

Chapter 4

Multilevel Data Structures



Abstract This chapter covers different data structures for which multilevel modelling is appropriate, giving examples of each. The first such structure is the strict hierarchy, which may be the structure that first comes to mind when you think about multilevel models: patients who are treated in hospitals or individuals living in certain areas. Then there are multistage sampling designs and the evaluations of community interventions, in which it is the study design that imposes the hierarchical structure on the data. There are studies that collect data over time, either through repeated cross-sections or through repeated measures on an individual. This introduces another hierarchy to the data. Such models can be expanded to include multiple responses: more than one measure on each individual. These can be analysed simultaneously and considered as being nested within individuals. Then there are structures which are not strictly hierarchical. Firstly, the cross-classified model, in which there is an overlap between different classifications meaning that units at one level are not nested neatly within units at another level. Secondly, the multiple membership model in which an individual at one level can be a member of a number of different units at a higher level. Thirdly, the correlated cross-classified model, used when cross-classifications are repeated over time. Finally, this chapter briefly covers some further structures that can be modelled as multilevel structures. The idea of including these further structures is to make the reader aware of the range of models that could potentially be fitted to data rather than to cover them in detail.

Keywords Multilevel analysis · Hierarchy · Community interventions · Time dependent data · Multiple responses · Cross-classified models · Multiple membership models

In Chap. 3, we considered why levels were important and what might constitute a level in your data. We now expand on these ideas as we show a wide range of data structures that can be considered to be hierarchical and for which MLA is therefore the appropriate form of analysis. We draw largely on the model classifications used by Duncan et al. (1996) and Subramanian et al. (2003).

Strict Hierarchies: The Basic Model

We start off with the strict hierarchies. A lot of the theory and practice of multilevel modelling was developed in educational research in which the aim was to determine whether the shared environment of the school that pupils attended contributed to educational attainment, after adjusting for differences between schools in pupil characteristics (Aitken and Longford 1986; Goldstein 1986). From there it is not a big leap to consider a design of, for example, patients nested within hospitals (Fig. 4.1). The hierarchies have a pyramid structure with patients at the lower level (level one) nested within hospitals at the higher level (level two). The lowest level—the patient level in this example—is the level at which the outcome is measured. The reason for considering a multilevel model for these data is because the outcome for an individual patient may be influenced by the hospital that they attend or, in general, the shared context means that the patient outcomes may well be correlated, violating the standard regression assumption of independence. So whilst there is variability between patient outcomes, some of this variability may be due to differences between hospitals. The ability to partition variation into that attributable to different levels is an important feature of multilevel models. It is easy to think of examples of these basic models, whether they be patients in hospitals, survey respondents in residential neighbourhoods or GPs nested within practices.

We might have a three-level model in which the individuals at level one are the persons for whom we have measured a response (Fig. 4.2). These individuals are clustered within households at level two and then within neighbourhoods at level three. The idea of all of these strict hierarchies is that we have many units at one level nested within fewer units at the next level. Of course, the real world is not restricted to two or three levels and nor need our multilevel models be; the inclusion of relevant contexts may increase the number of levels that we need to consider. For example, in a study of diagnostic practice style in Alberta, Canada, Yiannakoulis et al. (2009) used a model including not only the individual physicians, for whom the outcome of diagnostic style was recorded, and the facilities in which they worked but also the municipality and census division—a strict hierarchy of four levels. And when



Fig. 4.1 Basic two-level model

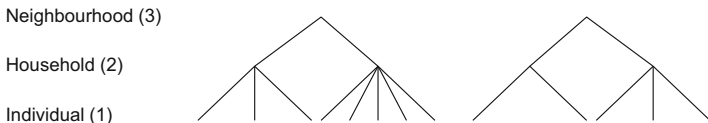


Fig. 4.2 Basic three-level model

exploring the consumption of tobacco in India, Subramanian et al. (2004) included household, local areas, districts and states as relevant contexts for the survey respondents in a five-level model.

It is important to note two features of these basic designs. Firstly, we do not need to have a balanced design. Our sample does not need to have the same number of patients in every hospital, or the same number of individuals in every household, or the same number of households in every neighbourhood. Secondly, the examples that we have discussed have the person as the lowest level, whether this is a patient, survey respondent or physician. Although this is a common occurrence, and there have been instances in previous chapters where we have referred to the individual and level one as though the two were synonymous, this need not necessarily be the case. For example, in a study of the variation in the use of drug-eluting stents (DESs) in the treatment of coronary heart disease in Scotland, Austin et al. (2008) took into account the fact that patients may have more than one lesion treated during a procedure by using lesions as the lowest level (the level at which the outcome, DES use, is measured) with these in turn nested within patients, operators and hospitals. The use of a multilevel model in this instance took into account the possible clustering of DES use within patients. And in periodontology, in a study of factors influencing the closure of pockets observed at different sites around teeth, Tomasi et al. (2007) used a hierarchy of sites within teeth within patients, patients forming the highest level in this analysis.

It may also be the case that data are not available at the individual level but rather are aggregated to an administrative area level. Such data restriction may reflect issues surrounding data confidentiality, whereby agencies are unwilling to release potentially identifiable individual data, or may just represent the constraints of official data systems. Cavalini and Ponce de Leon (2008) undertook an ecological analysis of the association between various socio-economic, political and healthcare indicators and differing morbidity and mortality outcomes in Brazil. With no data on individuals they used the levels of municipality, region and state; the outcomes were all measured at municipality level. No matter whether the data we have refer to individuals, aggregations of individuals or are collected within individuals, the lowest level is always the level at which the outcome is measured.

Multistage Sampling Designs

For a multistage sampling design, the hierarchy is imposed during data collection. The structure of the survey dictates the hierarchical design and straight away this implies that MLA is necessary. If the survey design is a simple random sample, individuals are selected from a sampling frame (for example, from a population register or hospital discharge register). In a two-stage sample high level sampling units are first selected, perhaps towns or municipalities, and then within each high level unit a sample of individuals is drawn. Individuals are nested within the higher-

level sampling units, and this nesting must be taken into account because of the potential for contextual influences on any outcomes. The data hierarchy will appear similar to those seen in Figs. 4.1 and 4.2. An example of such a design is the health interview survey in Belgium, as described by Demarest et al. (2013).

