Pergamon

Soc. Sci. Med. Vol. 46, No. 1, pp. 97–117, 1998

Printed in Great Britain
PII: S0277-9536(97)00148-2
0277-9536/98 $19.00 + 0.00

CONTEXT, COMPOSITION AND HETEROGENEITY: USING MULTILEVEL MODELS IN HEALTH RESEARCH

CRAIG DUNCAN,[1]* KELVYN JONES[1] and GRAHAM MOON[2]

[1]Department of Geography, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, U.K. and [2]School of Social and Historical Studies, University of Portsmouth, Milldam, Burnaby Road, Portsmouth PO1 3AS, U.K.

Abstract—This paper considers the use of multilevel models in health research. Attention focuses on the structure and potential of such models and particular consideration is given to their use in elucidating the importance of contextual effects in relation to individual level social and demographic factors in understanding health outcomes, health-related behaviour and health service performance. Four graphical typologies are used to outline the questions that multilevel models can address and the paper illustrates their potential by drawing on published examples in a number of different research areas. © 1998 Elsevier Science Ltd. All rights reserved

Key words—health outcomes, health-related behaviour, health service performance, context, multilevel model

INTRODUCTION

Within recent years, many areas of social science research have been increasingly concerned with tracing the connections between individuals and the contextual settings in which they lead their lives. This development has been particularly noticeable within the field of health research where considerable interest has focused on the role of factors associated with the external environment in influencing individual susceptibility to disease. Work in this area has, however, proved to be difficult and contentious. In Britain, for example, several studies have provided unequivocal evidence of variations in individual health status between different geographical settings or "contexts" (Britton, 1990; Macintyre, 1986). This evidence does not, however, necessarily mean that "contextual" effects directly associated with the general living environment have significance for health. Rather, the variations may arise from "compositional" effects with particular types of people, who are more likely to become ill due to their individual characteristics, being found more commonly in particular places.

Two very different interpretations are, therefore, possible and the key question is not whether variations between different settings exist but what is their origin. On the basis of a large-scale empirical study, Blaxter is convinced that contextual effects are important and that places matter in their own right. She writes, "while the health of manual men and women was almost always poorer than that of non-manual, it is clear that types of living area do make a difference" (Blaxter, 1990, p. 82). In contrast, Sloggett and Joshi's analysis leads them to conclude that "excess mortality associated with residence in areas designated as deprived...is wholly explained by the concentration in those areas of people with adverse personal or household socio-economic characteristics" (Sloggett and Joshi, 1994, p. 1470).

These notions of contextual and compositional effects have general relevance and they apply not only when the focus is upon context as geographical setting but also when context is seen in terms of administrative (for example, health authority districts), temporal (for example, different time periods), or institutional (for example, hospitals or clinics) settings. In terms of the latter, there are extremely important implications for health services research as they connect with recent moves towards the use of performance indicators. Variations in the performance of health service activities between different provider units (e.g. vaccination uptake rates in clinics; length of stay times in hospitals) can be attributed to both the type of clients particular units serve ("compositional" effects) as well as the nature of the environment from and in which the service is provided ("contextual" effects). In order to establish how well a particular institution is performing adjustment must be made for the type of people it serves (Goldstein and Spiegelhalter, 1996). Thus, performance measures are needed which are able to separate compositional from contextual effects.

In all areas of research there is, however, a need to go beyond separating contextual and compositional effects in simple overall terms as there is the

*Author for correspondence.

possibility of interactions between the two: contextual effects may not be the same for all types of people. For example, while context may matter for the health outcomes of one particular group of people, it may not have any influence upon the health experiences of other groups. Thus, contextual effects may be too complex to reduce to overall summary measures. At the same time, the differences between individuals may also be too complex to reduce to simple summary "averages". For example, one group of people, regardless of context, may have more variable health experiences than another. Or, to put it another way so as to emphasise the distinction, within particular contexts one group's health experiences may be more or less variable than another's. We must, therefore, anticipate heterogeneity, both between individuals and between contexts.

This paper shows how notions of between-context and between-individual heterogeneity can be developed and investigated in research based on extensive quantitative designs through the application of multilevel modelling techniques. The paper begins with a brief review of traditional statistical approaches to contextual variation, highlighting the problems associated with their use. It goes on to outline more recent multilevel modelling approaches showing how they are capable of circumventing many of these problems as well as offering a way of examining complex forms of variation between individuals. Four graphical typologies are used to develop the outline before a series of published examples are drawn upon to illustrate some of the ways in which multilevel models have been used in health research to date. Whilst the paper is aimed at a general audience, technical issues are unavoidable. Such issues will, however, be discussed in a relatively non-technical fashion and more technical detail can be found elsewhere (Goldstein, 1991, 1995; Jones and Bullen, 1994).

