Becoming familiar with multilevel modelling

Items of data vary independently of one another—or so the textbooks would have us believe. But what do you do when the data do not behave like that? Real-life populations can vary in complex ways that are dependent on each other. **Harvey Goldstein** introduces the mysteries of multilevel modelling.

Hierarchically structured data

Interesting real-life data rarely conform to classical textbook assumptions about data structures. Traditionally these assumptions are about observations that can be modelled with independently, and typically identically, distributed "error" terms. More often than not, however, the populations that generate data samples have complex structures where measurements on data units are not mutually independent, but depend on each other through complex structural relationships. For example, a household survey of voting preferences will typically show variation between households and voting constituencies. This implies that the replies from individual respondents within a household or constituency will be more alike than replies from individuals in the population at large. Another example of such "hierarchically structured data" would be measurements on students in different schools, where, for example, schools differ in terms of the average attainments of their students. In epidemiology we would expect to find differences in such things as fertility and disease rates across geographical and administrative areas.

To formalise this, consider the simple regression model:

$$y_i=\beta_0+\beta_1 x_i+e_i;\quad e_i\sim N(0,\sigma^2) \tag{1}$$

applied to a sample, say, of school students where i indexes students, the response y is an attainment measure and x is a predictor, such as a prior test score. The residuals e_i are assumed to have a normal distribution with variance σ^2, with the assumption that they are independently distributed. However, as pointed out above, this will not be true generally. Two randomly chosen students in the same school will tend to be more alike in their attainments, in this case adjusted for the predictor x, than two students chosen at random from different schools. This is the result of a number of factors, including those associated with schools selecting students on the basis of their apparent ability and/or social background. One way of recognising this is to extend model (1) as follows:

$$y_{ij}=\beta_0+\beta_1 x_{ij}+u_j+e_{ij};\ e_i\sim N(0,\sigma_e^2);\ u_j\sim N(0,\sigma_u^2) \tag{2}$$

where i indexes students as before and the additional subscript, j, indexes schools. We now have an explicit term, the school residual, u_j, which allows each school to contribute an "effect" to the response, i.e. to have a different intercept. The situation is illustrated in Figure 1 for four schools, where each school's line crosses the y axis at a different point.

Model (2) is often known as a "random intercept" or "variance components" model where σ_u^2 is the "between-school" variance and σ_e^2 is the "between-student" variance. We also assume that the school and student residuals are independent so that the total variance is given by $\sigma_u^2+\sigma_e^2$. Whereas in model (2) we have chosen to model the school effect as a

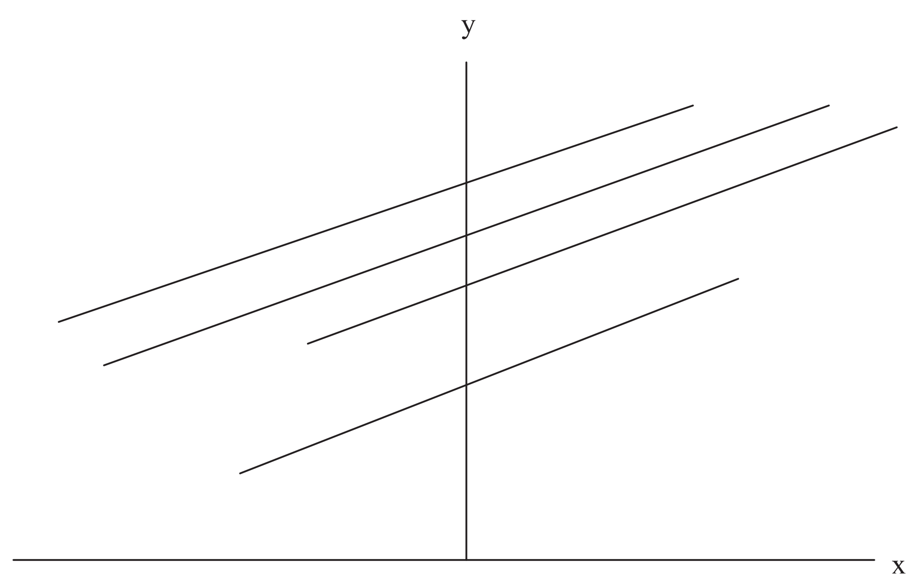


Figure 1 On a graph of results for students from four schools, each school has a different residual and y-intercept