The primary reason for using multistage sampling is usually related to cost. It may be considerably cheaper to send interviewers to conduct several interviews within selected municipalities than conducting single interviews across a number of municipalities. Statistical methods were developed to permit the analysis of data collected from multistage samples; relatively simple sandwich estimators can be used which correct the standard error of the estimates to take the clustered sample design into account (Froot 1989). As described in Chap. 3, one effect of a multilevel data structure is to reduce the effective sample size which will in turn increase standard errors and confidence intervals. We return to the impact of clustering on power calculations in Chap. 6. The use of techniques such as sandwich estimators assumes that the hierarchical data structure is a nuisance—something for which we must make allowances but in which we have no substantive interest. But this is an oversimplification and is rarely the case; social epidemiology as a discipline is built on such substantive interests as the reasons for variations in health between areas. This is where we can start to explore the role of composition—who lives in the areas—and the context, or what it is about the areas themselves that lead to differences in outcomes between areas. These issues are explored further in Chap. 7.

Evaluating Community Interventions and Cluster Randomised Trials

There are a number of reasons for conducting an intervention at the community level; that is, when the community (as opposed to the individual) is the unit of allocation or randomisation. These include the impossibility or impracticality of introducing the intervention at an individual level (for example, in the case of water fluoridation), the desire to avoid contamination between intervention and control subjects, or as a cheaper and non-stigmatising means of targeting higher risk groups (Leyland 2010). In health services research, a cluster randomised approach may be the only appropriate means of evaluating certain interventions such as those relating to organisational change (Campbell and Grimshaw 1998). But whatever the rationale underlying the design of the study, if the intervention is at the group level and outcomes are measured at the individual level, then the data are hierarchical and must be analysed using MLA (Koepsell et al. 1992). Sample size or power calculations for cluster randomised trials differ from those for standard trials and are covered in Chap. 6.

Designs Including Time

We can think of two different types of designs including time: repeated cross-sections and repeated measures or panel data (Duncan et al. 1996; Subramanian et al. 2003). A repeated cross-sectional design might be used as a means of assessing hospital performance and how that changes over time. In such a case the hospitals form the highest level, and within each hospital every year data are collected relating to patient outcomes as a measure of that hospital’s performance. The ambition is to use these data to learn how each hospital performs in comparison to its peers and how the performance of each hospital is changing over time. Since the outcomes are at the patient level, the patient forms the lowest level in the hierarchy. Figure 4.3 shows the nesting of patients within years, and years within hospitals, in a three-level model. Dee (2001) used a repeated cross-sectional design to investigate the impact of (economic) cyclical state-level income effects on individual alcohol consumption through the study of repeated cross-sectional surveys of individuals nested within states of the USA. As with previous models we have no requirement for a perfectly balanced data set and so there is no need for our samples to include the same number of patients every year. Moreover, we can include hospitals for which we do not have data in every year. This will come as a relief to those familiar with the changing patterns of health provision and the idea that hospitals may close or open during a period of data collection.

The repeated measures or panel design is similar to the repeated cross-sectional design except that the same individuals are observed on different occasions. This means that the outcome is not measured at the level of the individual but at the level of the measurement occasion nested within the individual. The outcome still refers to the individual but may differ from one moment in time to another. Figure 4.4 illustrates a study in which outcomes on individuals are assessed on an annual basis and, in this example, the individuals themselves are clustered within neighbourhoods. This means that we can analyse longitudinal data in a multilevel

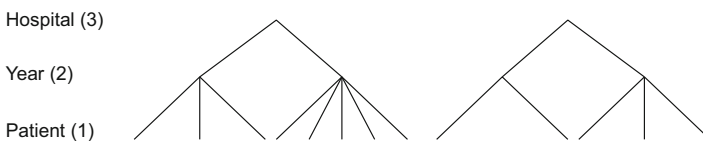


Fig. 4.3 Repeated cross-sectional design

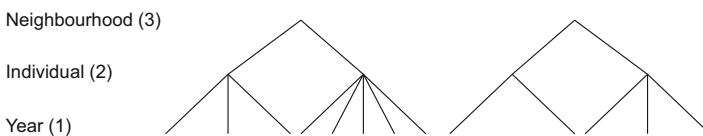


Fig. 4.4 Repeated measures or panel design

framework by taking into account the fact that measurement occasions are nested within individuals. In addition to any correlations that may exist between individuals within their contexts (hospitals, neighbourhoods, etc.), this design allows for the correlation between observations made on the same individual.

Haynes et al. (2008) looked at the risk of accidents in pre-school children using data from a longitudinal study, with measurement occasions nested within children and children nested within neighbourhoods. It is not necessary for individuals to be clustered within higher-level units; MLA can still be used to analyse repeated measures with individuals forming the higher level. Such a two-level model for changes in body mass index was used by Lipps and Moreau-Gruet (2010). Repeated measures do not have to be made on individuals; Kroneman and Siegers (2004) considered how reductions in the number of available hospital beds affected different measures of bed use using repeated measures on countries, with the outcomes (bed occupancy, average length of stay and admission rates) being observed in different years for each country. The example used in the first computing practical (Chap. 11) is based on the analysis of repeated measures of mortality rates made at the area level.

As with the previous models, it is not necessary to have information on every individual on every occasion; if we are able to make certain assumptions about missingness (that the data are *missing completely at random* or *missing at random*), then we can include individuals with incomplete data in the analysis. More detail about the different types of missing data and appropriate methods for their analysis can be found elsewhere (Carpenter et al. 2006; Little and Rubin 2002; Sterne et al. 2009).

When analysing repeated measures data, it is usually the case that we find more variation between individuals than within individuals (between measurement occasions) and so, unlike the basic models considered above, a larger proportion of the total variation may be at higher levels. This is easy to understand if you consider, for example, a study with repeated measures of people's weight; there is likely to be much less variability in individual weight from one measurement occasion to another than there is between the weights of individuals in the population. Such is the nature of individual heterogeneity.

