1

An introduction to multilevel models

1.1 Hierarchically structured data

Many kinds of data, including observational data collected in the human and biological sciences, have a *hierarchical, nested,* or *clustered* structure. For example, animal and human studies of inheritance deal with a natural hierarchy where offspring are grouped within families. Offspring from the same parents tend to be more alike in their physical and mental characteristics than individuals chosen at random from the population at large. For instance, children from the same family may all tend to be small, perhaps because their parents are small or because of a common impoverished environment. Many designed experiments, such as clinical trials carried out in several randomly chosen centres or groups of individuals, also create data hierarchies.

For now, we are concerned only with the *fact* of such hierarchies, not their provenance. The principal applications are those from the social and medical sciences, but the techniques are, of course, applicable more generally. In subsequent chapters, as we develop the theory and techniques with examples, we see how a proper recognition of these natural hierarchies allows us to obtain more satisfactory answers to important questions.

We refer to a hierarchy as consisting of *units* grouped at different *levels*. Thus offspring may be the level 1 units in a 2-level structure where the level 2 units are the families: students may be the level 1 units clustered or nested within schools that are the level 2 units.

The existence of such data hierarchies is neither accidental nor ignorable. Individual people differ, as do individual animals, and this differentiation is mirrored in all kinds of social activity where the latter is often a direct result of the former; for

Multilevel Statistical Models: 4th Edition Harvey Goldstein

example, when students with similar motivations or aptitudes are grouped in highly selective schools or colleges. In other cases, the groupings may arise for reasons less strongly associated with the characteristics of individuals, such as the allocation of young children to elementary schools, or the allocation of patients to different clinics. Once groupings are established, even if their establishment is effectively random, often they will tend to become differentiated. This differentiation implies that the group and its members both influence and are influenced by the group membership. To ignore this risks overlooking the importance of group effects, and may also render invalid many of the traditional statistical analysis techniques used for studying data relationships.

We look at this issue of statistical validity in the next chapter. For now, one simple example will show its importance. A well-known and influential study of the teaching styles used with primary (elementary) school children carried out in the 1970s (Bennett, 1976), claimed that children exposed to so-called ‘formal’ styles of teaching reading exhibited more progress than those who were not. The data were analysed using traditional multiple regression techniques which recognised only the individual children as the units of analysis and ignored their groupings within teachers and into classes. The results showed statistically significant differences. Subsequently, Aitkin *et al*. (1981) demonstrated that when the analysis accounted properly for the grouping of children into classes, the significant differences disappeared and the ‘formally’ taught children could not be shown to differ from the others.

This re-analysis is the first important example of a *multilevel* analysis of social science data. In essence what was occurring here was that the children within any one classroom, because they were taught together, tended to be similar in their performance. As a result they provided rather less information than would have been the case if the same number of students had been taught separately by different teachers. In other words, the basic unit for purposes of comparison should have been the teacher not the student. The function of the students can be seen as providing, for each teacher, an estimate of that teacher’s effectiveness. Increasing the number of students per teacher would increase the precision of those estimates but not change the number of teachers being compared. Beyond a certain point, simply increasing the numbers of students in this way hardly improves things at all. On the other hand, increasing the number of teachers to be compared with the same or an even smaller number of students per teacher considerably improves the precision of the comparisons.

Researchers have long recognised this issue. In education, there has been much debate (see Burstein *et al*., 1980) about the so-called ‘unit of analysis’ problem just outlined. Before multilevel modelling became well developed as a research tool, the problems of ignoring hierarchical structures were reasonably well understood, but they were difficult to solve because powerful general purpose tools were unavailable. Special purpose software, for example, for the analysis of genetic data, has been available longer but this was restricted to ‘variance components’ models (see Chapter 2) and was not suitable for handling general linear models. Sample survey workers have recognised this issue in another form. When population surveys are carried out, the sample design typically mirrors the hierarchical population structure, in terms of geography and household membership. Elaborate procedures have been developed to take such structures into account when carrying out statistical analyses. We look at this in more detail in Chapter 10.

The remainder of this chapter discusses some general issues and introduces the major topics explored in this book.