CONTEXT AND COMPOSITION: TRADITIONAL APPROACHES

By acknowledging the existence of compositional and contextual effects, it immediately becomes apparent that the outcomes under study are associated with processes operating at more than one "level": a lower level compositional effect and a higher level contextual one. To gain a more complete understanding, all the relevant levels of analysis need to be considered simultaneously. This requirement poses serious technical difficulties for traditional statistical modelling techniques as they operate only at a single level. Consequently, researchers have been forced to decide at which level they are going to work. There are two possibilities, both of them problematic. First, they could work at the aggregate/higher level aggregating up any individual level variables according to the higher level units of interest (contexts/institutions). For example, the number of people reporting limiting long term illness could be aggregated by geographical region and these aggregate rates could then be modelled as a function of the percentage unemployed in each geographical region. As the classic paper by Robinson showed, however, there is no reason to suggest that a relationship that occurs at the aggregate level can be held to exist at the individual level (Robinson, 1950). Taking the hypothetical example, while we may find that high levels of illness are associated with high levels of unemployment at the regional level, it may actually be that the people who are ill in regions with high-unemployment are those in work.

The second possibility is that the researcher chooses to work at the individual rather than the contextual level. Obviously, if no attempt is made to incorporate measures relating to higher-level characteristics then the "atomistic fallacy" (Alker, 1969) is committed as the research completely ignores the context in which individual action occurs. Consequently, researchers have developed the approach of incorporating contextual characteristics by disaggregating higher-level variables to the individual level, (that is, assigning the appropriate values of aggregate variables to individuals, the units of analysis), and performing an ordinary least squares (OLS) regression analysis (Hox and Kreft, 1994). This approach, however, is also problematic. Since the individuals are sampled within particular contexts all the unmodelled contextual variation will enter the single individual level random (error) term that is applied in OLS regression. This will mean that, potentially, the error terms of individuals in the same context are correlated, breaking the assumption of an independently distributed error term on which standard errors and significance tests are based in single level regression. This lack of independence produces inefficient estimates and an increased tendency to find differences and relationships where none exist (Skinner *et al.*, 1989). In addition, this approach assumes that the regression coefficients are equal in all contexts thus propagating the notion that processes work out in the same way in different contexts.

The restriction that all coefficients are equal in all contexts can be avoided if either an analysis of variance (ANOVA) or an analysis of covariance (ANCOVA) is used (Silk, 1977). These techniques involve including extra "fixed" terms in the model for each parameter that is allowed to take a different value in each contextual setting. Thus, if we wish to recognise 200 settings and allow each to have a different intercept and slope term there would be 400 parameters and a very large sample size would be needed to obtain reliable estimates. Consequently, these approaches are neither efficient nor parsimonious. Furthermore, inferences are only made on the basis of the contexts explicitly ident-

ified and not to the wider population from which they are drawn, making this approach both unrealistic and limited.

In more recent years, a new form of statistical modelling has been developed which is conceptually more realistic as it handles the micro-scale of people and the macro-scale of contexts simultaneously within one model. Several terms have been used to describe this new development: multilevel models (Goldstein, 1995), random coefficient models (Longford, 1993) and hierarchical linear models (Bryk and Raudenbush, 1992). (Hereinafter, we will only use the term "multilevel models"). By distinguishing different levels, multilevel models are able to treat the particular contexts identified as comprising a random sample drawn from a larger underlying population. The procedures then make inferences about the variation among all contexts in the population using this random sample of contexts. Consequently, the differences between contexts are not treated as being fixed and separate (or "unrelated") but are seen as coming from distributions that relate to a larger population (DiPrete and Forristal, 1994; Hox and Kreft, 1994).

The value of this approach is beginning to be recognised within many different research areas associated with health and health care, for example, epidemiology (Von Korff *et al.*, 1992), nursing (Wu, 1995), community psychology/health promotion (Hedeker *et al.*, 1994), resource allocation (Carr-Hill *et al.*, 1994, 1996), institutional performance (Leyland, 1995; Leyland and Boddy, 1997), medical practice (Davies and Gribben, 1995; Gatsonis *et al.*, 1993) and health geography (Congdon, 1995; Duncan *et al.*, 1995a; Ecob, 1996; Gould and Jones, 1996; Langford and Bentham, 1996; Shouls *et al.*, 1996). In the next section, four graphical typologies are used to outline this approach and it will be shown how it provides a framework for contextual analysis in health research which is not only technically stronger but which also has a much greater capacity for generality than traditional single-level statistical techniques.

