beta_1 + \beta_2 \text{Pre-THKS} + \beta_3 \text{CC} + \beta_4 \text{TV} + \beta_5 \text{CC} \times \text{TV} + b_k^{(3)} + b_{jk}^{(2)} + \epsilon_{ijk},$$

where $\epsilon_{ijk} \sim N(0, \sigma_1^2), b_{jk}^{(2)} \sim N(0, \sigma_2^2)$, and $b_k^{(3)} \sim N(0, \sigma_3^2)$. Once again, in a slight departure from the notation introduced previously, the first-, second-, and third-level variances are denoted by σ_1^2, σ_2^2, and σ_3^2, respectively.

The results of fitting this model to the data are presented in Table 22.3. The REML estimates of the three sources of variability indicate that there is variability at both classroom and school levels, with almost twice as much variability among class-

Table 22.3 Fixed and random effects estimates for the THKS scores from the Television, School and Family Smoking Prevention and Cessation Project.

Variable	Estimate	SE	Z
Intercept	1.702	0.1254	13.57
Pre-Intervention THKS	0.305	0.0259	11.79
CC	0.641	0.1609	3.99
TV	0.182	0.1572	1.16
CC × TV	−0.331	0.2245	−1.47
Level 3 Variance:			
σ_3^2	0.039	0.0253	
Level 2 Variance:			
σ_2^2	0.065	0.0286	
Level 1 Variance:			
σ_1^2	1.602	0.0591	

rooms within a school as among schools. The correlation among the THKS scores for classmates (or children within the same classroom within the same school) is approximately 0.061 (or $\frac{0.039+0.065}{0.039+0.06+1.602}$), while the correlation among the THKS scores for children from different classrooms within the same school is approximately 0.023 (or $\frac{0.039}{0.039+0.06+1.602}$). The estimates of the fixed effects for the intervention conditions, when compared to their standard errors, indicate that neither the mass-media intervention (TV) nor its interaction with the social-resistance classroom curriculum (CC) have an impact on adjusted changes in the THKS scores from baseline. There is a significant effect of the social-resistance classroom curriculum, with children assigned to the social-resistance curriculum showing increased knowledge about tobacco and health. The estimate of the main effect of CC, in the model that excludes the CC × TV interaction, is 0.47 (SE = 0.113, $p < 0.0001$).

The intra-cluster correlations at both the school and classroom levels are relatively small. The reader might be tempted to regard this as an indication that the clustering in these data is inconsequential. However, such a conclusion would be erroneous. Although the intra-cluster correlations are relatively small, they have a substantial impact on inferences concerning the effects of the intervention conditions. To illustrate this, consider the following model for the adjusted changes in THKS scores:

$$Y_{ijk} = \beta_1 + \beta_2 \text{Pre-THKS} + \beta_3 \text{CC} + \beta_4 \text{TV} + \beta_5 \text{CC} \times \text{TV} + \epsilon_{ijk},$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$. This model ignores clustering in the data at the classroom and school levels; it is a standard linear regression model and assumes independent

Table 22.4 Fixed effects estimates from analysis that ignores clustering in the THKS scores from the Television, School and Family Smoking Prevention and Cessation Project.

Variable	Estimate	SE	Z
Intercept	1.661	0.0844	19.69
Pre-Intervention THKS	0.325	0.0258	12.58
CC	0.641	0.0921	6.95
TV	0.199	0.0900	2.21
CC × TV	−0.322	0.1302	−2.47

observations and homogeneous variance. The results of fitting this model to the THKS scores are presented in Table 22.4. The estimates of the fixed effects are similar to those reported in Table 22.3. However, the model-based standard errors (assuming no clustering) are misleadingly small for the randomized intervention effects and lead to substantively different conclusions about the effects of the intervention conditions. This highlights an important lesson: the impact of clustering depends on both the magnitude of the intra-cluster correlation and the cluster size. For the data from the TVSFP, the cluster sizes vary from 1 to 13 classrooms within a school and from 2 to 28 students within a classroom. With relatively large cluster sizes, even very modest intra-cluster correlation can have a discernible impact on inferences.