random variable depending on a single parameter, the variance σ_u^2, an alternative to such a "random effects" model would be to fit school as a "fixed effect", using, for example, a set of $m - 1$ dummy variables where m is the number of schools. In some special circumstances this may be preferred, but more usually we would wish to consider the set of schools (or geographical areas or households) as a randomly chosen sample from a population of schools (or areas or households) about which we wish to make inferences. We may also wish to introduce school (level 2) predictors into the model, such as the resources available to the school or the average prior attainment of all the pupils in the school, to ascertain their effects on the response. An important advantage of the random effects approach is that it allows us to do this straightforwardly, and also to examine the effects on the between-school variance. With the fixed effects model we cannot introduce further school level effects at all, since the available degrees of freedom have been taken up completely by the dummy variables.

In a fixed effects model the coefficients of the dummy variables will provide direct estimates for the school effects. In the random effects model (2) we made the prior assumption that the u_j have a normal distribution so that an observation is more likely to be close to the mean than in the tails: this has the effect of moving estimates of the u_j towards the mean, and these estimates are often known, therefore, as "shrinkage" estimates.

From model (2) we can show that the correlation between two students in the same school is $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$, which is known as the intra-unit correlation or intra-cluster correlation, and for this particular model, but not more generally, is also the proportion of the total variance due to schools—the "variance partition coefficient" (VPC).

If we suppose that model (2) represents the "true" structure for the data but instead we fit model (1), it is well known that we shall obtain biased inferences. Although estimates of the regression coefficients are generally consistent, the apparent standard errors are too small so that confidence intervals will be too small and significance tests too optimistic, especially for level 2 predictors. This has been a long term concern in the sample survey literature where multi-stage or cluster sampling reflects population hierarchies, and procedures for "correcting" standard errors have been developed. There has been rather less emphasis here, however, on modelling the structure of the population itself. Multilevel modelling, in contrast, concentrates on modelling the complex population structure: where this structure is important it also generally provides a more efficient approach than the traditional survey sampling one.

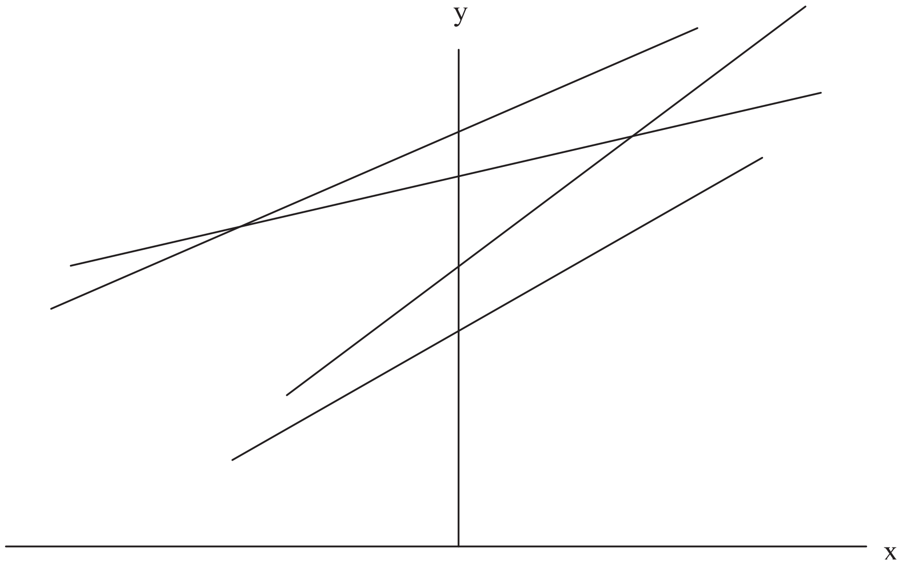


Figure 2 Here, more level in the data structure are reflected in the graph. The lines are no longer parallel

To illustrate the importance of proper modelling, suppose we have a two-level structure consisting of students (level 1 units) grouped within schools (level 2 units) where there are n students within each of m schools. If the unadjusted VPC values for y and x, when each of these is modelled as a response with just an intercept term, are 0.20 with 76 level 1 units per level 2 unit, then the true standard error of β is, on average, twice the simple regression estimate. Hence, confidence intervals based on that for the simple regression estimate will be too short and significance tests will reject the null hypothesis too often.

