correlation and variance components – are of major interest. The design then has to ensure that the standard errors for these main parameters are small enough.

To make an a priori assessment of standard errors of estimation for various parameters, one has to determine, or guess, the values of parameters such as variances and covariances of outcome and explanatory variables for each relevant level in the design. This was illustrated above by some examples. Sometimes such guesses can reasonably be made on the basis of existing data; otherwise, it is important to conduct a sensitivity analysis by varying the guessed parameter values and studying how this affects the standard errors of interest.

Given the complexity of design considerations for multilevel studies, it is often advisable to reduce the problem to its simplest form, ignoring control variables for which a minor impact is expected, and start with a random intercept model. If there are more than two levels involved and it is possible to point out the higher level that is expected to be associated with the largest random variability, then it may be advisable to ignore temporarily the other levels and start with a power analysis for a two-level model. If such a simplified first analysis has provided some rough insight into the effects of various combinations of sample sizes at the various levels on standard errors and/or power, and if there is sufficient information to make guesses about the values of additional parameters, then in a further analysis one could enter a random slope in the design considerations, and perhaps more than two levels. The general rule is to start design considerations as simply as possible, because if one tries to face the complexity right away one runs the risk of being put off and cancelling the a priori design considerations altogether – and wouldn't that be a pity?

CHAPTER 12

Further Topics in Multilevel Modelling

Harvey Goldstein
Mathematical Sciences, Institute of Education, University of London, UK

Alastair H. Leyland
MRC Social and Public Health Sciences Unit, University of Glasgow, UK

12.1 INTRODUCTION

This chapter covers further topics in multilevel modelling that will be of particular interest to those involved in research in the health sciences. The following sections therefore illustrate how the theory and methods underlying multilevel models may be extended to include meta-analysis, survival data modelling, and the ideas of context and composition. This is followed by a brief discussion of recent developments in multilevel modelling.

12.2 META-ANALYSIS

The purpose of meta-analysis is to provide an overall summary of results when information from several studies of the same topic are available. These 'studies' may be centres in a single clinical trial, distinct experimental studies, distinct (or possibly overlapping) observational surveys, or mixtures of these. Meta-analysis can therefore be regarded as a special case of the general hierarchical data model, where individual observations are nested within studies or centres. Viewing meta-analysis within this framework leads to some important and natural extensions.

In applied work, it is often assumed that the effect of interest is constant across the component studies (Thompson and Pocock, 1991), yielding the so-called 'fixed effect' model. The assumption of homogeneity can, however, be relaxed to allow for random variation between studies of the effects, yielding the so-called 'random effects' model (DerSimonian and Laird, 1986). Statistical

Multilevel Modelling of Health Statistics Edited by A.H. Leyland and H. Goldstein

models for this case can be fitted using a variance components multilevel model formulation. A general multilevel formulation (Goldstein, 1995), however, allows more general random coefficient models to be studied, and we describe this in more detail below. A straightforward extension is to include covariates in such a model and to observe the extent to which they account for between-study variation. An additional problem is when some studies provide individual-level data, while for others only summary results (such as means) are available and methods of meta-analysis that can combine such results efficiently are now available (Goldstein *et al.*, 2000).

For aggregate-level data, consider the following underlying model for individual-level data, for example a measure of attitude towards health education in schools where we have pupils grouped within studies with a treatment group that has been exposed to health education and a control group that has not. Suppose that we have a basic model, with the response Y_{ij} being the attitude score (suitably transformed to normality) for the ith pupil in the jth study, as

$$\left.\begin{aligned} y_{h,ij} &= \beta_0 + \beta_1 x_{ij} + \beta_2 t_{h,ij} + u_{h,j} + e_{h,ij}, \\ \mathrm{var}(u_{h,j}) &= \sigma^2_{hu}, \quad \mathrm{var}(e_{h,ij}) = \sigma^2_{he}, \end{aligned}\right\} \tag{12.1}$$

with the usual assumptions of normality and independence. The term x_{ij} is a covariate, in this case a baseline pretreatment measure of attitude. The subscript h indexes the treatment/control and the term $t_{h,ij}$ is 1 if treatment and 0 if control. The random effect $u_{h,j}$ is a study effect and the $e_{h,ij}$ are individual-level residuals. Clearly this model can be elaborated in a number of ways, by including further covariates at study or individual level, by allowing β_2 (or β_1) to vary at level 2 so that the effect of treatment varies across studies, and by allowing the level-1 variance to depend on other factors such as gender or ethnic origin. These generalisations are discussed in Goldstein *et al.* (2000).