22.4 MULTILEVEL GENERALIZED LINEAR MODELS

So far the discussion of multilevel models has focused on linear models for a continuous response, where clustering was accounted for through the introduction of random effects at different levels. Next we briefly describe how multilevel modeling can be extended to discrete response data. These models can be thought of as multilevel generalized linear models, and they extend in a natural way the conceptual approach described in Section 22.3. However, they differ in terms of assumptions concerning the distribution of observations at level 1. The level 1 observations are no longer required to have a normal distribution; instead, they are assumed to have a distribution belonging to the exponential family (e.g., Bernoulli or Poisson). We focus on models for two- and three-level data; the generalizations to more levels follow directly.

22.4.1 Two-Level Generalized Linear Models

The basic premise of multilevel generalized linear models is that clustering among units can be thought of as arising from their sharing a set of random effects. For example, with two-level binary data, it is assumed that the clustering of level 1 units

(within level 2 units) can be accounted for by heterogeneity across level 2 clusters in a subset of the regression coefficients from a generalized linear model (e.g., a logistic regression model with randomly varying intercepts). Conditional on the random effects, the level 1 observations are assumed to be independent and with a distribution belonging to the exponential family (e.g., Bernoulli).

In our description of two-level generalized linear models we adopt the notation used earlier. Let Y_{ij} denote the response on the i^{th} level 1 unit in the j^{th} level 2 cluster; the response can be continuous, binary, or a count. Associated with each Y_{ij} is a $1 \times p$ (row) vector of covariates, X_{ij}. We can formulate two-level models for discrete (and continuous) outcomes, Y_{ij}, using the familiar three-part specification of generalized linear mixed effects models outlined in Chapter 14:

1. We assume that the conditional distribution of each Y_{ij}, given a vector of random effects b_j (and the covariates), belongs to the exponential family of distributions and that $\text{Var}(Y_{ij}|b_j) = v\{E(Y_{ij}|b_j)\}\,\phi$, where $v(\cdot)$ is a known variance function, a function of the conditional mean, $E(Y_{ij}|b_j)$, and ϕ is a scale or dispersion parameter. In addition, given the random effects b_j, it is assumed that the Y_{ij} are independent of one another.

2. The conditional mean of Y_{ij} is assumed to depend on fixed and random effects via the following linear predictor:

$$\eta_{ij} = X_{ij}\beta + Z_{ij}b_j,$$

   with

$$g\{E(Y_{ij}|b_j)\} = \eta_{ij} = X_{ij}\beta + Z_{ij}b_j$$

   for some known link function, $g(\cdot)$.

3. The random effects are assumed to have some probability distribution. In principle, any multivariate distribution can be assumed for the b_j; in practice, for computational convenience, the random effects are usually assumed to have a multivariate normal distribution, with zero mean and covariance matrix, G.

These three components completely specify a broad class of two-level generalized linear models for different types of responses. Next, to clarify the main ideas, we consider two examples of multilevel generalized linear models in greater detail.

Example 1: Two-Level Generalized Linear Model for Counts

Consider a study comparing cross-national rates of skin cancer and the factors (e.g., climate, economic and social factors, regional differences in diagnostic procedures) that influence variability in the rates of disease. Suppose that we have counts of the number of cases of skin cancer in a set of well-defined regions, indexed by i, within countries, indexed by j. Let Y_{ij} be a count of the number of individuals who develop skin cancer within the i^{th} region of the j^{th} country during a given period of time (e.g., 5 years). The resulting counts have a two-level structure with regional units at the

lower level (level 1 units) nested within countries (level 2 units). Usually the analysis of count data requires knowledge of the denominator, the population at risk. That is, the *rate* at which the disease occurs is of more direct interest than the corresponding count.

Counts are often modeled as Poisson random variables using a log link function. This motivates the following illustration of a two-level generalized linear model for Y_{ij} given by the three-part specification:

1. Conditional on a vector of random effects b_j, the Y_{ij} are assumed to be independent observations from a Poisson distribution, with $\text{Var}(Y_{ij}|b_j) = E(Y_{ij}|b_j)$, (i.e., $\phi = 1$).