Estimation and extensions

The simple normal two-level model (2) can be extended in several ways.

The number of levels in our models is not restricted to two, and there are examples of applications with four, five or even more levels. For many data structures, a single random effect for a higher level unit will not be adequate. Thus, in the schools example it is commonly found that the relationship with x will vary from school to school, which we can show diagrammatically, as in Figure 2 where the lines are no longer parallel.

This gives us a "random coefficient" model that can be written as

$$y_{ij} = \beta_0 + \beta_{1j} x_{ij} + u_{0j} + e_{ij};$$
$$\beta_{1j} = \beta_1 + u_{1j} \tag{3}$$

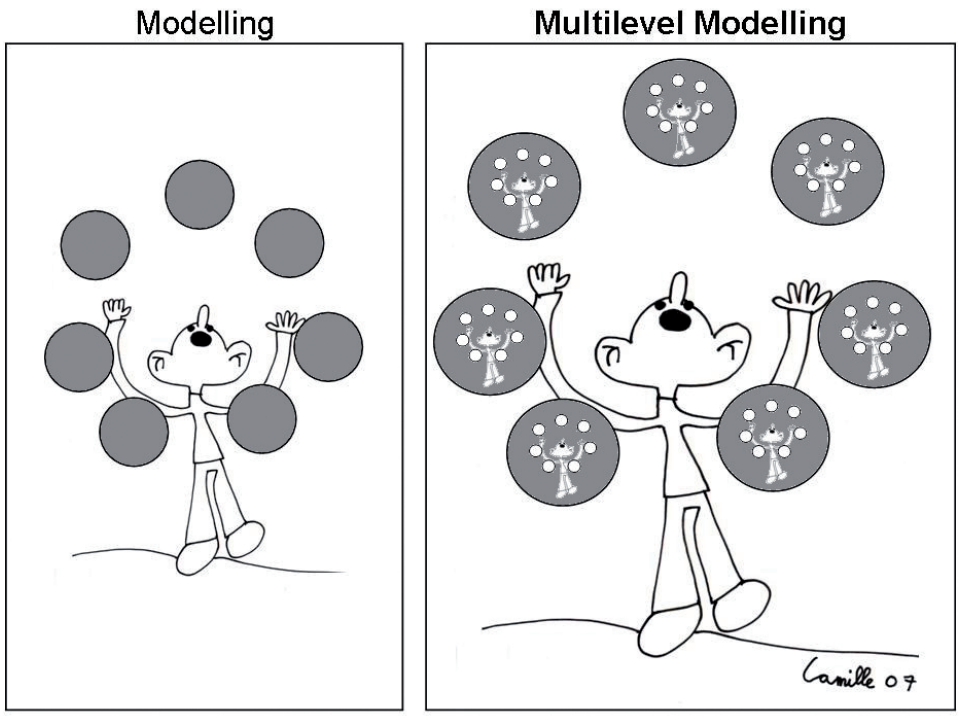
where we assume that (u_{0j}, u_{1j}) is bivariate normal with

$$\mathrm{var}(u_{0j}) = \sigma_{u0}^2,\ \ \mathrm{var}(u_{1j}) = \sigma_{u1}^2,\ \ \mathrm{cov}(u_{0j}, u_{1j}) = \sigma_{u01}$$

and the only difference from model (2) is that the coefficient of x is assumed to be random across schools, with mean β_1, variance σ_{u1}^2 and a covariance σ_{u01} with the random intercept term. Having a random coefficient also implies that the between-school variance, and hence the VPC, are (quadratic) functions of x.

An important application of random coefficient models is to repeated measures data where the level 1 units are the measurement occasions and the level 2 units are, for example, individual animals or human beings. In such models the response is typically a function of time or age where the parameters of the function vary across level 2 units. Formulated as a two-level model the data analyst is not restricted to having all units measured on the same set of occasions; nor that all units should have the same number of occasions.