1.2 School effectiveness

Schooling systems present an obvious example of a hierarchical structure, with pupils clustered within schools, which themselves may be clustered within education authorities or boards. It therefore provides a useful way to introduce some basic ideas of multilevel modelling. Educational researchers have long been interested in comparing schools and other educational institutions, most often in terms of the achievements of their pupils. Such comparisons have several aims, including the aim of public accountability (Goldstein, 1997) but, in research terms, interest is usually focused upon studying the factors that explain school differences.

Consider the common example where test or examination results at the end of a period of schooling are collected from a randomly chosen sample of schools. The researcher wants to know whether a particular kind of subject streaming practice in some schools is associated with improved examination performance. She also has good measures of the pupils' achievements when they started the period of schooling so that she can control for this in the analysis. The traditional approach to the analysis of these data would be to carry out a regression analysis, using performance score as the response, to study the relationship with streaming practice, adjusting for the initial achievements as covariates. This is very similar to the teaching styles analysis described in the previous section, and suffers from the same lack of validity through failing to take account of the school level clustering of students.

An analysis that explicitly models the manner in which students are grouped within schools has several advantages. Firstly, it enables data analysts to obtain statistically efficient estimates of regression coefficients. Secondly, by using the clustering information it provides correct standard errors, confidence intervals and significance tests, and these generally will be more 'conservative' than the traditional ones that are obtained simply by ignoring the presence of clustering – just as Bennett's previously statistically significant results became non-significant on reanalysis. Thirdly, by allowing the use of covariates measured at any of the levels of a hierarchy, it enables the researcher to explore the extent to which differences in average examination results between schools are accountable for by factors such as organisational practice or in terms of other characteristics of the students. It also makes it possible to study the extent to which schools differ for different kinds of students, for example to see whether the variation between schools is greater for initially high scoring students than for initially low scoring students (Goldstein *et al.*, 1993) and whether some factors are better at accounting for or 'explaining' the variation for the former students than for the latter. Finally, there may be interest in the relative ranking of schools, using the performances of their students after adjusting for intake achievements. This can be

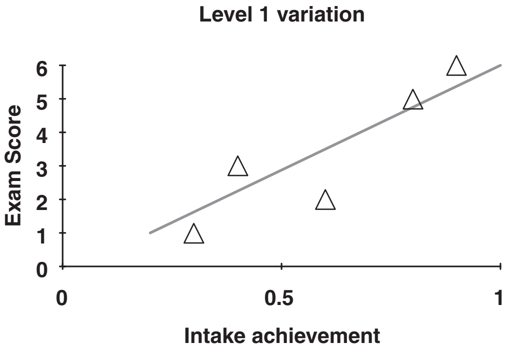

Figure 1.1 A simple regression.

done straightforwardly using a multilevel modelling approach and we shall see an example in Chapter 2.

To fix the basic notion of a level and a unit, consider Figures 1.1 and 1.2, which are based on hypothetical relationships.

Figure 1.1 shows the exam score and intake achievement scores for five students in a school, together with a simple regression line fitted to the data points. The residual variation in the exam scores about this line is the *level 1 residual variation*, since it relates to level 1 units (students) within a sample level 2 unit (school). In Figure 1.2 the three lines are the simple regression lines for three schools, with the individual student data points removed. These vary in both their slopes and their intercepts (where they would cross the exam axis), and this variation is the *level 2 variation*. It is an example of complex level 2 variation since both the intercept and slope parameters vary.

The other extreme to an analysis which ignores the hierarchical structure is one which treats each school completely separately by fitting a different regression model within each one. In some circumstances, for example where we have very few schools

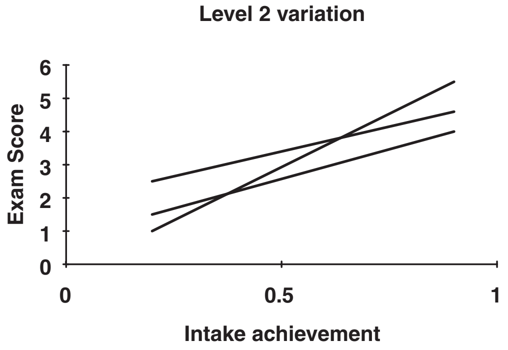

Figure 1.2 Complex level 2 variation.

and moderately large numbers of students in each, this may be efficient. It may also be appropriate if we are interested in making inferences about just those schools. If, however, we regard these schools as a (random) sample from a population of schools and we wish to make inferences about the variation between schools in general, then a full multilevel approach is called for. Likewise, if some of our schools have very few students, fitting a separate model for each of these will not yield reliable estimates: we can obtain more precision by regarding the schools as a sample from a population and using the information available from the whole sample data when making estimates for any one school. This approach is especially important in the case of repeated measures data where we typically have very few level 1 units per level 2 unit (Chapter 5).