2. The conditional mean of Y_{ij} depends on fixed and random effects via the following log link function,

$$\log\{E(Y_{ij}|b_j)\} = \log(T_{ij}) + X_{ij}\beta + Z_{ij}b_j,$$

   where T_{ij} is the population at risk in the i^{th} region of the j^{th} country and $\log(T_{ij})$ is an *offset.*

3. The random effects are assumed to have a multivariate normal distribution, with zero mean and covariance matrix G.

This is an example of a two-level log-linear model that assumes a linear relationship between the log rate of disease occurrence and the covariates.

Example 2: Two-Level Generalized Linear Model for Binary Responses

Consider a study of men with newly diagnosed prostate cancer. The study is designed to evaluate the factors that determine physician recommendations for surgery (radical prostatectomy) versus radiation therapy. In particular, it is of interest to determine the relative importance of patient factors (e.g., patient's age, level of prostate specific antigen) and physician factors (e.g., specialty training, years of experience) on physician recommendations for treatment. Many patients in the study seek the recommendation of the same physician. As a result patients (level 1 units) are nested within physicians (level 2 units). For each patient we have a binary outcome denoting the physician recommendation (surgery versus radiation therapy).

Let Y_{ij} be the binary response, taking values 0 and 1 (e.g., denoting surgery or radiation therapy) for the i^{th} patient of the j^{th} physician. An illustrative example of a two-level logistic model for Y_{ij} is given by the following three-part specification:

1. Conditional on a single random effect b_j, the Y_{ij} are independent and have a Bernoulli distribution, with $\text{Var}(Y_{ij}|b_j) = E(Y_{ij}|b_j)\{1 - E(Y_{ij}|b_j)\}$, (i.e., $\phi = 1$).

2. The conditional mean of Y_{ij} depends on fixed and random effects via the following linear predictor:

$$\eta_{ij} = X_{ij}\beta + b_j,$$

with

$$\log\left\{\frac{\Pr(Y_{ij}=1|b_j)}{\Pr(Y_{ij}=0|b_j)}\right\} = \eta_{ij} = X_{ij}\beta + b_j.$$

That is, the conditional mean of Y_{ij} is related to the linear predictor by a logit link function.

3. The single random effect b_j is assumed to have a univariate normal distribution, with zero mean and variance g_{11}.

In this example the model is a simple two-level logistic regression model with randomly varying intercepts. In principle, the linear predictor can include additional random effects. However, some caution must be exercised because there is usually not much information in binary data to estimate more than a single variance component unless the number of level 1 units is relatively large.

In Section 11.3 we discussed how the logistic regression model for binary data can also be developed through the notion of an underlying latent variable distribution. The same notion can be applied to multilevel models for binary responses. Suppose that L_{ij} is a latent (i.e., unobserved) continuous variable and that a positive response is observed only when L_{ij} exceeds some threshold. Consider the following two-level linear model for L_{ij},

$$L_{ij} = X_{ij}\beta + Z_{ij}b_j + \epsilon_{ij},$$

where the random effects are assumed to have a multivariate normal distribution, with mean zero and covariance matrix, G, and the ϵ_{ij} are assumed to have a standard logistic distribution, with mean zero and variance $\pi^2/3$. Without loss of generality, we can assume the threshold for categorizing L_{ij} is zero, with

$$Y_{ij} = 1 \text{ if } L_{ij} > 0,$$

$$Y_{ij} = 0 \text{ if } L_{ij} \leq 0.$$

Then the relationship between Y_{ij} and L_{ij} results in a logistic regression model for $\Pr(Y_{ij} = 1|b_j)$. That is, the two-level linear model for L_{ij} with standard logistic errors,

$$L_{ij} = X_{ij}\beta + Z_{ij}b_j + \epsilon_{ij},$$

implies the two-level logistic regression model for Y_i,

$$\log\left\{\frac{\Pr(Y_{ij}=1|b_j)}{\Pr(Y_{ij}=0|b_j)}\right\} = X_{ij}\beta + Z_{ij}b_j.$$

Using the notion of an underlying latent variable distribution, we can then compare the magnitudes of the between-cluster and within-cluster sources of variability of the L_{ij}. For example, in a two-level logistic regression model with a single random effect b_j (with variance g_{11}), the relative magnitudes of the between-cluster and within-cluster sources of variability can be summarized in terms of the intra-cluster correlation

$$\rho = \text{Corr}(L_{ij}, L_{i'j}) = \frac{g_{11}}{g_{11} + \pi^2/3}; \quad (\text{for } i \neq i').$$

Note that ρ is the marginal correlation among the latent variables, L_{ij}; it is not the marginal correlation among the binary variables, Y_{ij}.