Suppose now that we do not have individual data available but only means at the study level. If we average (12.1) to the study level, we obtain

$$y_{h,j} = \beta_0 + \beta_1 x_{.j} + \beta_2 t_{h,j} + u_{h,j} + e_{h,j}, \tag{12.2}$$

where $y_{h,j}$ is the mean response for the jth study for treatment/control (h). The residual variance for this model is given by

$$\sigma^2_{hu} + \sigma^2_{he}/n_{hj},$$

where n_{hj} is the number of pupils in treatment h for the jth study. It is worth noting at this point that we are ignoring, for simplicity, levels of variation within studies, which will add further levels to the model. If we have information on the relevant quantities in (12.2) then we shall be able to obtain estimates for the model parameters, so long as the n_{hj} differ. Such estimates, however, may not be very precise and extra information, especially about the value of the level-1 variances, will improve them.

Model (12.2) forms the basis for the multilevel modelling of aggregate-level data. In practice, the results of studies will often be reported in non-standard form, for example with no estimate of σ^2_{he}, but it may be possible to estimate this from reported test statistics. In some cases, however, the reporting may be such that the study cannot be incorporated in a model such as (12.2). Goldstein *et al.* (2000) give a set of minimum reporting standards in order that meta-analysis can subsequently be carried out.

12.2.1 Combining individual-level data with aggregate-level data

While it is possible to perform a meta-analysis with only aggregate-level data, it is clearly more efficient to utilise individual-level data where these are available. In general, therefore, we shall need to consider models that have mixtures of individual and aggregate data, even perhaps within the same study.

We can do this by specifying a model that is just the combination of (12.1) and (12.2), namely

$$\left.\begin{aligned} y_{h,ij} &= \beta_0 + \beta_1 x_{ij} + \beta_2 t_{h,ij} + u_{h,j} + e_{h,ij}, \\ y_{h,j} &= \beta_0 + \beta_1 x_{.j} + \beta_2 t_{h,j} + u_{h,j} + e_{h,j} z_{h,j}, \\ z_{h,j} &= \sqrt{n_{hj}^{-1}}, \quad e_{h,j} \equiv e_{h,ij}. \end{aligned}\right\} \tag{12.3}$$

What we see is that the common level-1 and level-2 random terms link together the separate models and allow a joint analysis that makes fully efficient use of the data. Several issues immediately arise from (12.3). One is that the same covariates should be involved. This is also a requirement for the separate models. If some covariate values are missing at either level then it is possible to use an imputation technique to obtain estimates, assuming a suitable random missingness mechanism. The paper by Goldstein *et al.* (2000) discusses generalisations of (12.3) and applies it to an analysis of class size studies.

12.2.2 Clinical trial meta-analysis

One of the most common applications of meta-analysis in medicine is to clinical trials with a basic binary response. This involves a series of choices. The decisions at each stage are similar whether the meta-analyst has only summary data from published results or full individual patient data, but the options available may differ. The first choice is between the fixed effect and random effects models, and in either case the method of estimation must be selected from a number of alternatives. If one is fitting a random effects model, more decisions arise: how to allow for uncertainty in estimation of the between trial variance when constructing a confidence interval for the treatment effect, how to obtain confidence intervals for the between-trial variance, how to incorporate trial-level covariates and how to investigate between-trial heterogeneity.