Multiple Responses

There are strong similarities between repeated measures and multiple response designs. In the former we measure the same item on individuals at a number of different measurement occasions; in the latter we measure a number of different items on individuals, often at the same measurement occasion. This can therefore be seen as a multilevel model—we have the different responses nested within each individual—and there may be a further level such as the neighbourhood of residence as illustrated in Fig. 4.5. The multiple responses may, for example, be drawn from a questionnaire focusing on health-related behaviours; a number of individuals may be surveyed about alcohol and tobacco consumption, diet and exercise. These behaviours may be correlated within individuals; high alcohol consumption may be

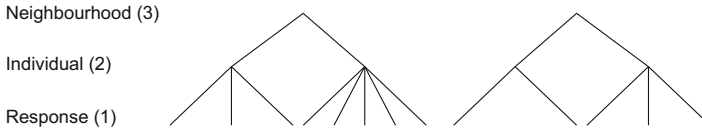


Fig. 4.5 Multiple responses

associated with poor diet, for example. This correlation may remain after adjustment for individual characteristics, particularly if an important characteristic associated with more than one behaviour is omitted or poorly recorded in the survey. But we also have the possibility of modelling and examining these correlations at higher levels. If alcohol consumption and diet both show variation between areas, is the nature of the relationship the same? That is, are those areas associated with above (below) average alcohol consumption also associated with poorer (better) diets?

Once again we can work with an unbalanced data set and so if some individuals have not responded to all questions, and provided that we can make the usual assumptions about the data being missing at random, we can include all the data that we have and do not have to consider the deletion of cases or responses. An example of a multiple response model includes a joint analysis of self-rated health and happiness on individuals nested within communities (Subramanian et al. 2005). In addition to showing the different effects of various socio-demographic variables on the two outcomes, the authors demonstrated a modest positive correlation at the individual level and a stronger positive correlation at area level, interpreting this as meaning that communities that were unhealthy were also likely to be unhappy.

It is possible to combine the analysis of different response types in a multilevel multiple response model; for example, we could include a continuous response such as blood pressure alongside a dichotomous response such as smoking status. The fact that there is no requirement for the data to be balanced or complete means that we can have structurally missing values: data which may or may not be collected depending on the response to another question. Duncan et al. (1996) looked at smoking behaviour among individuals living in areas (electoral wards) in England, considering two aspects of smoking: smoking status (whether an individual currently smoked or not) and the number of cigarettes smoked per day. For those who do not smoke the number of cigarettes smoked per day must be zero and can be ignored, removing a large peak in the (bimodal) distribution. Smoking status is therefore treated as a dichotomous outcome and the number of cigarettes smoked per day (among those who smoke) as a continuous measure. In addition to noting differences in the factors related to smoking status and cigarette consumption, the authors found a positive correlation between the two at the area level suggesting that cigarette consumption tends to be higher for individuals who live in areas in which people are more likely to smoke. A similar example is given in a study of the use of tranquilizers (benzodiazepines) in neighbourhoods in a Dutch city (Groenewegen et al. 1999). In this case the dichotomous outcome was whether or not people received a prescription and the dose of the drug, if given a prescription, was treated as a continuous response. Once again the model permitted not only the analysis of factors

associated with both prescription and dose but also the analysis of the relationship between these outcomes at the area level.

Any data showing an excessive number of observations at zero are amenable to these types of mixed response models. Tooze et al. (2002) considered a range of factors associated with medical expenditure based on a sample of individuals nested within households. They interpreted the strong positive correlation between the occurrence of healthcare expenditure (dichotomous) and the intensity of expenditure (continuous) as indicating that, after adjusting for any differences in covariates, households that were more likely to seek medical care were also likely to have greater healthcare expenditure.

Non-hierarchical Structures

The data structures that we have considered up to this point are all strict hierarchies; that is, a number of units at one level are nested within one and only one unit at the level above. The reality is that healthcare systems or the social contexts affecting individuals are often more complex than this, and if we have data that reflects this complexity then this leads to hierarchies that do not have such a neat structure. Below we discuss three types of non-hierarchical structures that can be fitted using MLA: *cross-classified models*, *multiple membership models* and *correlated cross-classified models*.

Cross-Classified Models

A cross-classified model is one in which units at one level are simultaneously nested within two separate, non-nested hierarchies (Goldstein 1994). For example, we may want to examine how the outcome for an individual patient varies according both to the hospital the patient attended and to the general practitioner (GP) that referred the patient to hospital. Figure 4.6 shows how the hierarchy may appear for such a model. Although all patients are referred by one and only one GP, and each attends one and only one hospital, there is no strict nesting of GPs within hospitals; certain GPs may refer different patients to different hospitals. Similarly, hospitals are not nested within GPs since hospitals receive referrals from several different GPs. We say in

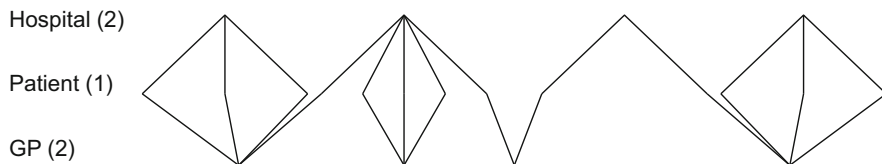


Fig. 4.6 Cross-classified model

such a case that patients are nested within a cross-classification of GPs and hospitals (Browne et al. 2001; Rasbash and Browne 2001). The way in which the computational aspects of fitting cross-classified models are handled varies according to the software used for analysis; some of the statistical packages used to fit multilevel models treat cross-classified models no differently from strict hierarchies, whilst other packages may require a distinct specification for this class of model. Readers are advised to check the reference manuals of their chosen software for further details.