Although in both of the examples of two-level generalized linear models we have chosen canonical link functions to relate the conditional mean of Y_{ij} to η_{ij}, in principle, any suitable link function can be selected. These two examples are intended to be purely illustrative. They demonstrate how the choices of the three components might differ according to the type of response variable.

So far our discussion of two-level models has closely paralleled the description of the generalized linear mixed effects model given in Chapter 14 (albeit, with the indices i and j reversed). In Chapter 14 we focused on two-level models where measurement occasions are the level 1 units and subjects are the level 2 units; this is a special case of two-level data. However, the methods in Chapter 14 can be applied more broadly to different types of two-level data and also extend naturally to more than two levels.

22.4.2 Three-Level Generalized Linear Models

The extension of two-level generalized linear models to three or more levels is straightforward and follows from the previous sections. With three-level data the variability of the response is accounted for by the introduction of random effects at both levels 2 and 3. For a three-level data structure, we adopt the same notation as in Section 22.3, except that the response can be continuous, binary, or a count. Let Y_{ijk} denote the response on the i^{th} level 1 unit within the j^{th} level 2 cluster within the k^{th} level 3 cluster. Associated with each Y_{ijk} is a $1 \times p$ (row) vector of covariates, X_{ijk}, with the covariates defined at different levels. A three-level generalized linear model for Y_{ijk} is given by the following three part specification:

1. We assume that the conditional distribution of each Y_{ijk}, given vectors of random effects, $b_k^{(3)}$ and $b_{jk}^{(2)}$ (defined at levels 3 and 2, respectively), belongs to the exponential family of distributions and that $\text{Var}(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)}) = v\{E(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)})\}\,\phi$, where $v(\cdot)$ is a known variance function, a function of the conditional mean, $E(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)})$, and ϕ is a scale or dispersion parameter. In addition, given $b_k^{(3)}$ and $b_{jk}^{(2)}$, it is assumed that the Y_{ijk} are independent of one another.

2. The conditional mean of Y_{ijk} is assumed to depend on fixed and random effects via the following linear predictor:

$$\eta_{ijk} = X_{ijk}\beta + Z_{ijk}^{(3)} b_k^{(3)} + Z_{ijk}^{(2)} b_{jk}^{(2)},$$

with

$$g\{E(Y_{ijk}|b_k^{(3)}, b_{jk}^{(2)})\} = \eta_{ijk} = X_{ijk}\beta + Z_{ijk}^{(3)} b_k^{(3)} + Z_{ijk}^{(2)} b_{jk}^{(2)},$$

for some known link function, $g(\cdot)$.

3. The random effects are assumed to have multivariate normal distributions, with mean zero and covariance matrices, $\text{Cov}(b_k^{(3)}) = G^{(3)}$ and $\text{Cov}(b_{jk}^{(2)}) = G^{(2)}$. Although random effects may be correlated within a given level, random effects at different levels are assumed to be independent of each other.

These three components completely specify a broad class of three-level generalized linear models.

22.4.3 Estimation

The multilevel generalized linear models described in the previous section fully specify the joint distribution of the responses at level 1 and the random effects at all higher levels. As a result we can base estimation and inference on the likelihood function. However, unlike the case with a continuous response assumed to have a normal distribution, maximum likelihood (ML) estimation for multilevel generalized linear models is not straightforward and will, in general, require numerical quadrature.