The usual fixed effect model for meta-analysis assumes the true treatment effects to be homogeneous across trials, and accordingly estimates the common treatment effect θ by a weighted average of the trial-specific estimates, with weights equal to the reciprocals of their within-trial variances. The random effects two-level model assumes that the true treatment effects vary randomly between trials. This model therefore includes a between-trial component of

variance, say τ^2. A commonly used measure of treatment effect in binary event data is the log odds ratio; the normality assumption required is more easily satisfied for this than for alternative measures such as the risk difference. We write a model analogous to (12.2) as

$$\begin{aligned} y_j &= \theta + v_j + e_j, \\ v_j &\sim N(0, \tau^2), \quad \operatorname{var}(e_j) = \sigma_{ej}^2, \end{aligned} \tag{12.4}$$

where y_j is the estimated log odds ratio in trial j with variance σ_{ej}^2 (which is assumed known). Under the assumption of normality, a confidence interval may be calculated for the average treatment effect θ.

For individual-level data, the conventional fixed effects model for p studies can be written as

$$\left.\begin{aligned} \operatorname{logit}(\pi_{ij}) &= \beta_0 + \sum_{k=1}^{p-1} \beta_k \delta_{ijk} + (\theta + v_j) x_{ij}, \\ v_j &\sim N(0, \tau^2), \\ y_{ij} &\sim \text{Binomial}(\pi_{ij}, 1), \end{aligned}\right\} \tag{12.5}$$

where y_{ij} is the binary response, π_{ij} is the probability of a positive response for the ith subject in the jth study, the δ_{ijk} are dummy variables for study membership, and x_{ij} is a dummy variable for treatment/control.

A particularly troublesome issue in all meta-analyses is that of publication bias, whereby certain kinds of studies tend not to get published. To allow for this, it is common to assign to each study a weight as a function of the selection probability for that study. Such models require assumptions on the specific form taken by the selection probabilities, and may involve rather arbitrary decisions for which robustness is lacking (Hedges and Vevea, 1996). Copas (1999) has recommended a sensitivity approach to the problem of publication bias, as an alternative to explicit estimation of corrected estimates. The proposed method involves examination of the extent to which the estimation of θ depends on parameters describing the selection probabilities. This procedure yields a range of plausible estimates of θ rather than a single corrected estimate, and sensitivity analyses using this procedure would seem to be useful.

12.3 SURVIVAL DATA MODELLING

This class of models, also known as event duration models, have as the response variable the length of time between 'events'. Such events may be, for example, birth and death, or the beginning and end of a period of employment, with corresponding times being length of life or duration of employment. There is a considerable theoretical and applied literature, especially in the field of biostatistics, and a useful summary is given by Clayton (1988).

The multilevel structure of such models arises in two general ways. The first

have repeated spells of various kinds of employment, of which unemployment is one, or women may have repeated spells of pregnancy. In this case, we have a two-level model with individuals at level 2, often referred to as a renewal process. We can include explanatory dummy variables to distinguish different kinds or 'states' of employment or pregnancy, such as the sequence number. The second kind of model is where we have a single duration for each individual, but the individuals are grouped into level-2 units. In the case of employment duration, the level-2 units would be firms or employers. If we had repeated measures on individuals within firms then this would give rise to a three-level structure

A characteristic of duration data is that for some observations we may not know the exact duration but only that it occurred within a certain interval. This is known as interval censored data: if less than a known value, left censored data; if greater than a known value, right censored data. For example, if we know at the time of a study that someone began her pregnancy before a certain date then the information available is only that the duration is longer than a known value. Such data are known as right censored. In another case, we may know that someone entered and then left employment between two measurement occasions, in which case we know only that the duration lies in a known interval.

There are a variety of models for duration times, and we here mention only three of the most common. We shall merely sketch the model without going into details of estimation. A full description of estimation procedures is given by Goldstein (1995).

Perhaps the most commonly used is the proportional hazards model, also known as a semiparametric proportional hazards model. Consider the two-level proportional hazards model for the jkth level-1 unit:

$$h(t_{jk}; X_{jk}) = \lambda(t_{jk}) \exp(X_{jk} \beta_k), \tag{12.6}$$

where X_{jk} is the row vector of explanatory variables for the level-1 unit and some or all of the β_k are random at level 2.

We suppose that the times at which a level-1 unit comes to the end of its duration period or 'fails' are ordered, and at each of these we consider the total 'risk set'. At failure time t_{jk}, the risk set consists of all the level-1 units that have been censored or for which a failure has not occurred immediately preceding time t_{jk}.