We introduce the basic procedures for fitting multilevel models to hierarchically structured data in Chapter 2 and discuss the design problem of choosing the numbers of units at each level in Chapter 3.

1.3 Sample survey methods

We have already mentioned sample survey data. The standard literature on surveys, reflected in survey practice, recognises the importance of taking account of the clustering in complex sample designs. Thus, in a household survey, the first stage sampling unit will often be a well-defined geographical unit. From those which are randomly chosen, further stages of random selection are carried out until the final households are selected. Because of the geographical clustering exhibited by measures such as political attitudes, special procedures have been developed to produce valid statistical inferences, for example, when comparing mean values or fitting regression models (Skinner *et al*., 1989).

While such procedures usually have been regarded as necessary they have not generally merited serious substantive interest. In other words, the population structure, insofar as it is mirrored in the sampling design, is seen as a ‘nuisance factor’. By contrast, the multilevel modelling approach views the population structure as of potential interest in itself, so that a sample designed to reflect that structure is not merely a matter of saving costs as in traditional sample design, but can be used to collect and analyse data about the higher level units in the population.

Although the direct modelling of clustered data is statistically efficient, it will generally be important to incorporate weightings in the analysis which reflect the sample design or, for example, patterns of non-response, so that robust population estimates can be obtained and so that there will be some protection against serious model misspecification. A procedure for introducing external unit weights into a multilevel analysis is discussed in Chapter 3 and survey data are discussed in Chapter 10.

1.4 Repeated measures data

A different example of hierarchically structured data occurs when the same individuals or units are measured on more than one occasion, as occurs in studies of animal and human growth. Here the occasions are clustered within individuals that represent the level 2 units with measurement occasions as the level 1 units. Such structures are typically strong hierarchies because there is much more variation between individuals in general than between occasions within individuals. In the case of child height growth, once we have adjusted for the overall trend with age, the variance between successive measurements on the same individual is generally no more than about 5 % of the variation in height between children.

The traditional literature on the analysis of such repeated measurement data (see, for example, Goldstein, 1979), has more or less successfully confronted the statistical problems. It has done so, however, by requiring that the data conform to a particular, balanced, structure. Broadly speaking these procedures require that the measurement occasions are the same for each individual. This may be possible to arrange, but often in practice individuals will be measured irregularly, some of them a great number of times and some perhaps only once. By considering such data as a 2-level structure, however, we can apply the standard set of multilevel modelling techniques that allow for any pattern of measurements while providing statistically efficient parameter estimation. At the same time, modelling such data as a 2-level structure presents a simpler conceptual understanding and leads to a number of interesting extensions (see Chapter 5).

One particularly important extension occurs in the study of growth where the aim is to fit growth curves to measurements over time. In a multilevel framework this involves, in the simplest case, each individual having their own straight line growth trajectory with the intercept and slope coefficients varying between individuals (level 2). When the level 1 measurements, considered as deviations from each individual's fitted growth curve, are not independent but have an autocorrelated or time series structure, neither the traditional procedures nor the basic multilevel ones are adequate. This situation may occur when measurements are made very close together in time so that a 'positive' deviation from the curve at one time implies also a positive deviation after the short interval before the next measurement. Chapter 5 considers methods for handling such data.

1.5 Event history and survival models

Modelling the time spent in various states or situations is important in a number of areas. In industry the 'time to failure' of components is a key factor in quality control. In medicine, the survival time is a fundamental measurement in studying certain diseases. In economics, the duration of employment periods is of great interest. In education, researchers often study the time students spend on different tasks or activities.