For example, for three-level data, inference about β, $G^{(2)}$ and $G^{(3)}$ is based on the marginal likelihood function. The marginal likelihood can be expressed as the product of the probability density functions, $f(y_k)$, for the level 3 units. Specifically, the marginal log-likelihood function is given by the following sum:

$$\sum_{k=1}^{n_3} \log f(y_k),$$

where $f(y_k)$ can be obtained by recognizing that observations on level 2 units (within level 3 units) are conditionally independent of one another given the level 3 random effects, $b^{(3)}$; similarly, observations on level 1 units (within level 2 units) are conditionally independent of one another given the level 3 and level 2 random effects, $b^{(3)}$ and $b^{(2)}$. The k^{th} level 3 unit's contribution to the likelihood function is

$$f(y_k) = \int \prod_{j=1}^{n_{2k}} \left\{ \int \prod_{i=1}^{n_{1jk}} f(y_{ijk}|b_k^{(3)}, b_{jk}^{(2)},) f(b_{jk}^{(2)})\, db_{jk}^{(2)} \right\} f(b_k^{(3)})\, db_k^{(3)},$$

where $f(b_{jk}^{(2)})$ and $f(b_k^{(3)})$ denote the multivariate normal distributions for the random effects at levels 2 and 3, respectively. The ML estimates of β, $G^{(2)}$ and $G^{(3)}$ are simply those values of β, $G^{(2)}$ and $G^{(3)}$ that maximize the marginal log-likelihood function.

The primary reason for displaying the expression given above is to highlight that multivariate integrals must be evaluated to compute the marginal log-likelihood. That is, the log-likelihood function is obtained by integrating out or averaging over the distribution of the random effects, $b_{jk}^{(2)}$ and $b_k^{(3)}$. Because integrals (denoting averaging over the distribution of the random effects) appear in the log-likelihood function, there are no simple, closed-form solutions. Instead, numerical integration techniques, for instance, Gaussian quadrature, are required for maximizing the log-likelihood function. ML estimation, using Gaussian quadrature, for two-level and higher-level generalized linear models is implemented in some of the major statistical software

packages (e.g., PROC GLIMMIX in SAS, the `glmer` function in the `lme4` package in R, and the `xtmelogit` and `xtmepoisson` commands in Stata). Various alternative approximations to ML estimation for the extensions to three or more levels are implemented in more specialized, stand-alone programs that have been specifically developed for multilevel modeling (e.g., MLwiN and HLM).

22.4.4 Case Studies

We illustrate the main ideas underlying multilevel modeling by conducting analyses of two-level data where the observations at level 1 are counts and binary outcomes. The first example analyzes two-level count data from a study of malignant melanoma mortality and ultraviolet (UV) radiation exposure. The second example analyzes two-level data on fetal malformations, a binary outcome, from the developmental toxicity study of ethylene glycol (EG) described in Section 22.3. For the latter, we present a traditional multilevel analysis of the fetal malformation data and contrast the results with those obtained from a marginal model that accounts for clustering in the data in a different way.

Malignant Melanoma Mortality and Ultraviolet Light Exposure

In a study of the effects of ultraviolet (UV) light exposure on malignant melanoma mortality (Langford et al., 1998), counts of the number of deaths due to malignant melanoma were recorded for males of all ages in the United Kingdom. The counts of the number of deaths between 1975 and 1980 were aggregated over areas that correspond to counties or shires; hereafter referred to as counties. Data were collected on 70 counties nested within 11 regions of the United Kingdom. The resulting data structure is multilevel, with counties at level 1 (indexed by i) nested within regions at level 2 (indexed by j). The main predictor of interest is exposure to ultraviolet light in the B band (UVB). An index of UVB dose reaching the earth's surface was calculated for each county. The mean UVB index in the United Kingdom was 10.9, with a standard deviation of 1.5.