Another model in common use is the accelerated life model, where the distribution function for duration is commonly assumed to be of the form

$$f(t; X, \beta) = f_0(t e^{X\beta}) e^{X\beta},$$

where f_0 is a baseline function (Cox and Oakes, 1984). For a two-level model, this can be written as

$$l_{ij} = \log t_{ij} = X_{ij} \beta_j + e_{ij}, \tag{12.7}$$

which is in the standard form for a two-level model. We shall assume normality for the random coefficients at level 2 (and higher levels), but at level 1 we may

have other distributional forms for the e_{ij}. The level-1 distributional form is important where there are censored observations.

The third model, often used in demographic studies (Steele *et al.*, 1996), is the piecewise duration model. We suppose that the total time interval is divided into short intervals during which the probability of failure, given survival up to that point, is effectively constant. Denote these intervals by t $(1, 2 \ldots, T)$. We define the hazard at time t as the probability that, given survival up to the end of time interval $t-1$, failure occurs in the next interval. At the start of each interval, we have a 'risk set' n_t consisting of the survivors, and, during the interval, r_t fail. If censoring occurs during interval t then this observation is removed from that interval (and subsequent ones) and does not form part of the risk set. A simple, single-level, model can be written as

$$\pi_{i(t)} = f[\alpha_t z_{it}, (\beta X)_{it}], \tag{12.8}$$

where $z_t = \{z_{it}\}$ is a dummy variable for the tth interval and α_t is a 'blocking factor' defining the underlying hazard at time t. The second term is a function of covariates. A common formulation would be the logit model, and a simple such model, in which the first blocking factor has been absorbed into the intercept term could be written as

$$\text{logit}(\pi_{i(t)}) = \beta_0 + \alpha_t z_{it} + \beta_1 x_{1i}, \quad (z_2, z_3, \ldots, z_T). \tag{12.9}$$

Since the covariate varies across individuals, in general the data matrix will consist of one record for each individual within each interval, with a (0,1) response indicating survival or failure. The model can be fitted using standard procedures, assuming a binomial error distribution.

As it stands (12.9) involves the fitting of $T-1$ blocking factors. However, this can be avoided, (Goldstein, 1995, Chapter 9) by fitting a low-order polynomial to the sequentially numbered time indicator, $Z^* = 1, 2, \ldots, T$, so that (12.9) becomes

$$\text{logit}(\pi_{i(t)}) = \beta_0 + \sum_{h=1}^{p} \alpha_h^* (z_{it}^*)^h + \beta_1 x_{1i}, \tag{12.10}$$

where p is typically 3 or 4.

The logit function can be replaced by, for example, the complementary log–log function, which gives a proportional hazards model, or, say, the probit function. We note that we can incorporate time-varying covariates such as age. A 'competing risks' model with several different kinds of survival can be constructed by extending the response to become a multinomial vector representing the various risks.

Consider the two-level extension where we suppose that level 1 is individual (pregnancy length) and level 2 is community. A simple generalisation is

$$\text{logit}(\pi_{ij(t)}) = \beta_0 + \sum_{h=1}^{p} \alpha_h^* (z_{it}^*)^h + \beta_1 x_{1ij} + u_j, \tag{12.11}$$

where u_j is the 'effect' for the jth community, and is typically assumed to be distributed normally with zero mean and variance σ_u^2. We can elaborate this using random coefficients, resulting in a heterogeneous variance structure, further levels of nesting etc. This is just a two-level binary response model and can be fitted using, for example, quasi-likelihood or Markov-chain Monte Carlo (MCMC) methods (for details about using these in MLwiN, see Rasbash *et al.*, 1999a, b). The data structure has two levels, so that individuals will be grouped (sorted) within communities, but within each community the record order is again immaterial. For the competing risks model we use the multinomial two-level formulation (Goldstein, 1995). The setting up and fitting of such a model in MLwiN is described in Yang *et al.* (1999).