As with the strictly hierarchical multilevel models, cross-classified models may be used to reflect the observed hierarchy (in which case the levels themselves may not be of substantive interest) or they may be used to explore variation and determine the relative importance of different contexts. This distinction relates to the range of hypotheses that can be tested using MLA discussed in Chap. 3. Downing et al. (2007) explored the association between deaths and hospital admissions for a range of conditions and scores assigned to GP practices through the UK’s Quality and Outcomes Framework (QOF). Their data comprised patients nested within a cross-classification of GPs and residential areas, with covariates available on both contexts. Urquia et al. (2009) considered the relative impacts of neighbourhood of residence and country of origin on the birthweight of children born to recent immigrants in Ontario, Canada, following adjustment for a variety of individual factors, and concluded that the country of origin made a much larger contribution to the variation in outcomes. Virtanen et al. (2010) separated the effects of teachers’ neighbourhood of residence and the neighbourhood in which the school was located on the sickness absences of teachers and found significant relationships with both (in terms of a contextual variable—mean neighbourhood income—and the variances at the two levels).

Multiple Membership Model

The second type of non-hierarchical structure used in MLA is the multiple membership model (Hill and Goldstein 1998). This model is appropriate when units at one level may belong to (or *be members of*) more than one unit at a higher level. For example, consider a patient who receives a course of treatment such as chemotherapy over a period of time. Certain patients may receive their treatment at more than one hospital as shown in Fig. 4.7. If the outcome for each patient is survival at 12 months, then we may be interested in determining whether patient survival varies between hospitals. For those patients who were treated in more than one hospital, we

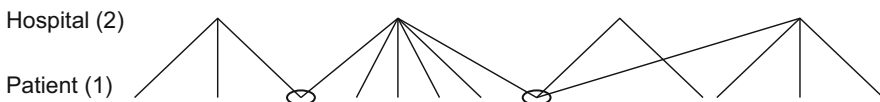


Fig. 4.7 Multiple membership model

must make assumptions about the relative contributions of different hospitals to the patients' care. This comes down to assigning a weight attributed to each hospital with the weights summing to one for each individual (so the weights are, in fact, proportions). If we know the proportion of time that a patient spent in each hospital, then these proportions may make suitable weights; otherwise, it may be sufficient to give equal weight to each hospital attended (so weights of 0.5 if a patient was seen in two hospitals, 0.33 if seen in three hospitals, etc.). The impact of different weighting schemes on the results can be examined as a form of sensitivity analysis.

Ryan et al. (2006) examined the influence of caseworkers on two child welfare outcomes: the length of stay in foster care and the probability of family reunification. Most youths in the study from Illinois were assigned more than one caseworker; multiple membership models allowed the authors to account for the complex data structure when testing hypotheses about the association of certain key caseworker characteristics on the child outcomes. Another use for a multiple membership model is to account for changes in geographical boundaries over the course of time; Leyland (2004) assigned weights based on resident populations to take account of changes in the number and boundaries of areas following administrative restructuring. Falster et al. (2018) used a multiple membership model to analyse the between-hospital variation in patient admission for preventable hospitalisations. Although the hospital of admission was known for those patients who were admitted to hospital, the population who were not admitted to any hospital were assigned to multiple hospitals based on observed admission patterns.

Correlated Cross-Classified Model

The correlated cross-classified model should be used for the analysis of repeated classifications (Leyland and Næss 2009). Such data structures are typically encountered when contextual information at regular intervals is linked to an outcome measured at the end of the study, although they may also be appropriate when different aspects of the same context are being measured such as place of residence and place of work. Figure 4.8 provides a simple example of individuals living in four areas at two different time points. The difference between this model and the cross-classified model (Fig. 4.6) is that instead of independent contexts such as GP and

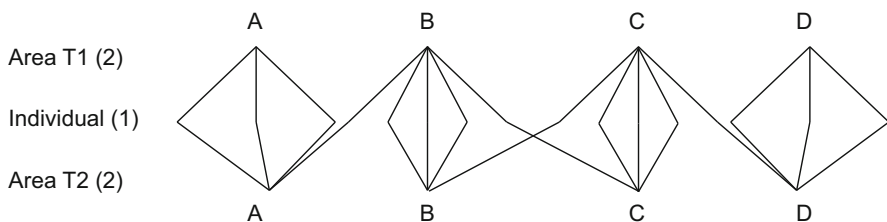


Fig. 4.8 Correlated cross-classified model

hospital, the areas are the same at each time (denoted areas A, B, C, and D). One of the assumptions underlying MLA is that the contexts are independent, whether these are the GPs and hospitals in Fig. 4.6 or the neighbourhood and households in Fig. 4.2. Standard multilevel models, including the cross-classified model, therefore assume no correlation between contexts. The multiple membership model described above is appropriate when individuals move between contexts but the contexts (e.g. areas) are the same at different points in time. The correlated cross-classified model comes somewhere between the cross-classified and multiple membership models, recognising that contexts may not be identical (due, for example, to the way neighbourhoods may change over time) but at the same time that the contexts are not completely independent of each other (the poorest area at one time point is unlikely to become the richest area at another time).

The cross-classified, multiple membership and correlated cross-classified models are described and the implications of the different assumptions underlying each are analysed from the perspective of life course epidemiology by Næss and Leyland (2010).

An example of the use of a correlated cross-classified multilevel model is based on analysis of the Oslo Mortality Study (Leyland and Næss 2009). Area of residence was known for inhabitants of Oslo at the time of the 1960, 1970, 1980 and 1990 Censuses and individuals were followed up in the mortality register until 1998. The models were used to determine the relative contribution of residence at different stages of the life course—based on known residence at the Censuses—on subsequent mortality for different birth cohorts.

Other Multilevel Models

There is a broad range of data types that can be analysed using MLA and of models that can be constructed in a multilevel framework. Some of these are dependent on the availability of specialist software, whilst others may be implemented in most packages that can be used for multilevel modelling. In this section, we briefly describe some of these models.

We have said little about the response types that can be analysed using MLA, but most of the examples presented in this chapter have assumed continuous outcomes to be normally distributed or have used logistic regression for dichotomous outcomes. Multilevel Poisson or negative binomial regression models may be used when the data take the form of counts, either because individual data are aggregated to an area level in studies of disease incidence or prevalence (Cavalini and Ponce de Leon 2008) or when the data represent counts made on individuals, such as the number of carious, extracted or filled teeth (Levin et al. 2010) or the frequency of contact with GPs (Cardol et al. 2005). Multilevel Poisson regression is also appropriate for modelling incidence or prevalence on individual data as a means of adjusting for exposure or person time at risk (Martikainen et al. 2003). Multilevel logistic regression can easily be extended to multilevel multinomial

regression if the responses form unordered categories, such as place of birth being categorised as home, private hospital or public hospital in a study of maternity care provision in Ghana (Amoako Johnson and Padmadas 2009), or ordered categories, such as a measure of self-rated health (Oshio and Kobayashi 2009). Note, however, that in the presence of five or more ordered categories it may be appropriate to analyse the data as though the response was continuous and normally distributed (Mansyur et al. 2008).