In studying employment histories, any one individual will generally pass through several periods of employment or unemployment, while at the same time changing his characteristics, for example his level of qualifications. From a modelling point of view we need to consider the length of time spent in each type of employment, relating this both to constant factors such as an individual's social origins or gender, and to changing or time dependent factors such as qualifications and age. In this case the multilevel structure is analogous to that for repeated measures data, with periods taking the place of occasions. Furthermore, generally we would have a further, higher level of the hierarchy, since individuals, which are the level 2 units, are themselves typically clustered into workplaces, which now constitute level 3 units.[1] The structure may be even more complicated if these workplaces change from period to period; to include this level in our model, we need to consider cross-classifications of the units (see below). There are particular problems that arise when studying event duration data that are encountered when some information is 'censored' in the sense that instead of being able to observe the actual duration we only know that it is longer than some particular value, or in some cases less than a particular value (see Chapter 11).

1.6 Discrete response data

Until now, we have assumed implicitly that our response or dependent variable is continuously distributed; for example, an exam score or anthropometric measure such as height. Many kinds of statistical models, however, deal with categorised responses, in the simplest case with proportions. Thus, we might be interested in a mortality rate, or an examination pass rate and how these vary from area to area or from school to school.

In studying mortality rates in a population, it is often of great concern to try to understand the factors associated with variations from area to area or community to community. This produces a basic 2-level structure with individuals at level 1 and communities at level 2. A typical study might record deaths over a given time period together with the characteristics of the individuals concerned, and level 2 characteristics of the communities, such as their sizes or social compositions. One analysis of interest would be to see whether any of these explanatory variables could explain between-community variation. Another interest might be in studying whether mortality rate differences, say between men and women, varied from community to community.

Such models, part of the class known as generalised linear models, have been available for some time for single level data (McCullagh and Nelder, 1989), with associated software. In Chapter 4 we show how to fit multilevel models with different types of categorical response. Chapter 7 extends this to consider multivariate models with mixtures of different response types.

1.7 Multivariate models

An interesting special case of a 2-level model is the multivariate linear (or generalised linear) model. Suppose we have taken several measurements on an individual, for

[1] Formally, we can regard unemployment for this purpose as a particular workplace.

example their systolic and diastolic blood pressure and their heart rate. If we wish to analyse these together as response variables we can do so by setting up a multivariate, in this case 3-variate, model with explanatory variables such as age, gender, social background, smoking exposure, etc. We can think of this as a 2-level model by considering each individual as a level 2 unit, with the three measurements constituting the level 1 units, rather as occasions did for the repeated measures model. Chapter 6 shows how this formal device for specifying a multivariate model yields considerable benefits. By considering further higher levels, such as clinics, we have a simple way of specifying a multivariate multilevel model and we can also have models where the responses can be measured at different levels of a data hierarchy; see Chapter 5 for an example. Also, if some individuals do not have all the measurements, for example if they are randomly missing a blood pressure measurement, then this is automatically taken account of in the analysis, without the need for special procedures for handling missing data.

A particularly important application occurs where measurements are missing by design rather than at random. In certain kinds of surveys, known as rotation designs, and in certain kinds of educational assessments known as matrix sample designs, each individual unit has only a subset of measurements made on it. For example, in large-scale testing programmes, the full range of tests may be too extensive for any one student, so that each student responds to only one, randomly assigned, combination. Such designs can be viewed as having a multivariate response, with the full set of tests constituting the complete multivariate response vector, and every student having some tests missing. Such designs can become rather complex, especially since the students themselves are clustered into schools. By viewing the data as a single hierarchy in which the multivariate responses are level 1, we obtain an efficient and readily interpretable analysis.

The multivariate multilevel model is also used as the basis for dealing with missing data in multilevel models and this is developed in Chapter 16.

1.8 Nonlinear models

Some kinds of data are better represented in terms of nonlinear rather than linear models. For example, the modelling of discrete response data is considered formally as a case of modelling nonlinear data. Many kinds of growth data are conveniently modelled in this way, especially during periods of rapid and complex growth such as early infancy and at the approach to adulthood when growth approaches an upper asymptote (Goldstein, 1979). Other examples arise when the response variable has inherent constraints. For example, biochemical activity patterns in patients may exhibit asymptotic behaviour, or cyclical patterns, both of which may be difficult to model using purely linear models. Chapter 9 introduces such models and shows how to extend the linear multilevel model to this case. It also considers cases where variances and covariances can be modelled as nonlinear functions of explanatory variables (see also Chapter 17).