12.4 CONTEXT AND COMPOSITION

Multilevel modelling has been an important advance in health service and public health research since it has enabled a focus on both microlevel and macrolevel relationships simultaneously, as well as the relationships between them (Groenewegen, 1997). The questions facing researchers concern the degree to which observed differences at the macrolevel – typically hospitals or areas – reflect genuine contextual differences between those areas or whether they do little more than reflect the composition of those areas in terms of the microlevel (typically the individual). For example, Jones (1997) questions whether the relationship between voting behaviour and place is *contextual* – meaning that '... something about the social and economic milieux of [an] area ... produces a distinctive political culture' – or whether it merely reflects the *composition* of an area, with particular relevance to social class composition.

Duncan *et al.* (1998), discussing institutional performance (see Chapter 9 for further discussion of this subject), suggest that the average performance of a clinic can be seen to comprise three elements:

average clinic performance	=	composition of the clinic	+	contextual clinic difference	+	composition/ contextual interaction.

The composition of the clinic in this example refers to the make-up of the clinic in terms of the net characteristics of the people who attend the clinic; the contextual differences are the additional effect that a clinic has once its composition has been taken into account, and the interaction then reflects differential performance across patient groups. At a microlevel, the composition of the clinic could mean no more than taking individual patient characteristics into account; the interaction with the context then reflects a microlevel variable that is random across the macrolevel (clinics). However, this section considers compositional variables at both the micro- and macro-levels; that is, it considers the individual patient in relation to the overall composition of the practice.

Consider a hypothetical example in which the objective is to determine what effect different hospitals have on a patient outcome – for example, a rating of health following surgery. For every patient in every hospital data are collected as to their age, sex and whether or not the patient is receiving private medical care. Do these data then refer to patients or hospitals? Since they were collected for every patient, they must refer to the patient; however, in addition, they may provide hospital-level information. If the health system under study has two types of hospital – private and public – then this is a hospital-level variable. Moreover, if all patients receiving private medical care do so at private hospitals, and all other patients are treated at public hospitals, then there is no information at patient level (since every patient within each hospital will have the same classification). However, it may be that some privately funded patients receive their care in public hospitals; in this case, a comparison of interest may be between the outcomes of privately funded patients who are treated in public hospitals as opposed to those who are treated in private hospitals. Alternatively, the health system may have three types of hospital – private, public and mixed – the composition of the hospital can be seen separately from the individual-level variable from which it is derived.

In a similar manner, it may be important to draw comparisons between single-sex hospitals and mixed-sex hospitals (or hospital wards for a particular diagnosis), so sex may be considered a descriptor of hospital composition as well of the individual patients. This example can be developed further by considering the patient's age. This is not necessarily a question of comparing categories of hospitals – such as those providing paediatric or geriatric care – but may involve a more complex relationship between individual and hospital composition, such as the average age of the patients treated in that hospital. In this manner, the influence of the age of each patient on the outcome can be separated from the way in which it is influenced by the operational context of the hospital. Do older patients fare better in a hospital that predominately treats older patients, or one that generally treats younger patients? In a similar manner it is possible to consider the proportion of privately funded patients within a hospital rather than categorising all hospitals treating both types of patients as being mixed. The same is true for the patient's sex; in general, any microlevel variable – continuous or categorical – can also be considered at the macrolevel. The mean is a common way of summarising the data, but, depending on the particular research question, the minimum, maximum or another measure may be a more appropriate description of the composition of the higher-level units.

Duncan *et al.* (1998) give illustrations of a variety of ways in which individual and compositional variables may interact in cross-level relationships; these have been adapted in Figure 12.1. The two lines can be thought of as representing, by way of example, the predicted response for patients of different ages – say 50 years (broken lines) and 80 years (solid lines). With the vertical axis indicating the level of the response, the horizontal axis reflects the mean age of patients in the hospital. Figure 12.1(a) therefore illustrates a situation in which the 50-year-old patients are generally healthier than the 80-year-olds, and this