Let Y_{ij} denote the count of the number of deaths in the i^{th} county in the j^{th} region due to malignant melanoma. Within a given region, we assume Y_{ij} has a Poisson distribution to account for level 1 variation in the counts. Variation in the counts across regions is accounted for by the inclusion of a random region effect, b_j. That is, conditional on a random region effect b_j, the counts are assumed to have a Poisson distribution with conditional mean related to UVB dose via a log link function,

$$\log\{E(Y_{ij}|b_j)\} = \log(T_{ij}) + \beta_1 + \beta_2 \mathrm{UVB}_{ij} + b_j,$$

where UVB_{ij} is the UVB index (centered at the mean UVB index in the United Kingdom) in the i^{th} county of the j^{th} region. For each county, T_{ij} is the number of deaths that would be expected were U.K. national age- and gender-specific death rates to apply to the population of the county. Note that T_{ij} is known and $\log(T_{ij})$ is an *offset* in this model. The ratio of the observed number of deaths to the expected

Table 22.5 Fixed and random effects estimates for the malignant melanoma mortality data for males in the United Kingdom.

Variable	Estimate	SE	Z
Intercept	−0.0365	0.0352	−1.04
UVB	0.1301	0.0279	4.67
Level 2 Variance:			
$\sigma^2 \times 100$	0.6222	0.5087	

Note: ML estimation is based on 50-point adaptive Gaussian quadrature.

number of deaths, Y/T, is referred to as the *standardized mortality ratio* (SMR) in each county. Our model assumes a linear relationship between the log SMR due to malignant melanoma and county level UV radiation exposure. Finally, we assume the random region effect has a normal distribution, $b_j \sim N(0, \sigma^2)$.

The results of fitting this model to the UK malignant melanoma mortality data, using maximum likelihood estimation, are presented in Table 22.5. There is a significant positive relationship between the SMRs and exposure to UVB. Recall that the standard deviation of UVB in the United Kingdom is 1.5. Therefore the estimated effect of UVB dose indicates that the SMR is approximately 1.5 times larger (or $e^{3\times 0.13}$, with 95% confidence interval: 1.25 to 1.74) when comparing a county with UVB index 1 standard deviation above the UK average to a county with UVB index 1 standard deviation below. Finally, in interpreting these results, it should be remembered that the UVB covariate used here is simply an index of exposure for each county; it is the potential, not the actual, UVB dose experienced by the population of residents in each county.

Developmental Toxicity Study of Ethylene Glycol

Next we consider a two-level logistic regression model for binary data on fetal malformations from the developmental toxicity study of ethylene glycol (EG). Recall that in this study, EG was administered (0, 750, 1500, or 3000 mg/kg/day) to 94 pregnant mice (dams) over the period of major organogenesis, beginning just after implantation (Price et al., 1985). Following sacrifice, each live fetus was examined for evidence of malformations, recorded as present or absent. The primary question of scientific interest is the effect of dose on fetal malformations.

Summary statistics (ignoring clustering in the data) for fetal malformations for the 94 litters (composed of a total of 1028 live fetuses) are presented in Table 22.6. The percentage of fetal malformations increases monotonically with increasing dose, with less than 1% in the control group and almost 60% in the group administered the highest dose (3000 mg/kg).

Table 22.6 Descriptive statistics on fetal malformations from the ethylene glycol (EG) experiment.

Dose			Fetal Malformations	
(mg/kg)	Dams	Fetuses	Number	Percentage
0	25	297	1	0.34
750	24	276	26	9.42
1500	22	229	89	38.86
3000	23	226	129	57.08

Table 22.7 Fixed and random effects estimates for the fetal malformation data from the ethylene glycol (EG) experiment.

Variable	Estimate	SE	Z
Intercept	−4.360	0.440	−9.92
Dose / 750	1.336	0.166	8.06
Level 2 Variance:			
σ_b^2	2.517	0.685	

Note: ML estimation is based on 50-point adaptive Gaussian quadrature.

Letting $Y_{ij} = 1$ denote the presence of fetal malformations in the i^{th} live fetus from the j^{th} litter (and $Y_{ij} = 0$ otherwise), we considered the following logistic model relating the log odds of fetal malformations to a linear effect of dose:

$$\text{logit}\{E(Y_{ij}|b_j)\} = \beta_1 + \beta_2 \, \text{d}_j + b_j,$$

where $\text{d}_j = \text{Dose}_j/750$ denotes the dose (in units of 750 mg/kg) administered to the j^{th} dam (cluster). The random effect b_j is assumed to vary independently across litters, with $b_j \sim N(0, \sigma_b^2)$. This model assumes that fetuses within a litter are exchangeable and the positive association among the fetal malformation outcomes is accounted for by their sharing a common random effect, b_j.