Several different models have been developed for the analysis of multilevel data when the outcome of interest is the time to an event or a survival time. The simplest of these is the accelerated lifetime or log duration model, which centres on modelling the logarithm of the survival time. Such a model has been used to assess area-based inequalities in a 30-year follow-up of a large Swedish cohort (Yang et al. 2009). An alternative approach is to fit multilevel Cox proportional hazard models; these have been used, for example, to examine contextual influences on the hazard of mortality (Chaix et al. 2007). Such models have the advantage of providing answers even if a large proportion of the data are censored and of enabling the inclusion of time-varying covariates (Goldstein 2003). For example, Sear et al. (2000) examined the effect of maternal grandmothers on the survival of children in rural Gambia; the presence of the grandmother is clearly an effect which may change during a child's life. Multilevel Cox regression models require data expansion that can quickly render a dataset large and unwieldy; an alternative approach is therefore to use multilevel Weibull survival models, as employed by Chaix et al. (2008) to examine the impact of individual perception of safety and neighbourhood cohesion on mortality from acute myocardial infarction.

A multilevel repeated measures model takes into account the fact that observations made on the same individual are likely to be correlated. A *time series* model can take this one stage further by modelling the correlation between observations as a function of time such that the correlation between two measures made on the same person close together in time will be higher than the correlation between two measures made a long time apart. There are a number of different ways in which this correlation can be included (Goldstein et al. 1994). An example of the application of such methods is for the analysis of smoking cessation data in which adjustment was made for the serial dependence of observations on individuals' smoking status (Wang et al. 2006).

A similar principle applies to multilevel spatial models as to the multilevel time series models. It is possible to take geography into account to some extent by using a series of areas of increasing size. This relates to the so-called 'modifiable areal unit problem' or MAUP (Openshaw 1984). Geographical units are to some extent artificial and changing from one geographical division to another might influence the results of a study. MLA facilitates a meaningful analysis of this problem (Groenewegen et al. 1999; Jones 1993; Merlo 2011). Some of the difference between small areas (such as neighbourhoods) may be attributable to differences between larger areas such as municipalities, and the differences between municipalities may in part be due to differences between larger areas such as counties or regions. Including these different geographies in a single multilevel model ensures that

there is a correlation between neighbourhoods in the same municipality and between municipalities in the same county. But this ignores the detail in the geography; the exact geographical positioning of neighbourhoods within a municipality or of municipalities within a county is not taken into account. A spatial multilevel model allows for a greater degree of correlation between areas that are geographically close than between areas that are geographically distant. A simple means of fitting such spatial dependencies is to use a *multiple membership model* (see above) in which, in addition to heterogeneous area effects, areas are modelled as multiple members of the set of their neighbours. Bartolomeo et al. (2010) used such a model to investigate the geographical patterning of hospitalisations for lung cancer and chronic obstructive pulmonary disease. Spatial modelling will also provide geographically smoothed estimates, overcoming some of the problems associated with small areas and rare outcomes leading to volatile rates and allowing the identification of clusters of disease. The methodology underlying such modelling may be complex and is described in detail elsewhere (Best et al. 2005; Lawson et al. 2003; Leyland and Davies 2005). Næss et al. (2007) used a spatial multilevel model to separate the effect of air pollution from that of social deprivation, both measured at the neighbourhood level, on individual mortality following adjustment for individual socio-economic status.

Other data which lend themselves to multilevel analysis include meta-analysis, for example a meta-analysis of the results of several clinical trials. The idea of meta-analysis is to combine information from separate studies. A fixed effects approach to meta-analysis is based on the assumption that there is a single 'true' effect which is observed with error in each study. The random effects or multilevel approach to meta-analysis assumes that there is heterogeneity between studies in the effect size. Published information on the original trials will often be extremely limited; for example, a randomised controlled trial may report the numbers of deaths and total number of patients in the treatment and control wings of a trial. In such circumstances, and if the original data cannot be made available, it is important to take into account the precision of the estimate of the effect size by giving more weight to larger studies. It is also possible to combine summary outcomes from trials with complete data on individuals from those trials for which full individual data are available or to combine trial data with observational data. Examples of multilevel meta-analyses include a study of the effectiveness of interventions to promote advance directives (such as living wills and durable power of attorney for healthcare) among the elderly (Bravo et al. 2008) and a quantification of the effects of education on self-reported health (Furnée et al. 2008).

Multilevel models have been extended to include factor analysis, latent class analysis and structural equation models. These expand upon their single-level counterparts to take into account the clustering of individuals within higher-level units. For example, Franzini et al. (2005) used multilevel structural equation models to investigate whether latent variables such as collective efficacy (comprising social cohesion, trust and helpfulness) or neighbourhood disorder (comprising physical and social disorder) mediated the relationship between neighbourhood impoverishment and self-rated health after adjusting for individual characteristics. Curry et al. (2008)

used multilevel path analysis to determine whether objectively measured neighbourhood crime rates impacted directly on individual depression or whether the impact was indirect, being mediated by subjective perceptions of neighbourhood problems. And Vermunt (2007) identified three classes of doctors and two classes of hospital on the basis of their prescribing behaviour when treating children with acute respiratory tract infection; responses for individual children were coded as indicating appropriate use, abuse of a single antibiotic or abuse of multiple antibiotics.

Multilevel latent variable analysis will be considered more extensively in Chap. 8. The reason for this is that this approach is increasingly used to construct characteristics of higher-level units on the basis of individual responses to a series of scale items. These scale items try to measure a latent variable at the higher level. For example, items about neighbourhood disorder, collected from residents in a survey, can collectively be used to indicate disorder at the neighbourhood level. This approach is also known as *ecometrics*.