MULTILEVEL ANALYSIS: FOUR GRAPHICAL TYPOLOGIES

Graphs of varying relationships

Consider a simple regression model in which it is hypothesized that the probability of an individual reporting limiting long term illness (the response variable) is a function of their age (the predictor variable). An analysis using this model may generate the relationship shown in Fig. 1(a). Here the limiting long term illness/age relationship is shown as a straight line with a positive slope: older people are more likely to report being sick. In such a model the context in which the behaviour occurs is completely ignored: the same single straight line relationship is held to exist everywhere. In effect the model has explained "everything in general and nothing in particular". In a multilevel framework this can be rectified by recognising the communities in which individuals live and using a two-level model with individuals at level 1 nested within communities at level 2. One possible result is shown in Fig. 1(b), a two-level "random-intercepts" model.

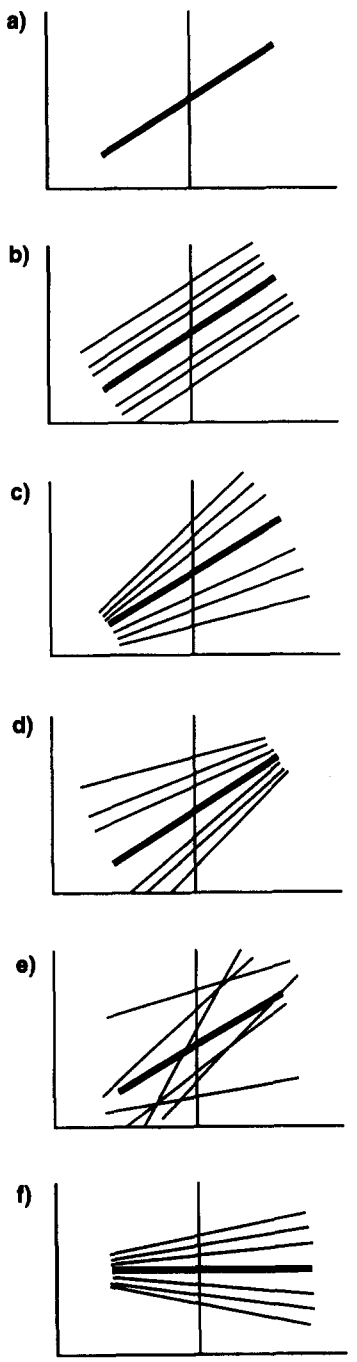

Fig. 1. Varying relationships between illness and age.

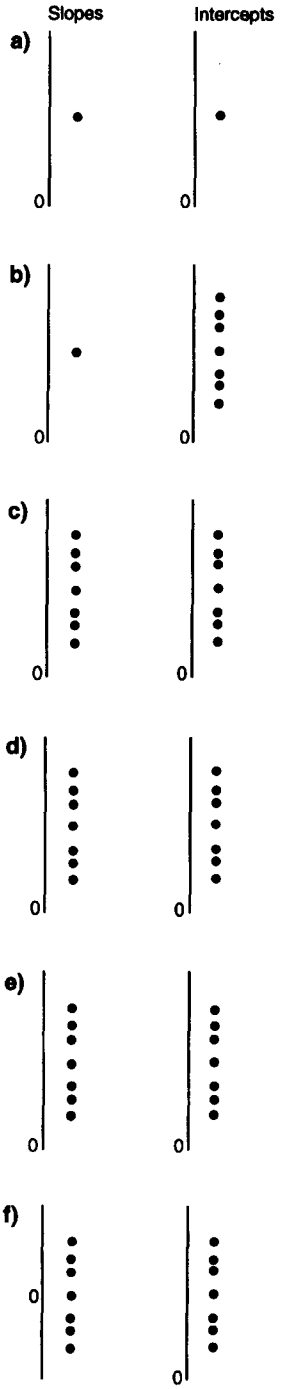

Fig. 2. Dotplots of the higher-level distributions underlying Fig. 1.

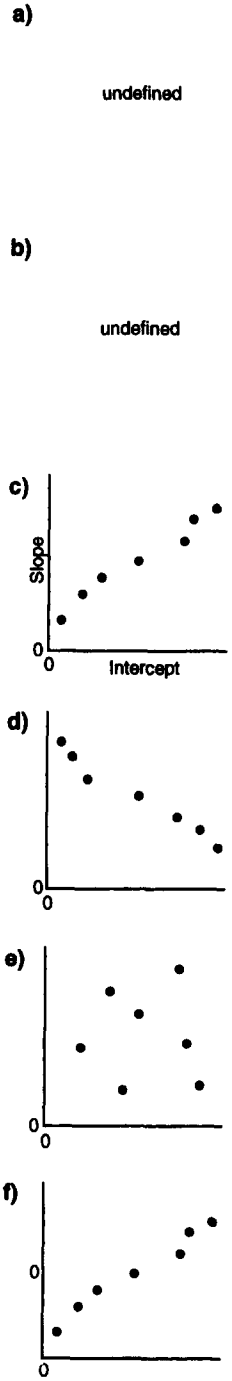

Fig. 3. Scatterplots of the higher-level distributions underlying Fig. 1.