The results of fitting the model to the fetal malformation data, using maximum likelihood estimation, are presented in Table 22.7. The estimated regression parameter for dose indicates that the log odds of malformation increases with increasing dose. In particular, the odds ratio for malformation, comparing the highest dose group

Table 22.8 Estimates of regression parameters from marginal model for the fetal malformation data from the ethylene glycol (EG) experiment.

Variable	Estimate	SE	Z
Intercept	−3.190	0.220	−14.53
Dose / 750	0.960	0.099	9.66
Log Odds Ratio	1.447	0.221	6.56

to the control group, is 209.2 (or $e^{4\times 1.336}$, with 95% confidence interval: 56.1 to 779.9). This provides overwhelming evidence of the increased risk of malformations at the highest dose of EG. The odds ratio for malformations, comparing the lowest dose group to the control group, is 3.80 (or $e^{1.336}$, with 95% confidence interval: 2.75 to 5.26). Finally, the estimate of σ_b^2 indicates that there are moderate litter effects, with heterogeneity across dams in the underlying risk of producing fetuses with malformations. For example, in the control group, 95% of dams have a risk of producing fetuses with malformations between 0% and 22%. Alternatively, if we appeal to the notion of a latent variable distribution and assume an underlying two-level linear model for the latent variable with standard logistic errors, the estimated intra-cluster correlation is

$$\widehat{\rho} = \frac{\widehat{\sigma}_b^2}{\widehat{\sigma}_b^2 + \pi^2/3} = \frac{2.517}{2.517 + 3.290} = 0.43.$$

For illustrative purposes we also consider a marginal logistic regression model relating the log odds of fetal malformations to a linear effect of dose

$$\text{logit}\{E(Y_{ij})\} = \beta_1 + \beta_2 \, \text{d}_j,$$

and account for the intra-litter association by a common log odds ratio,

$$\log\{\text{OR}(Y_{ij}, Y_{ik})\} = \log\left\{\frac{\Pr(Y_{ij}=1, Y_{ik}=1)\Pr(Y_{ij}=0, Y_{ik}=0)}{\Pr(Y_{ij}=1, Y_{ik}=0)\Pr(Y_{ij}=0, Y_{ik}=1)}\right\}.$$

This can be thought of as a marginal model analogue of the random intercepts model for the within-litter association. The results of fitting this marginal model, using GEE methods, are presented in Table 22.8. The estimate of the effect of dose indicates that the odds ratio for malformations, comparing the highest dose group to the control group, is 46.4 (or $e^{4\times 0.960}$, with 95% confidence interval: 21.3 to 101.2). These results also provide strong evidence of the increased risk of malformations at the highest dose of EG. The within-litter odds ratio of 4.26 (or $e^{1.447}$) indicates that there is clustering in the fetal malformations data.

Note that the estimated effect of dose in the marginal model is discernibly smaller than that reported in Table 22.7. This should not be too surprising given the important

distinctions between the regression parameters in marginal models and generalized linear mixed effects models that were highlighted in Chapter 16. The regression parameters for dose in the two models have quite different interpretations. In the logistic regression model with a random litter effect, β_2 describes the change in the risk of producing fetuses with malformations for *any given dam*; this change in the risk for a single-unit change in dose depends on b_j, a specific dam's random effect or underlying propensity for producing fetuses with malformations. In the marginal model, β_2 describes changes in the prevalence of fetal malformations when subpopulations of dams exposed to different doses of EG are compared. Although both models account for clustering in the data, the targets of inference are different and the two analyses address distinct scientific questions.

22.5 SUMMARY

In previous sections we described models for data with a hierarchical structure, where lower-level units are nested within higher-level units. The dominant approach for modeling such data is regression models where random effects are introduced at different levels. A central feature of multilevel modeling is the incorporation of covariates that can be measured at any level of the hierarchy, thereby allowing the effects at each level to be disentangled. By combining covariates that have been measured at different levels within a single regression model, their relative importance can be determined. For example, multilevel models can address questions about the effects of individuals' characteristics (e.g., disease severity) while adjusting for their context (e.g., being treated at a large university teaching hospital versus a rural clinic).