Pseudo-levels

In Chap. 3, we considered what constitutes a level. In particular, we made a distinction between a level—comprising units which could be sampled—and the characteristics of a level. Although this is true in the strictest sense, it is sometimes useful to introduce characteristics as a *pseudo-level* at any level apart from the highest level in the hierarchy. This is particularly important if we want to test hypotheses about (or just to explore) variation between subgroups, as was discussed in Chap. 3. For example, suppose we have health data on a number of individuals attending different hospitals, and one focus of our interest is whether the variance in our outcome differs between men and women. Although the individual's sex is a characteristic of the individual and not a level, we can include sex as a pseudo-level in our model so that patients are nested within sex within hospitals, and then condition on the mean difference between men and women. (*Conditioning on the mean* means that we include a dummy variable to take account of the mean difference in health between men and women. This dummy variable is then a characteristic of the pseudo-level rather than the individual level since it applies to all individuals within that group.) Figure 4.9 shows how the inclusion of this pseudo-level changes the structure of our dataset. The groups at the pseudo-level are often referred to as *cells*, and sometimes individual responses are aggregated over these

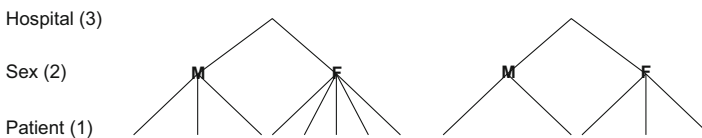


Fig. 4.9 Model with pseudo-levels

cells which then form the lowest level. For example, Judge et al. (2009) examined the rates of joint replacement in England using a hierarchy of cells defined by 5-year age group and sex (at level 1) nested within small areas (at level 2) and districts (at level 3). For each cell they had a count of the number of procedures undertaken and included an offset to adjust for differences in the population at risk in each cell whilst controlling for age and sex. And Turrell et al. (2007) investigated associations between area deprivation and mortality using cells defined by a combination of age, sex and individual occupational social class nested within a hierarchy of areas.

Incomplete Hierarchies

In general, we know to which unit at each higher level a lower-level unit belongs and so we have complete information on the hierarchy. There are two notable exceptions when this will not be the case. The first exception concerns multiple responses; the hierarchies may differ for different responses. This may be because the responses are actually measured at different levels. Goldstein gives an example of a multiple response model combining longitudinal measures (during childhood) of height and bone age with a measure of adult height (Goldstein 2003). Whilst the repeated measures during childhood are clustered within the individual, the one adult measurement is effectively at the level of individual rather than measurement occasion. The hierarchy may vary according to the number in each cluster. Dundas et al. (2014) give an example of individual children nested within sibling groups living in small areas; sibling group was omitted as a level for the 71% of children who had no siblings in the study. Alternatively, the *structured missingness* detailed under the earlier section on multiple response models may lead to differing hierarchies; Leyland and Boddy (1998) describe a model of mortality following acute myocardial infarction in which they consider the influences of both area of residence and hospital attended. Their data include both sudden deaths (death before reaching hospital) and deaths in hospital or within 30 days of discharge from hospital. These two responses (sudden death and death in or shortly after discharge from hospital) were nested within patients. The sudden deaths are clearly not affected by hospital attended; indeed, for such deaths there is no hospital attended. The second exception is when the higher-level membership is unknown. In such a situation, it is possible to use a multiple membership model with different probabilities of membership attached to the higher-level units (Hill and Goldstein 1998). Each higher-level unit (e.g. hospital) could be given equal weight or weight proportional to the total number of patients seen by that hospital in the absence of any knowledge as to group membership. However, it may be that more detailed information is available and that the precise membership of higher-level units is only partially missing; for example, it may be that a patient living in a given area is most likely to attend one of a number of local facilities.

A slightly different situation may arise when two levels are indistinguishable. Figure 4.2 illustrates a hierarchy that includes individuals nested within households.

In general, there will not be many individuals per household and many households may only contain one person. To an extent this does not matter; as long as there is at least one household comprising two or more people, then we can start to describe variation within households as well as between households. (In practice, the more households in the study in which there are at least two people, the more precise our estimate of the variance within households will be.) And clearly excluding single person households from our analysis is likely to introduce considerable bias into our sample. But our sample design may have included just one person in each household. In such a case, although it is correct to think of individuals as being nested within households, we are unable to distinguish between the individual and household levels. Not really a missing hierarchy, we are forced in practice to work with a joint individual/household level.

Conclusion

This chapter has introduced the reader to a variety of structures that can be thought of as multilevel or hierarchical. In addition to the strict hierarchies that perhaps constitute the common understanding of a multilevel model, we have discussed the appropriateness of multilevel modelling for designs including time, multiple responses and non-hierarchical structures. Furthermore, we have covered the concept of a pseudo-level and circumstances in which the unit of membership at a particular level may be missing.

When working out the data structure in your own research, it is important to bear in mind what has been said in Chaps. 2 and 3. The first step would be to analyse your research problem and specify which levels would be relevant to include from a theoretical perspective. You might end up working with data that are readily available, and the structure of these data might differ from what you would have wanted based on an analysis of your research problem. Of particular importance is whether you are missing information about a level in your data that seems to be important from a theoretical point of view. If this is the case, then your statistical model may be misspecified as a consequence. An example of this, which is discussed in more detail in Chap. 7, is the situation where you consider a health outcome of people living in neighbourhoods but omit the fact that your subjects are also clustered in families or households. This would lead to an overestimation of individual level or neighbourhood variation or both; see, for example, Sacker et al. (2006).

Some data structures may be quite complex, especially since the structures that have been discussed in this chapter can be combined. The more complicated the data structures are, the more difficult they will be to analyse and interpret. For readers who are keen to work with more complex data structures, we offer two pieces of advice. Firstly, we suggest that you simplify the data structure into a less complex, simple hierarchical structure and analyse the data in this manner before proceeding. In Chap. 9, we discuss ways of simplifying data structures as part of the modelling

process. Our second piece of advice in the event of more complex data structures is to consult a colleague with experience in running and interpreting the analysis or to read some of the more technical multilevel modelling texts to gain further understanding of such analyses (for example, De Leeuw and Meijer 2008; Gelman and Hill 2007; Goldstein 2010; Snijders and Bosker 2012).