1.9 Measurement errors

Many variables of interest in the human sciences contain ‘noise’ or random measurement error. This may be due to observer error as when measuring the weight of an animal, or an inherent result of being able to measure only a small sample of behaviour as in educational testing. It is well known that when variables in statistical models contain relatively large components of such error the resulting statistical inferences can be very misleading unless careful adjustments are made (Fuller, 2006). In the case of simple regression, when the explanatory or independent variable is measured with error, the usual estimate of the regression line slope is an underestimate compared to that which would result if the measurement were available without error. This is particularly important in studies of school effectiveness where the fitting of intake achievement scores is important but where such scores often have large components of measurement error.

An important case is where we have a level 2 variable that is a ‘compositional’ variable. That is, it is a measurement aggregated from the characteristics of the level 1 units within the level 2 units. Thus, for example, the mean intake achievement and the standard deviation of the intake achievements of all the pupils in a school are compositional variables that may, and indeed sometimes do, affect the final achievements of each individual student. Likewise, in a household survey, we may consider that a measure of the average social status or the percentages of households in each social group, using all the households in the immediate community, are important explanatory variables to fit in a model. The problem arises when it is possible to collect data only upon some of the level 1 units, this often being the case with household sample surveys. What we then have is an estimate of a compositional variable that is measured with error, in the case of household surveys typically with a very large error. In many educational studies, this also occurs where only a small proportion of students within a class or school are sampled. Chapter 14 discusses the problems of dealing with measurement errors in multilevel models.

1.10 Cross classifications and multiple membership structures

We have already alluded to examples where units are cross-classified as well as clustered. In area studies the definition of an individual’s geographical area is contingent upon the context being considered. Thus, the relevant location unit for purposes of leisure may not be the same as that surrounding the environment of work or schooling. We can conceive, formally, of individuals belonging simultaneously to both types of unit, each of which may have an influence on a person’s life.

In most schooling systems, students move from elementary to secondary or high school. We might expect that both the elementary and secondary schools attended will influence a student’s achievements or attitudes measured at the end of secondary school. Thus the level 2 units are of two types, elementary school and secondary school, where each ‘cell’ of their cross classification contains some, or possibly no students. In this example, a third classification could be the area or neighbourhood where the student lives. Chapter 12 explores such cross-classified structures.

An interesting situation occurs where for a single level 2 classification, level 1 units may belong to more than one level 2 unit. A sociological example concerns childrens’ and adults’ friendship patterns. An individual may belong to several ‘friendship groups’ simultaneously. The characteristics of the members of each group will influence such an individual, in relation to the individual’s exposure to the group. In a longitudinal study of schooling, many students will change schools during the course of the study. The contribution to the response from schools will therefore reflect, for these students, the ‘effect’ of every school they have attended. With a suitable set of weights to reflect the time spent in each school this can be taken into account in the analysis. Such ‘multiple membership models’ are discussed in Chapter 13.

To handle the complexity of multiple membership and cross-classified structures, as well as mixtures of these, a special notation and set of diagrams will be introduced that allows a complete specification of such models (see Notation).

1.11 Factor analysis and structural equation models

In many areas of the social sciences where measurements are difficult to define precisely, an investigator might suppose that there is some underlying construct which cannot be measured directly but nevertheless can be assessed indirectly by measuring a number of relevant indicators. Structural equation modelling, and in particular the special case of factor analysis, was developed for this purpose, typically dealing with individuals’ behaviour, attitudes or mental performance. Where individuals are grouped within hierarchies, for the reasons already discussed, it is important to carry out such analyses in a multilevel framework. For example, we may be interested in underlying individual attitudes based upon a number of indicators. Data on such indicators may be available over time and we can postulate a model whereby the underlying attitude varies from individual to individual (level 2) and also varies randomly over time within individuals (level 1). The model can then be further elaborated by studying whether there is any systematic change over time and whether this varies across individuals. Chapter 8 discusses such models.

1.12 Levels of aggregation and ecological fallacies

When studying relationships among variables, there has often been controversy about the appropriate ‘unit of analysis’. We have alluded to this already in the context of ignoring hierarchical data clustering and, as we have seen, the issue is resolved by explicit hierarchical modelling.