Multilevel data can be challenging to analyze for at least two main reasons. First, the covariates can be measured at different levels, and the same covariate can operate at many different levels. As a result somewhat greater care is required in the interpretation of regression parameters in multilevel models. It is not always transparent how best to combine covariates measured at different levels within a single model so that the regression parameters have useful interpretations. In our brief overview of multilevel models, we have not touched on this important topic; for more information, the interested reader is directed to the references at the end of this chapter.

The second challenge in the analysis of multilevel data is how best to account for the clustering that can arise at different levels of the hierarchy. In the multilevel modeling literature the dominant approach for accounting for the intra-cluster correlations at different levels is via the introduction of random effects at different levels. This gives rise to mixed effects models that can be extended in a very natural way to any number of levels of clustering in the data. For linear models, this is certainly a very natural way to account for clustering. However, for generalized linear models for discrete data, it does raise subtle issues concerning the interpretation of the fixed effect regression parameters and questions about what is the relevant target of inference. The same issues that were given an airing in Chapter 16 apply equally to multilevel models for discrete data. These issues were highlighted in our analyses of the two-level clustered data on fetal malformations in Section 22.4, where the

estimated effect of dose was discernibly different depending on how the clustering was accounted for. The estimates of the effect of dose from the marginal and random effects logistic regression models differed because the corresponding regression parameters have distinct interpretations and address somewhat different scientific questions. In general, the fixed effects parameters in a two-level model for discrete data represent changes in the (transformed) mean response, for a single-unit change in the corresponding covariate, *for any given level 2 unit*. In Chapters 14 and 16, in the context of longitudinal data, these regression coefficients were referred to as "subject-specific"; here they are "cluster-specific" and describe covariate effects for an individual cluster. In contrast, the regression parameters in a marginal model represent changes in the (transformed) mean response when sub-populations defined by different values of the corresponding covariate are compared. The regression parameters in marginal models address the dependence of the population-averaged response (where averaging is over all possible units in the hierarchy) on the covariates. These regression parameters do not have any direct interpretation for an individual cluster when there is heterogeneity across clusters.

Although much of the multilevel literature on the analysis of discrete data is dominated by the use of generalized linear mixed effects models, we note that marginal models can also be used to account for clustering at different levels. All the issues discussed in Chapter 16 for two-level longitudinal data apply equally to two-level and higher-level data more broadly defined. In general, the choice between the two classes of models should not be driven by the availability of software for multilevel modeling but on the basis of careful thought about the questions of scientific interest.

22.6 FURTHER READING

There is an extensive literature on multilevel models that appears in the statistical, psychometric, and educational literature. A comprehensive description of multilevel models, and their application to a wide range of problems, can be found in the books by Raudenbush and Bryk (2002), Longford (1993), and Goldstein (2003). For readers who find the level of mathematical difficulty in these books too challenging, the books by Hox (2002), Kreft and De Leeuw (1998), and Snijders and Bosker (1999) provide a more introductory and accessible presentation of similar topics targeted at empirical researchers. An engaging and non-technical introduction to multilevel modeling can be found in the excellent text by Gelman and Hill (2007).

For illustrations of the application of multilevel models in the biomedical and health sciences, we recommend the edited volume of articles in Leyland and Goldstein (2001) and the review articles by Sullivan et al. (1999), Goldstein et al. (2002), and Subramanian et al. (2003).

Bibliographic Notes

Although most of the statistical literature on marginal models has focused on two-level data, Qaqish and Liang (1992) discuss the use of marginal models for multilevel binary data, with multiple levels of nesting.

Daniels and Gatsonis (1999) describe multilevel modeling in a Bayesian framework; also see Lindley and Smith (1972), Zeger and Karim (1991), Browne et al. (2002), Carlin and Louis (2000), and Chapters 15 and 16 of Gelman et al. (2003). Bayesian methods for multilevel modeling can be implemented using the publicly available software WinBUGS (Spiegelhalter et al., 1999).