References

- Aitken M, Longford N (1986) Statistical modelling in school effectiveness studies. *J R Stat Soc Ser A* 149:1–43
- Amoako Johnson F, Padmadas SS (2009) On the spatial inequalities of institutional versus home births in Ghana: a multilevel analysis. *J Community Health* 34:64–72
- Austin D, Oldroyd KG, McConnachie A, Slack R, Eteiba H, Flapan AD, Jennings KP, Northcote RJ, Pell ACH, Starkey IR, Pell JP (2008) Hospital and operator variations in drug-eluting stent use: a multi-level analysis of 5967 consecutive patients in Scotland. *J Public Health* 30:186–193
- Bartolomeo N, Trerotoli P, Serio G (2010) Lung cancer and COPD rates in Apulia: a multilevel multimember model for smoothing disease mapping. *Int J Health Geogr* 9:15
- Best N, Richardson S, Thomson A (2005) A comparison of Bayesian spatial models for disease mapping. *Stat Methods Med Res* 14:35–59
- Bravo G, Dubois M-F, Wagneur B (2008) Assessing the effectiveness of interventions to promote advance directives among older adults: a systematic review and multi-level analysis. *Soc Sci Med* 67:1122–1132
- Browne WJ, Goldstein H, Rasbash J (2001) Multiple membership multiple classification models. *Stat Model* 1:103–124
- Campbell MK, Grimshaw JM (1998) Cluster randomised trials: time for improvement. *Br Med J* 317:1171–1172
- Cardol M, Groenewegen PP, de Bakker DH, Spreuwenberg P, van Dijk L, van den Bosch W (2005) Shared help seeking behaviour within families: a retrospective cohort study. *Br Med J* 330:882–884
- Carpenter JR, Kenward MG, Vansteelandt S (2006) A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *J R Stat Soc Ser A* 169:571–584
- Cavalini LT, Ponce de Leon ACM (2008) Morbidity and mortality in Brazilian municipalities: a multilevel study of the association between socioeconomic and healthcare indicators. *Int J Epidemiol* 37:775–785
- Chaix B, Rosvall M, Merlo J (2007) Assessment of the magnitude of geographical variations and socioeconomic contextual effects on ischaemic heart disease mortality: a multilevel survival analysis of a large Swedish cohort. *J Epidemiol Community Health* 61:349–355
- Chaix B, Lindström M, Rosvall M, Merlo J (2008) Neighbourhood social interactions and risk of acute myocardial infarction. *J Epidemiol Community Health* 62:62–68
- Curry A, Latkin C, Davey-Rothwell M (2008) Pathways to depression: the impact of neighborhood violent crime on inner-city residents in Baltimore, Maryland, USA. *Soc Sci Med* 67:23–30
- De Leeuw J, Meijer E (eds) (2008) *Handbook of multilevel analysis*. Springer, New York
- Dee TS (2001) Alcohol abuse and economic conditions: evidence from repeated cross-sections of individual-level data. *Health Econ* 10:257–270
- Demarest S, Van der Heyden J, Charafeddine R, Drieskens S, Gisle L, Tafforeau J (2013) Methodological basics and evolution of the Belgian health interview survey 1997–2008. *Arch Public Health* 71:24
- Downing A, Rudge G, Cheng Y, Tu Y-K, Keen J, Gilthorpe MS (2007) Do the UK government's new Quality and Outcomes Framework (QOF) scores adequately measure primary care performance? A cross-sectional survey of routine healthcare data. *BMC Health Serv Res* 7:166

- Duncan C, Jones K, Moon G (1996) Health-related behaviour in context: a multilevel modelling approach. *Soc Sci Med* 42:817–830
- Dundas R, Leyland AH, Macintyre S (2014) Early-life school, neighborhood, and family influences on adult health: a multilevel cross-classified analysis of the Aberdeen Children of the 1950s Study. *Am J Epidemiol* 180:197–207
- Falster MO, Jorm LR, Leyland AH (2018) Using weighted hospital service area networks to explore variation in preventable hospitalization. *Health Serv Res* 53(Suppl 1):3148–3169
- Franzini L, Caughy M, Spears W, Esquer MEF (2005) Neighbourhood economic conditions, social processes, and self-rated health in low-income neighborhoods in Texas: a multilevel latent variables model. *Soc Sci Med* 61:1135–1150
- Froot KA (1989) Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in financial data. *J Financ Quant Anal* 24:333–355
- Furnée CA, Groot W, Maassen van den Brink H (2008) The health effects of education: a meta-analysis. *Eur J Pub Health* 18:417–421
- Gelman A, Hill J (2007) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge
- Goldstein H (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73:43–56
- Goldstein H (1994) Multilevel cross-classified models. *Sociol Methods Res* 22:364–375
- Goldstein H (2003) *Multilevel statistical models*, 3rd edn. Arnold, London
- Goldstein H (2010) *Multilevel statistical models*, 4th edn. Wiley, Chichester
- Goldstein H, Healy MJR, Rasbash J (1994) Multilevel time series models with applications to repeated measures data. *Stat Med* 13:1643–1655
- Groenewegen PP, Leufkens HG, Spreeuwenberg P, Worm W (1999) Neighbourhood characteristics and use of benzodiazepines in The Netherlands. *Soc Sci Med* 48:1701–1711
- Haynes R, Jones AP, Reading R, Daras K, Emond A (2008) Neighbourhood variations in child accidents and related child and maternal characteristics: does area definition make a difference? *Health Place* 14:693–701
- Hill PW, Goldstein H (1998) Multilevel modeling of educational data with cross-classification and missing identification for units. *J Educ Behav Stat* 23:117–128
- Jones K (1993) ‘Everywhere is nowhere’: multilevel perspectives on the importance of place. The University of Portsmouth Inaugural Lectures, Portsmouth
- Judge A, Welton NJ, Sandhu J, Ben-Shlomo Y (2009) Geographical variation in the provision of elective primary hip and knee replacement: the role of socio-demographic, hospital and distance variables. *J Public Health* 31:413–422
- Koepsell TD, Wagner EH, Cheadle AC, Patrick DL, Martin DC, Diehr PH, Perrin EB (1992) Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annu Rev Public Health* 13:31–57
- Kroneman M, Siegers JJ (2004) The effect of hospital bed reduction on the use of beds: a comparative study of 10 European countries. *Soc Sci Med* 59:1731–1740
- Lawson AB, Browne WJ, Vidal Rodeiro CL (2003) *Disease mapping with WinBUGS and MLwiN*. Wiley, Chichester
- Levin KA, Davies CA, Douglas GVA, Pitts NB (2010) Urban-rural differences in dental caries of 5-year old children in Scotland. *Soc Sci Med* 71:2020–2027
- Leyland AH (2004) Increasing inequalities in premature mortality in Great Britain. *J Epidemiol Community Health* 58:296–302
- Leyland AH (2010) Methodological challenges in the evaluation of community interventions. *Eur J Pub Health* 20:242–243
- Leyland AH, Boddy FA (1998) League tables and acute myocardial infarction. *Lancet* 351:555–558
- Leyland AH, Davies CA (2005) Empirical Bayes methods for disease mapping. *Stat Methods Med Res* 14:17–34
- Leyland AH, Næss Ø (2009) The effect of area of residence over the life course on subsequent mortality. *J R Stat Soc Ser A* 172:555–578