If a categorical predictor variable has a random coefficient then this can be interpreted as each category having a different level 2 variance—a particular instance of where the level 2 variance is a function of a level 1 predictor. This leads on to the idea of more general ways of modelling variation. Thus, for example, if we have a predictor variable "gender" then not only could we model different between-school variances for boys and girls, we could also consider modelling different between-student variances for boys and girls: the between-boy variance is allowed to be different from the between-girl variance. It turns out that we can formulate directly quite complex linear and nonlinear models for the variance between units at any level of the data hierarchy.



Another extension is to generalised linear models such as a logistic model for a categorical response. Suppose that we have a binary response, e.g. whether or not a hospital patient is infected with a virus. A basic two-level model would have two components: the first expresses the probability of a positive response as a function of patient and hospital characteristics in the "fixed part" of the model together with a level 2 residual:

$$\text{logit}(\sigma_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_j x_{2j} + u_j;\ u_j \sim N(0, \sigma_u^2) \quad (4)$$

where x_1 could be a patient characteristic and x_2, say, a level 2, hospital, characteristic. Our actual response, y, is (0,1) and we specify the second part of the model where this has a Bernoulli distribution or binomial distribution with denominator 1 and probability π_{ij}. Model (4) can be extended to ordered or unordered multicategory responses and also specified for other link functions.

For all of the models considered thus far we can have multivariate versions where a set of responses depends on a vector of fixed predictors with residual covariance matrices at each level of the hierarchy. Important special cases of such models are multilevel factor analysis and structural equation models. Recent work has also extended these models to allow responses to be measured at different levels of the data hierarchy and also to allow mixtures of response types so that normal and categorical responses can be modelled jointly.

More complex structures

So far only purely hierarchical models have been described. In practice, however, data structures are often more complicated. Consider an educational example where students are followed through both their primary and their secondary education with the response being attainment at the end of secondary school. For any given primary school, students will generally move to different secondary schools, and any given secondary school will draw students from a number of primary schools. We therefore have a *cross-classification* of primary by secondary schools where each cell of the classification will be populated by students (some may be empty). When we model such a structure we have a contribution to the response that is the sum of an effect from the primary and an effect from the secondary school attended by a student.

Pursuing this example further we know that students do not all remain in the same secondary or primary school. Thus, a student may attend two or three primary schools so the "effect" of primary school on the response is the average effect of all the primary schools attended. These models are referred to as "multiple-membership models" since a lower level unit can belong to more than one higher level unit. In our model, therefore, we would apply a set of weights to the relevant primary school u_j to reflect, for example, the time spent in each school. Such models are also useful for studying spatial correlation structures where individuals can be viewed as belonging to several areas simultaneously with appropriate weightings.

Estimation

In the space of this article there is little room to describe estimation methods. Briefly, however, these are somewhat more complex than those for single-level models, especially for the more complex structures. Over the last 20 years, software for fitting these models has become widespread and the basic models can be fitted readily using any of the major statistical packages. Software also exists for fitting the more complex models mentioned: a comparative analysis of software can be found on the Centre for Multilevel Modelling website at Bristol (`www.cmm.bristol.ac.uk`). Both maximum likelihood and Bayesian estimation via Markov chain Monte Carlo methods are available, and the latter becomes particularly important as the complexity of the models increases with many more parameters. For very large datasets estimation time can often become very lengthy, and this can limit the amount of time for model exploration, although as computers become more powerful this is becoming less of a problem.

Concluding comments

Multilevel models have found applications in many discipline areas and have been developed for many kinds of data. These include time series data, survival or event history data, smoothing splines, bootstrapping, measurement errors, missing data and more. In the space of a short article it is not possible to give more than an introduction to the wide applicability of multilevel models and their potential for modelling the actual complexity of real data (a more technical account is available[1], and Snijders and Bosker[2] provide a good detailed introduction to the subject). Using these models will often lead to quite different inferences when compared with the use of simpler models, and the existence of easily accessible software and expository texts has removed practical objections to their use. Many university courses now incorporate multilevel modelling, and these models are rapidly becoming a standard feature of every applied statistician's toolkit.

References

1. Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn. London: Arnold.
2. Snijders, T. and Bosker, R. (1999) *Multilevel Analysis*. London: Sage.

Harvey Goldstein is Professor of Social Statistics at the University of Bristol. His main research is developing methodology, software and training materials for multilevel modelling.