One of the best known early illustrations of what is often known as the ecological or aggregation fallacy was the study by Robinson (1950) of the relationship between literacy and ethnic background in the United States. When the mean literacy rates and mean proportions of Black Americans for each of nine census divisions are correlated the resulting value is 0.95, whereas the individual-level correlation ignoring the grouping is 0.20. Robinson was concerned to point out that aggregate-level relationships could not be used as estimates for the corresponding individual-level relationships and this point is now well understood. In Chapter 3, we discuss some of the statistical consequences of modelling only at the aggregate level.

Sometimes the aggregate level is the principal level of interest, but nevertheless a multilevel perspective is useful. Consider the example (Derbyshire, 1987) of predicting the proportion of children socially ‘at risk’ in each local administrative area for the purpose of allocating central government expenditure on social services. Survey data are available for individual children with information on risk status so that a prediction can be made using area based variables as well as child and household based variables. The probability (π) of a child being ‘at risk’ was estimated by the following (single level) equation

$$\text{logit}(\pi) = -6.3 + 5.9x_1 + 2.2\,x_2 + 1.5x_3$$

where x_1 is the proportion of children in the area in households with a lone parent, x_2 is the proportion of households in each area which have a density of more than 1.5 persons per room and x_3 is the proportion of households whose ‘head’ was born in the British ‘New Commonwealth’ or Pakistan. All these explanatory variables are measured at the aggregate area level and the response is the proportion of children at risk in each area. Although we can regard this analysis as taking place entirely at the area level (with suitable weighting for the number of children in each area), there are advantages in thinking of it as a 2-level model with each child being a level 1 unit and the response variable being the binary response of whether or not the child is at risk.

Firstly, this allows us to incorporate possibly important variables that are measured at the child level, for example whether or not each child’s household is overcrowded. Including such level 1 variables may greatly improve the predictive power of the model. With the results of such a model we can then form a prediction for each area by aggregating over the known numbers of children living in overcrowded households. Secondly, the possibility of modelling the characteristics of children or their households allows the possibility of an allocation formula that can take account of costs and benefits related to the actual composition of each area in terms of these child characteristics.

1.13 Causality

In the natural sciences, experimentation has a dominant position when making causal inferences. This is both because the units of interest can be manipulated experimentally, typically using random allocation, and because there is a widespread acceptance that the results of experiments are generalisable over space and time. The models described in this book can be applied to experimental or non-experimental data, but the final causal inferences can differ. Nevertheless, most of the examples used are from non-experimental studies in the human sciences and a few words on causal inferences from such data may be useful.

If we wish to answer questions about a possible causal relationship between, say, class size and educational achievement, an experimental study would need to assign different numbers of level 1 units (students) randomly to level 2 units (classes or teachers) and study the results over a time period of several years. This would be time consuming and could create ethical problems. In addition to such practical problems, any single study would be limited in time and place, and require extensive replication before results could be generalised confidently. The specific context of any study is important; for example, the state of the educational system and the resources available at the time of the study. The difficulty from an experimental viewpoint is that it is practically impossible to allocate randomly with respect to all such possible confounding factors.

A further limitation of randomised controlled trails (RCTs) is that they cannot necessarily deal with situations where the *composition* of a higher level unit interacts with the treatment of interest, to affect the responses of lower level units. Thus, in schooling studies the size of class may affect the progress of students only when the proportion of ‘low achieving’ students is above a certain threshold. Randomisation will tend to eliminate classes with extreme proportions so that such effects may not be discovered. Goldstein (1998) looks at this case in more detail.

None of this is to say that randomised experiments should never be undertaken, rather that on their own they may have limited potential for making general statements about causality. Whether an experiment fails or succeeds in demonstrating a relationship, there will almost always be further explanations for the findings which require study. Even if an experiment appears to eliminate a possible relationship, for example, demonstrating a negligible relationship between class size and attainment, it may be legitimate to query whether a relationship nevertheless exists for specific subgroups of the population.

In the pursuit of causal explanations, it is desirable to have some guiding underlying principles or theories. It is these which will tell us what kinds of things to measure and how to be critical of findings. For example, in studies of the relationship between perinatal mortality and maternal smoking in pregnancy (Goldstein, 1976) we can attempt to adjust for confounding factors, such as poverty, which may be responsible for influencing both smoking habits and mortality. We can also study how the relationship varies across groups and seek measures which explain such variation. We might also, in some circumstances, be able to carry out randomised experiments, assigning, for example, intensive health education to a randomly selected ‘treatment’ group and comparing mortality rates with a ‘control’ group.