- Lipps O, Moreau-Gruet F (2010) Change of individual BMI in Switzerland and the USA: a multilevel model for growth. *Int J Public Health* 55:299–306
- Little RJ, Rubin DB (2002) *Statistical analysis with missing data*, 2nd edn. Wiley, New York
- Mansyur C, Amick BC, Harrist RB, Franzini L (2008) Social capital, income inequality, and self-rated health in 45 countries. *Soc Sci Med* 66:43–56
- Martikainen P, Kauppinen TM, Valkonen T (2003) Effects of the characteristics of neighbourhoods and the characteristics of people on cause specific mortality: a register based follow up study of 252 000 men. *J Epidemiol Community Health* 57:210–217
- Merlo J (2011) Contextual influences on the individual life course: building a research framework for social epidemiology. *Psychosoc Interv* 20:109–118
- Næss Ø, Leyland AH (2010) Analysing the effect of area of residence over the life course in multilevel epidemiology. *Scand J Public Health* 38:119–126
- Næss Ø, Piro FN, Nafstad P, Davey Smith G, Leyland AH (2007) Air pollution, social deprivation, and mortality. A multilevel cohort study. *Epidemiology* 18:686–694
- Openshaw S (1984) *The modifiable areal unit problem*. Geo Books, 38, Norwich
- Oshio T, Kobayashi M (2009) Income inequality, area-level poverty, perceived aversion to inequality, and self-rated health in Japan. *Soc Sci Med* 69:317–326
- Rasbash J, Browne W (2001) Modelling non-hierarchical structures. In: Leyland AH, Goldstein H (eds) *Multilevel modelling of health statistics*. Wiley, Chichester
- Ryan JP, Garnier P, Zyphur M, Zhai F (2006) Investigating the effects of caseworker characteristics in child welfare. *Child Youth Serv Rev* 28:993–1006
- Sacker A, Wiggins RD, Bartley M (2006) Time and place: putting individual health into context. A multilevel analysis of the British household panel survey, 1991–2001. *Health Place* 12:279–290
- Sear R, Mace R, McGregor IA (2000) Maternal grandmothers improve nutritional status and survival of children in rural Gambia. *Proc R Soc B* 267:1641–1647
- Snijders TAB, Bosker RJ (2012) *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage, Los Angeles
- Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Br Med J* 338:b2393
- Subramanian SV, Jones K, Duncan C (2003) Multilevel methods for public health research. In: Kawachi I, Berkman LF (eds) *Neighborhoods and health*. Oxford University Press, New York
- Subramanian SV, Nandy S, Kelly M, Gordon D, Davey Smith G (2004) Patterns and distribution of tobacco consumption in India: cross sectional multilevel evidence from the 1998–9 national family health study. *Br Med J* 328:801–806
- Subramanian SV, Kim D, Kawachi I (2005) Covariation in the socioeconomic determinants of self rated health and happiness: a multivariate multilevel analysis of individuals and communities in the USA. *J Epidemiol Community Health* 59:664–669
- Tomasi C, Leyland AH, Wennström JL (2007) Factors influencing the outcome of non-surgical periodontal treatment: a multilevel approach. *J Clin Periodontol* 34:682–690
- Tooze JA, Grunwald GK, Jones RH (2002) Analysis of repeated measures data with clumping at zero. *Stat Methods Med Res* 11:341–355
- Turrell G, Kavanagh A, Draper G, Subramanian SV (2007) Do places affect the probability of death in Australia? A multilevel study of area-level disadvantage, individual-level socioeconomic position and all-cause mortality, 1998–2000. *J Epidemiol Community Health* 61:13–19
- Urquia ML, Frank JW, Glazier RH, Moineddin R, Matheson FI, Gagnon AJ (2009) Neighborhood context and infant birthweight among recent immigrant mothers: a multilevel analysis. *Am J Public Health* 99:285–293
- Vermunt JK (2007) Latent class and finite mixture models for multilevel data sets. *Stat Methods Med Res* 17:33–51
- Virtanen M, Kivimäki M, Pentti J, Oksanen T, Ahola K, Linna A, Kouvonen A, Salo P, Vahtera J (2010) School neighborhood disadvantage as a predictor of long-term sick leave among teachers: prospective cohort study. *Am J Epidemiol* 171:785–792

- Wang K, Lee AH, Hamilton G, Yau KKW (2006) Multilevel logistic regression modelling with correlated random effects: application to the Smoking Cessation for Youth study. *Stat Med* 25:3864–3876
- Yang M, Eldridge S, Merlo J (2009) Multilevel survival analysis of health inequalities in life expectancy. *Int J Equity Health* 8:31
- Yiannakoulis N, Hill MD, Svenson LW (2009) Geographic hierarchies of diagnostic practice style in cerebrovascular disease. *Soc Sci Med* 68:1985–1992

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

