

Hierarchical Models

Garrett M. Fitzmaurice

Laboratory for Psychiatric Biostatistics, McLean Hospital
Department of Biostatistics, Harvard School of Public Health

Outline

- Background
- Examples
- Regression Models for Hierarchical Data
- Case Studies
- Summary and Concluding Remarks

Background: Hierarchical Data

Hierarchical or multilevel data arise when there is a *clustered/grouped* structure to the data.

Data of this kind frequently arise in the social, behavioral, and 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 hierarchical/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.

Clustering in hierarchical data can be due to a naturally occurring hierarchy in the target population or a consequence of study design (or sometimes both).

Examples of naturally occurring data hierarchies:

Studies of nuclear families: observations on the mother, father, and children (level 1 units) nested within families (level 2 units).

Studies of health services/outcomes: observations on patients (level 1 units) nested within clinics (level 2 units).

Studies of education: observations on children (level 1 units) nested within classrooms (level 2 units).

Note: Naturally occurring hierarchical data structures can have more than two levels, e.g., children (level 1 units) nested within classrooms (level 2 units), nested within schools (level 3 units).

Examples of clustering as consequence of study design:

Longitudinal Studies: the clusters are composed of the repeated measurements obtained from a single individual at different occasions. In longitudinal studies the level 1 units are the repeated occasions of measurement and the level 2 units are the subjects.

Cluster-Randomized Clinical Trials: Groups (level 2 units) of individuals (level 1 units), rather than the individuals themselves, are randomly assigned to different treatments or interventions.

Complex Sample Surveys: Many national surveys use multi-stage sampling.

For example, in 1st stage, “primary sampling units” (PSUs) are defined based on counties in the United States. A first-stage random sample of PSUs are selected. In 2nd stage, within each selected PSU, a random sample of census blocks are selected. In 3rd stage, within selected census blocks, a random sample of households are selected.

Resulting data can be regarded as hierarchical, with households being the level 1 units, area segments the level 2 units, and counties the level 3 units.

Finally, clustering can be due to both study design and naturally occurring hierarchies in the target population.

Example: Clinical trials are often conducted in many different centers to ensure sufficient numbers of patients and/or to assess the effectiveness of the treatment in different settings.

Observations from a multi-center longitudinal clinical trial can be regarded as hierarchical data with 3 levels, with repeated measurement occasions (level 1 units) nested within subjects (level 2 units) nested within clinics (level 3 units).

Distinctive Feature of Hierarchical Data

Distinctive feature of hierarchical data is that they are *clustered*.

A consequence of this clustering is that measurement on units within a cluster are more similar than measurements on units in different clusters.

For example, two children selected at random from the same family are expected to respond more similarly than two children randomly selected from different families.

The clustering can be expressed in terms of correlation among the measurements on units within the same cluster.

Statistical models for hierarchical data must account for the intra-cluster correlation at each level; failure to do so can result in misleading inferences.

Models for Hierarchical Data

Focus mainly on linear regression models for hierarchical data.

Basis of dominant approaches for modelling hierarchical data: account for clustering via introduction of *random effects* at different levels in hierarchy.

In addition, hierarchical models permit estimation of the effects of covariates, measured at any levels of the hierarchy, on the outcome.

Note: Although response is obtained on the lowest level (or level 1) units, covariate information can be measured at any level.

Combining covariates measured at different levels of the hierarchy within a single regression model is central to hierarchical modelling.

Two-Level Linear Models

Notation: 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 $j = 1, \dots, n_2$) is composed of n_{1j} level 1 units.

Example: 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.

Clinics are the level 2 units ($j = 1, \dots, 20$) and patients are the level 1 units ($i = 1, \dots, n_{1j}$), 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).

Let Y_{ij} denote the response on the i^{th} level 1 unit within the j^{th} level 2 cluster.

Associated with each Y_{ij} is a (row) vector of covariates, X_{ij} .

These can include covariates defined at each of the two levels and can also include so-called “compositional” or “contextual” covariates, 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 regression model relating the mean response to the covariates:

$$E(Y_{ij}) = X_{ij}\beta = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp}. \quad (1)$$

The model given by (1) specifies how the mean response depends on covariates, where the covariates can be defined at level 2 and/or level 1.

A hierarchical model accounts for the variability in Y_{ij} , around its mean, by allowing for random variation across both level 1 and level 2 units.