difference is constant no matter what the composition of the hospital. Any differences between hospitals are therefore contextual rather than reflecting differences in composition. This is not to say that there is no difference between hospitals in terms of their composition, merely that their composition has no additional bearing on the patient outcomes. It may not be that the same average difference will be seen at all hospitals; it is possible that the age effect varies randomly across hospitals but there is no relationship between these random age effects and the hospital composition. Figure 12.1(b) suggests a scenario in which there is little difference between the health rating of the two age groups, but the ratings tend to be higher in hospitals in which the average age is high. So the health of the individual is determined not by the individual patient's age, but by the average age of patients treated in the hospital – this situation is the converse of Figure 12.1(a) in that it is the macrolevel compositional variable that is important rather than the microlevel patient characteristic. Figure 12.1(c) and (d) reflect a combination of these two situations in which both individual and compositional factors have an important impact on the response. The health of 50-year-olds is generally better than that of 80-year-olds in a hospital of the same composition, and this difference is independent of the hospital composition. Figure 12.1(c) presents a scenario in which the average health of all patients is improved in a hospital with a high mean age, whilst in Figure 12.1(d), the health of patients at hospitals with a young mean age is improved relative to patients of the same age who are treated in hospitals with a high mean age. The remaining four graphs illustrate some possible interactions between individual characteristics and hospital composition. In Figure 12.1(e), there is little difference between the health of those patients treated at hospitals with a high mean age, irrespective of the age of the individual patients. There are, however, substantial differences in hospitals with a low mean age, with younger patients faring very much better than the older patients. In Figure 12.1(f), the situation is the same in hospitals with a low mean age, but in hospitals with a high mean age it is the 80-year-olds whose health is better than the 50-year-olds. There are no age differences in the health of patients with composition in the middle of the age range. Figures 12.1(g) and (h) present more complex interactions; in both situations, the health of the 50-year-old patients is better than that of the 80-year-olds in hospitals with either a low or a high average age. However, in Figure 12.1(g) these differences disappear when patients are treated at hospitals with composition in the middle of the age range, with the average health of the older patients being better and that of the younger patients being worse; in figure 12.1(h), on the other hand, the differences between the two age groups are accentuated in these hospitals.

A common situation that gives rise to compositional effects is the differences between institutions in the recording of information. Leyland and Boddy (1998) considered the differences between Scottish hospitals in 30-day mortality rates following acute myocardial infarction (AMI). One patient characteristic strongly associated with death following AMI is the recording of a secondary diagnosis of other (non-ischaemic) heart disease (odds ratio = 1.77; 95% confidence interval (CI) 1.56–2.00). However, the large variation between hospitals