A multilevel approach could be useful here in two different ways. Firstly, pregnant women will be grouped hierarchically, geographically and by medical institution, and the between-area and between-institution variation may affect mortality and the relationship between mortality and smoking. Secondly, we will be able often to obtain serial measurements of smoking, so allowing the kind of repeated measures

2-level modelling discussed earlier. This will allow us to study how changes in smoking are related to mortality, and permit a more detailed exploration of possible causal mechanisms.

Multilevel models can often be used to identify units with extreme values. In school effectiveness studies, an exploration of school-level residual estimates (see Chapter 2) may identify those which are highly atypical, having adjusted for ‘contextual’ variables such as the intake characteristics of their students. These can then be selected for further scrutiny, for example by means of intensive case studies, so forming a link between the quantitatively based multilevel analysis and a more qualitatively based investigation which would seek to identify detailed causal processes.

The notion of causation, especially in non-randomised studies, is controversial and has a long philosophical history. Recent work has extended the range of tools available for studying causality and a useful introduction is given by Sobel (2000); Rosenbaum (1995) provides a detailed discussion of issues. A particular assumption in deriving causal inferences that is important in multilevel modelling, is the ‘stable unit treatment assumption’ (SUTVA). This states that the response to a ‘treatment’ assigned to an individual, for example being placed in a small rather than large class for teaching purposes, does not depend on the assignments to other individuals – that there is no interference between units. Where there are hierarchical structures, such as schools, within which different treatments may occur, this ‘non-interference’ assumption may not hold. It may be possible to model such dependencies or to redesign studies to avoid this problem. A discussion of this problem in the context of class size studies is given by Blatchford *et al*. (1998).

Finally, many of the concerns addressed by multilevel models are to do with straightforward prediction. Thus, for example, in Chapter 6 we use a 2-level model of children’s growth for the purpose of predicting adult height. In studies of school effectiveness we may be interested in understanding the causes of school differences, but we may be concerned also with the less ambitious task of predicting which school is likely to produce the best (on average) examination result for a student with given initial characteristics and achievements.

1.14 The latent normal transformation and missing data

In Chapter 7, we show how various discrete responses can be modelled simultaneously by transforming them jointly to an underlying multivariate normal distribution. This has an important application to ways of handling missing data, as discussed in Chapter 16. In particular, Chapter 16 shows how a multiple imputation approach can incorporate not only mixtures of different types of response but also responses at different levels of a data hierarchy. This provides a very general procedure for handling missingness in complex data structures.

1.15 Other texts

While the present volume aims to provide a comprehensive coverage of the topic of multilevel models, there are now many other texts which deal with specialised areas. Many of these are referenced in the appropriate chapters, but there are also several books which provide good introductions as well as detailed worked examples and technical details. Among these are Snijders and Bosker (1999), Little *et al*. (2000), Heck and Thomas (2000), McCulloch and Searle (2001), Hox (2002), Bryk and Raudenbush (2002), Skrondal and Rabe-Hesketh (2004), Lee *et al*. (2006) and Gelman and Hill (2007). There are also edited collections of articles on particular application areas. Leyland and Goldstein (2001) bring together a collection of papers on the multilevel modelling of health statistics and Courgeau (2007) brings together a number of perspectives on the interpretation of multilevel data.

1.16 A caveat

The purpose of this book is to bring together techniques for the analysis of highly structured multilevel data. The application of such techniques has already begun to yield new and important insights in a number of areas as the following chapters illustrate. Software is now widely available (Chapter 18), so that the application of these techniques should become relatively straightforward, even routine.

All this is welcome, yet despite their usefulness, models for multilevel analysis cannot be a universal panacea. In circumstances where there is little structural complexity, they may be hardly necessary and traditional single level models may suffice both for analysis and presentation. On the other hand multilevel analyses can bring extra precision to attempts to understand causality, for example, by making efficient use of student achievement data in attempts to understand differences between schools. They are not, however, substitutes for well grounded substantive theories, nor do they replace the need for careful thought about the purpose of any statistical modelling. Furthermore, by introducing more complexity they can extend but not necessarily simplify interpretations.

Multilevel models are tools to be used with care and understanding.