Hierarchical models assume 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

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_j + e_{ij}, \quad (2)$$

where Z_{ij} is a design matrix for the random effects at level 2, formed from a subset of the appropriate components 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, e_{ij} , are also assumed to be independent across level 1 units, with mean zero and variance, $\text{Var}(e_{ij}) = \sigma^2$.

In addition, the e_{ij} 's are assumed to be independent of the b_j 's, with $\text{Cov}(e_{ij}, b_j) = 0$.

That is, the level 1 units are assumed to be conditionally independent given the level 2 random effects (and the covariates).

Simple Illustration:

Consider the following two-level model with a single random effect that varies across level 2 units:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + b_j + e_{ij}.$$

Here $Z_{ij} = 1$ for all i and j .

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 (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.

The two-level model given by (2) can also be written in terms of two models, one for each level of the hierarchy.

Level 1 model:

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

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

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 (2)

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

where $X_{ij} = Z_{ij}A_j$.

An advantage of specifying a hierarchical model in stages 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.

Key Points: The two-level linear model given by (2) accounts for the clustering of the level 1 units by incorporating random effects at level 2.

Model explicitly distinguishes two main sources of variation in the response: (a) variation across level 2 units and (b) 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 the degree of clustering.

Simple Illustration:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + b_j + e_{ij},$$

where e_{ij} are assumed to be independent across level 1 units, with mean zero and variance, $\text{Var}(e_{ij}) = \sigma_e^2$; b_j are assumed to vary independently across level 2 units, with mean zero and variance, $\text{Var}(b_j) = \sigma_b^2$.

Then, the correlation (or clustering) for a pair of level 1 units (within a level 2 unit) is given by:

$$\text{Corr}(Y_{ij}, Y_{i',j}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}.$$

The larger the variance of the level 2 random effect (σ_b^2), relative to the level 1 variability (σ_e^2), the greater the degree of clustering (or correlation).

Finally, the two-level model given by (2) can be extended in a natural way to three or more levels.

Clustering in three or higher level data is accounted for via the introduction of *random effects* at each of the different levels in the hierarchy.

Conceptually, no more complicated than in the two-level model.

Estimation of Parameters in Hierarchical Models

Parameters of hierarchical models are the fixed effects, β , and the covariance (or variance) of the random effects at each level.

For hierarchical linear models, it is common to assume random components have multivariate normal distributions.

For example, in the two-level model, it is usually assumed that $b_j \sim N(0, G)$, and $e_{ij} \sim N(0, \sigma^2)$.

Given these distributional assumptions, maximum likelihood (ML) estimation of the hierarchical model parameters is relatively straightforward.

The ML estimator of β is the generalized least squares (GLS) estimator.

For the two-level model, the GLS estimator of β has a closed-form expression:

$$\hat{\beta} = \left\{ \sum_{j=1}^{n_2} (X_j' V_j^{-1} X_j) \right\}^{-1} \sum_{j=1}^{n_2} (X_j' V_j^{-1} Y_j),$$

where Y_j is a column vector, of length n_{1j} , formed by stacking responses for all level 1 units within j^{th} cluster; X_j is a matrix formed in a similar way.

Finally, V_j is the covariance among observations on first-level units within the j^{th} cluster and has a random effects covariance structure, expressed as a function of G , and σ^2 .

Restricted maximum likelihood (REML) estimates of G and σ^2 are obtained by maximizing the restricted log-likelihood with respect to G and σ^2 .

In general, it is not possible to write down simple, closed-form expressions for the REML (or ML) estimators of G and σ^2 ; instead, estimates must be obtained using iterative techniques.

Once the REML (or ML) estimates of G and σ^2 have been obtained, the estimate of $V_j(G, \sigma^2)$, say $V_j(\hat{G}, \hat{\sigma}^2)$, is substituted into the generalized least squares estimator of β to obtain the REML (or ML) estimate of β .

REML estimation for hierarchical linear models has been implemented in many major statistical software packages (e.g., PROC MIXED in SAS and the *lme* function in S-PLUS) and in stand-alone programs that have been specifically tailored for hierarchical modelling (e.g., MLwiN and HLM).

Case Study 1: Developmental Toxicity Study of Ethylene Glycol

Developmental toxicity studies of laboratory animals play a crucial role in the testing and regulation of chemicals.

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.

Consider an analysis of data from a development toxicity study of ethylene glycol (EG).

Ethylene glycol 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) beginning just after implantation.