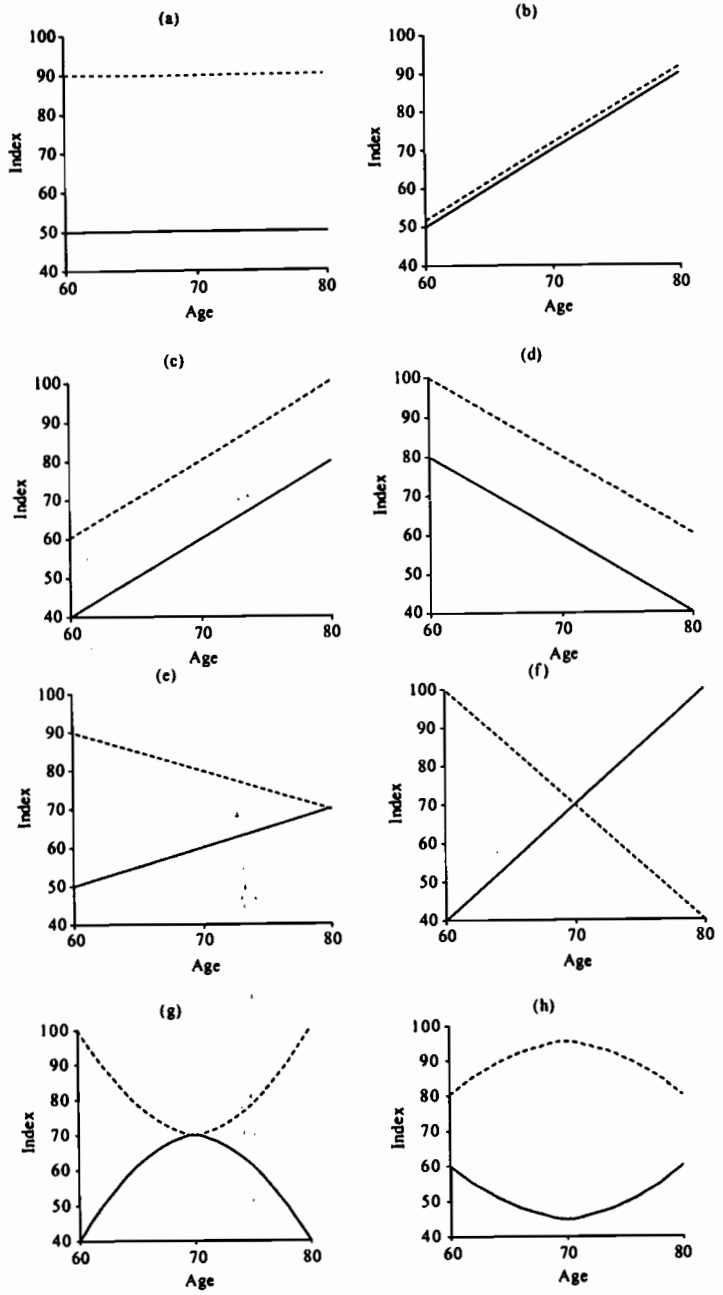

Figure 12.1 Illustration of cross-level interactions between individual and compositional variables. Reprinted from *Social Science and Medicine* **46**. Duncan, C., Jones, K. and Moon, G. Context, composition and heterogeneity: using multilevel models in health research. pp. 97–117. Copyright (1998), with permission from Elsevier Science.

in the proportion of patients for whom such a secondary diagnosis was recorded raises doubts as to whether this was really reflecting the patients' condition or the hospital's recording practice. Among the 31 hospitals that saw more than 50 AMI patients in the year of study, the recording of other heart disease ranged from 8% to 33%. If some hospitals have a lower threshold for a patient to be classified as having other heart disease (i.e. they are classifying a high proportion of patients in such a manner) then it is likely that the mortality rate among such patients will be lower in that hospital than in hospitals that have a higher threshold (those with a lower proportion of patients with other heart disease). However, if the classification of patients is still related to the severity of their condition then it is also likely that the mortality among patients who do not have a secondary diagnosis of other heart disease will also be lower among hospitals that have higher rates of recording of the condition. It is therefore possible to fit a model that indicates at the patient level the odds of mortality associated with the presence of other heart disease (1.74; 95% CI 1.51–1.98) and also the relationship between mortality and the composition of the hospital such that a 10% increase in the proportion of patients with other heart disease is associated with an odds ratio of 0.85 (0.72–1.01) among patients with other heart disease and 0.76 (0.68–0.86) among patients not classified as having other heart disease.

Table 12.1 shows the combined effect of other heart disease as both a patient characteristic and as a compositional variable. Three levels of the percentage of patients with other heart disease recorded are used: 8.8%, 17.0% and 32.7%, which correspond to the 5th, 50th and 95th percentiles for patients (that is, 5% of all patients are in hospitals in which no more than 8.8% have a recorded secondary diagnosis of other heart disease, etc.). The reference category is patients without a diagnosis of other heart disease in a hospital that records 17.0% of cases as having this diagnosis. The odds of mortality associated with other heart disease are 1.75 in such a hospital; this figure decreases to 1.58 in hospitals with the lower level of recording, and increases to 2.09 in hospitals recording 32.7% of patients as having other heart disease. This therefore corresponds to a situation somewhat similar to Figure 12.1(e), but with the lines diverging rather than converging as the compositional variable increases. Patients with other heart disease are always at greater risk than patients without this recorded (if the vertical axis is the odds of mortality, the broken line

Table 12.1 Odds of mortality (95% confidence intervals in parenthese) associated with hospital composition (percentage of patients with other heart disease) and the recording of other heart disease for individual patients.