Following sacrifice, fetal weight and evidence of malformations were recorded for each live fetus.

In our analysis, we focus on the effects of dose on fetal weight.

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 1.

Fetal weight decreases monotonically with increasing dose, with the average weight ranging from 0.97 (gm) in the control group to 0.70 (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 $\sqrt{\text{dose}}$.

Table 1: Descriptive statistics on fetal weight.

Dose		Weight (gm)			
(mg/kg)	$\sqrt{\text{Dose}/750}$	Dams	Fetuses	Mean	St. Deviation [†]
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

[†]Calculated ignoring clustering.

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:

$$Y_{ij} = \beta_0 + \beta_1 d_j + b_j + e_{ij},$$

where $d_j = \sqrt{\text{Dose}_j/750}$ 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, e_{ij} , are assumed to vary independently across fetuses (within a litter), with $e_{ij} \sim N(0, \sigma_1^2)$.

In this model, the clustering or correlation among the fetal weights within a litter 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

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

Table 2: Fixed and random effects estimates for the fetal weight data.

Parameter	Estimate	SE	Z
Intercept	0.984	0.016	61.32
$\sqrt{\text{Dose}/750}$	-0.134	0.012	-10.85
$\sigma_2^2(\times 100)$	0.726	0.119	6.11
$\sigma_1^2(\times 100)$	0.556	0.026	21.55

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 highest dose group to control group, is 0.27 (or 2×-0.134 , 95% confidence interval: -0.316 to -0.220).

The estimate of the intra-cluster correlation, $\hat{\rho} = 0.57$, indicates that there are moderate litter effects.

Finally, adequacy of the linear dose–response trend assessed by considering model with quadratic effect of (transformed) dose.

Both Wald and likelihood ratio tests of the quadratic effect indicated that linear trend is adequate (Wald $W^2 = 1.38$, with 1 df, $p > 0.20$; likelihood ratio $G^2 = 1.37$, with 1 df, $p > 0.20$).

Case Study 2: Television School and Family Smoking Prevention and Cessation Project (TVFSP)

Although smoking prevalence has declined among adults, substantial numbers of young people begin to smoke and become addicted to tobacco.

TVFSP designed to determine efficacy of school-based smoking prevention curriculum in conjunction with television-based prevention program.

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 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.

Consider linear model for the post-intervention THKS score, with the baseline or pre-intervention THKS score as a covariate.

Model the adjusted change in THKS scores as function of main effects of CC and TV and the $CC \times TV$ interaction.

School and classroom effects modelled by incorporating random effects at levels 3 and 2, respectively (level 1 units are the children).

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

$$Y_{ijk} = \beta_0 + \beta_1 \text{Pre-THKS} + \beta_2 \text{CC} + \beta_3 \text{TV} + \beta_4 \text{CC} \times \text{TV} + b_k^{(3)} + b_{jk}^{(2)} + e_{ijk},$$

where $e_{ijk} \sim N(0, \sigma_1^2)$, $b_{jk}^{(2)} \sim N(0, \sigma_2^2)$, and $b_k^{(3)} \sim N(0, \sigma_3^2)$.

Table 3: Fixed effects estimates for the THKS scores.

Parameter	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 \times TV	-0.331	0.2245	-1.47

Table 4: Random effects estimates for the THKS scores.

Parameter	Estimate	SE	Z
Level 3 Variance:			
σ_3^2	0.039	0.0253	1.52
Level 2 Variance:			
σ_2^2	0.065	0.0286	2.26
Level 1 Variance:			
σ_1^2	1.602	0.0591	27.10

Consider REML estimates of the three sources of variability.

Comparing their relative magnitudes, there is variability at both classroom and school levels, with almost twice as much variability among classrooms within a school as among schools themselves.

Correlation among THKS scores for classmates (or children within same classroom within same school) is approximately 0.061 (or $\frac{0.039+0.065}{0.039+0.06+1.602}$).

Correlation among THKS scores for children from different classrooms within same school is approximately 0.023 (or $\frac{0.039}{0.039+0.06+1.602}$).

Next, consider REML estimates of fixed effects for the interventions.

When compared to their SEs, indicate that neither mass-media intervention (TV) nor its interaction with social-resistance classroom curriculum (CC) have an impact on adjusted changes in 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 \times 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.

It is very tempting to regard this as an indication that the clustering in these data is inconsequential.

However, such a conclusion would be erroneous.

Although intra-cluster correlations are relatively small, they have an impact on inferences concerning the effects of the intervention conditions.

To illustrate this, consider analysis that ignores clustering in the data:

$$Y_{ijk} = \beta_0 + \beta_1 \text{Pre-THKS} + \beta_2 \text{CC} + \beta_3 \text{TV} + \beta_4 \text{CC} \times \text{TV} + e_{ijk},$$

The results of fitting this model to the THKS scores are presented in Table 5 and the estimates of the fixed effects are similar to those reported in Table 3.

However, SEs (assuming no clustering) are misleadingly small for intervention effects and lead to substantively different conclusions about effects of 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 1–13 classrooms within a school and from 2–28 students within a classroom.

With relatively large cluster sizes, even very modest intra-cluster correlation can have a discernible impact on inferences.

Table 5: Fixed effects estimates from analysis that ignores clustering in the THKS scores.

Parameter	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 \times TV	-0.322	0.1302	-2.47

Extensions of Hierarchical Linear Models

The main focus has been on linear model for two-level hierarchical data.

As noted earlier, these models generalize in a direct way when there are three or more levels.

In addition, the models and conceptual approach can be extended to analyses of discrete outcomes, e.g., hierarchical generalized linear models.

Hierarchical Generalized Linear Models

Same conceptual approach as before.

The only difference is 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).

Complications:

Estimation is somewhat more difficult because multivariate integrals must be evaluated to compute the marginal log-likelihood.

Numerical integration techniques, for instance, Gaussian quadrature, can be used for maximizing the log-likelihood function.

ML estimation, using Gaussian quadrature, for two-level generalized linear models is implemented in some of the major statistical software packages (e.g., PROC NLMIXED in SAS).

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 hierarchical modelling (e.g., MLwiN and HLM).

Case Study 3: Developmental Toxicity Study of Ethylene Glycol

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) beginning just after implantation.

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.

Table 6: Descriptive statistics on fetal malformations from the ethylene glycol (EG) experiment.

Dose (mg/kg)	Dams	Fetuses	Fetal Malformations	
			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

Percentage of fetal malformations increases monotonically with dose, with less than 1% in control group and almost 60% in highest dose group.

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), consider logistic model relating log odds of fetal malformations to dose:

$$\text{logit}\{E(Y_{ij}|b_j)\} = \beta_0 + \beta_1 d_j + b_j,$$

where $d_j = \text{Dose}_j/750$ 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 clustering of fetal malformations within a litter.

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 ML estimation, are presented in Table 7.

Table 7: Fixed and random effects estimates for the fetal malformation data.

Parameter	Estimate	SE	Z
Intercept	-4.360	0.440	-9.92
Dose	1.336	0.166	8.06
Level 2 Variance:			
σ_b^2	2.517	0.685	3.68

ML estimation based on 50-point adaptive Gaussian quadrature.

The estimated regression parameter for dose indicates that the log odds of malformation increases with increasing dose.

The odds ratio for malformation, comparing the highest dose group 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).

Estimate of σ_b^2 indicates there are moderate litter effects, with heterogeneity across dams in underlying risk of producing fetuses with malformations.

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

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

Summary

We have discussed models for data with a hierarchical structure, where lower-level units are nested within higher-level units.

Hierarchical data can be challenging to analyze for at least two main reasons.

The first challenge in the analysis of hierarchical data is how best to account for the clustering that can arise at different levels of the hierarchy.

In hierarchical modelling literature, dominant approach for accounting for intra-cluster correlations is via random effects introduced 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, it raises subtle issues concerning interpretation of fixed effects and what is the relevant target of inference.

Second, the covariates can be measured at different levels, and the same covariate can operate at many different levels.

It is not always transparent how best to combine covariates measured at different levels within a single model.

As a result somewhat greater care is required in the interpretation of regression parameters in hierarchical models.

This is both their challenge and their reward!

Further Reading

There is an extensive literature on hierarchical/multilevel models that appears in the statistical, psychometric, and educational literature.

A comprehensive description of hierarchical/multilevel models, and their application to a wide range of problems, can be found in the following books:

Goldstein, H. (2003). *Multilevel Statistical Methods*, 3rd ed. London: Edward Arnold.

Longford, N. (1993). *Random Coefficient Models*. Oxford, UK: Oxford University Press.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Newbury Park, CA: Sage Publications.