Percentile	Percentage of patients with other heart disease	Patients without other heart disease	Patients with other heart disease
5th	8.8%	1.25 (1.14–1.38)	1.98 (1.64–2.45)
50th	17.0%	1.00	1.75 (1.51–1.98)
95th	32.7%	0.65 (0.54–0.79)	1.36 (1.06–1.76)

corresponds to patients with this diagnosis). In both groups of patients, the odds of mortality decrease as the compositional variable increases (so both lines slope down from left to right), and the increased risk becomes more pronounced as the level of recording of the diagnosis in a hospital increases (so the distance between the two lines increases).

12.5 RECENT DEVELOPMENTS IN MULTILEVEL MODELLING

The topic of measurement errors is a complex one, and there appears to be little attempt to make adjustments when models include such errors in either the response or predictor variables. It is known, however, that, in the presence of such measurement error, inferences may be biased. We refer the reader to a detailed discussion, with an example taken from education, given by Woodhouse *et al.* (1996).

One area that has been receiving some attention recently is multilevel covariance structure analysis or structural equation modelling. Hox (1995) describes this general approach as one that encompasses both path models and factor models; the former structural model is used to describe predictive relationships between observed variables and latent factors, whilst the latter factor model describes the construction of the latent factors from the observed variables. For introductory texts on this subject the reader is referred to Muthén (1994) and McDonald (1994). Software capable of fitting multilevel structural equation models includes Mplus (Muthén and Muthén, 1998) and STREAMS (Gustafsson and Stahl, 1997).

A further common issue in statistics concerns the analysis of data sets where some of the data are missing, and this concern extends to multilevel data sets. The fact that a balanced design is not a prerequisite of a multilevel model means that subjects with some outcomes missing may still be included in an analysis (see Chapter 5 on multivariate regression analysis). When the explanatory variables are incomplete, there are typically two options as discussed in Goldstein (1995). If the data can be viewed as being missing completely at random (MCAR) or missing at random (MAR) conditional on other explanatory variables but independently of the outcome (Rubin, 1976) then a two-stage approach may be adopted. At the first stage a multivariate model is used to obtain predictions for all missing data values. The second stage resorts to multiple imputation (Rubin, 1987); a number of complete data sets are formed by repeatedly sampling from the predicted distributions of the missing values. The data sets are then analysed in turn, and the estimates obtained are based on data with the correct distributional properties. If the data are not missing at random – that is, the fact that the data are missing is related to the response and is in itself informative (see, e.g. Best *et al.*, 1996) – it is common practice to model the missingness mechanism and then proceed as if the data were missing at random.

CHAPTER 13

Software for Multilevel Analysis

Jan de Leeuw
UCLA Department of Statistics, Los Angeles, CA, USA

Ita G.G. Kreft
Health and Human Services, CSLA, Los Angeles, CA, USA

13.1 INTRODUCTION

In this chapter we review some of the more important software programs and packages that are designed for, or can be used for, multilevel analysis. These programs differ in many respects. Some are parts of major statistics packages such as SAS or BMDP. Others are written in the macro language of a major package. And some are stand-alone special-purpose programs that can do nothing but multilevel analysis. We have been involved in a number of these comparisons before. The first (Kreft *et al.*, 1990), comparing HLM, ML3, VARCL, BMDP5-V and GENMOD, was published in Kreft *et al.* (1994). The second comparison (van der Leeden *et al.*, 1991), comparing HLM, ML3 and BMDP5-V on repeated measures data, was published in van der Leeden *et al.* (1996). We give both the reference to the internal report version and to the published version, because the unpublished version usually has much more material. Giving both references also shows the unfortunate time interval between the two, which is especially annoying in the case of software reviews. The reviews were summarised briefly in our book (Kreft and de Leeuw, 1998, Section 1.6).

Since our last publication on the subject, there have been many major changes. The program GENMOD, which was never easy to obtain, has more or less completely disappeared. VARCL, which was one of the leading contenders in the early 1990s, is no longer actively supported or developed, which means that it has rapidly lost ground. BMDP, as a company, went out of business, which had serious consequences for its software products. Programs such as HLM, written originally for DOS, were upgraded for Windows. ML3

Multilevel Modelling of Health Statistics Edited by A.H. Leyland and H. Goldstein