Here each of six different communities have their own limiting long term illness/age relation represented by a separate line. The single, thicker line represents the general relationship across all six communities. The parallel lines imply that, while the limiting long term illness/age relation in each place is the same, some places have uniformly higher illness rates than others.

In Fig. 1(c) and (d) the situation is more complicated as the steepness of the lines varies from place to place. In Fig. 1(c) the pattern is such that place makes very little difference for the young but there is a high degree of between-community variation in illness amongst the elderly. In contrast, Fig. 1(d) shows large place-specific differentials exist for the young. In Fig. 1(e) there is a complex interaction

between age and place. In some communities it is the young who have relatively high rates; in others it is the old. While the final plot, Fig. 1(f), is unlikely to occur in terms of the present example, it can be expected in other health research areas. Across all the communities there is no relationship between the response and the predictor (the single thicker line is horizontal) but in specific communities there are distinctive relationships. This situation is similar to Fig. 1(c) but here the differences result from some communities having high rates for the elderly, while in others they have the lowest rates.

Similar patterns of varying slopes and intercepts could be obtained by treating them as fixed and unrelated through such techniques as ANOVA and ANCOVA. As mentioned earlier, however, in the multilevel approach the differences between contexts —the varying slopes and intercepts—are treated as coming from two higher-level distributions that relate to a larger underlying population. Thus, the models do not fit a slope and intercept to each community separately but instead they estimate the statistical characteristics of the higher-level population distributions. Since the vertical axis is centred at the mean age of individuals, the intercept represents the probability that a person of average age reports limiting long-term illness. The slope measures the increase in reporting limiting long-term illness associated with a unit increase in age. A multilevel analysis summarizes these higher-level distributions in terms of a fixed part and a random part. The fixed part gives the mean value for each distribution and consists of two fixed, unchanging terms—the average slope and intercept across all communities (shown by the thick lines in Fig. 1). The random part of the model, meanwhile, is expanded to include two extra variance terms, which summarize the variability of slopes and intercepts across communities, and a covariance which assesses the degree to which the two distributions are related.

Returning to Fig. 1, the cases shown would each have the same fixed part—an average intercept and an average slope (which would be zero in Fig. 1(f))—and the same level 1 random part—a variance term summarizing residual individual variation—but they would differ in terms of their level 2 random parts. These differences are represented in Figs 2 and 3 which give the higher-level distributions that correspond to each of the different graphs in Fig. 1. Figure 2 shows a dotplot for the distributions of the slopes and intercepts separately, while Fig. 3 shows a "scatterplot" of the joint distribution. To reiterate, these distributions concern communities, not individuals, and result from treating places as a sample drawn from a population.

As can be seen, Fig. 1(a) is the result of a single non-zero intercept and slope. This is a single-level model and so there are no higher-level distributions. Figure 1(b) has a set of intercepts but a single slope. Figure 1(c)–(d) have sets of both intercepts and slopes. The different form of each of these results from the different ways in which the intercepts and slopes are associated. In Fig. 1(c) the illness/age relation is strongest in places where illness is highest on average; a steep slope is associated with a high intercept. Or, put another way, there is positive association between the intercepts and slopes as shown in Fig. 3(c). In Fig. 1(d) this

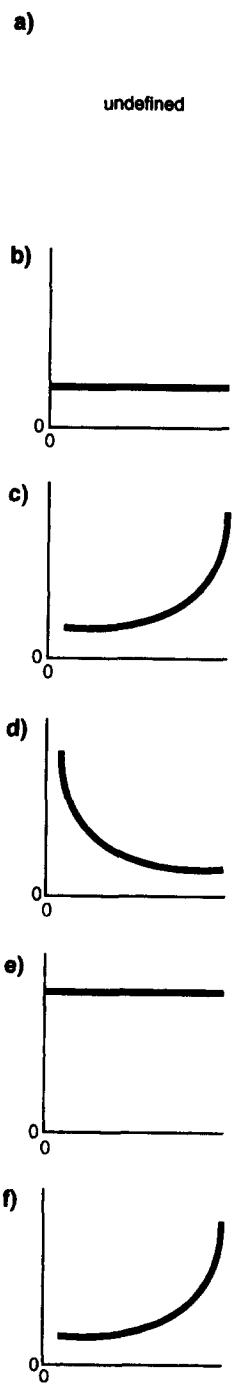

Fig. 4. Between-context variation.

situation is reversed as the illness/age relation is strongest in places where illness is lowest on average; a steep slope is associated with a low intercept. Consequently, Fig. 3(d) shows a negative association between intercepts and slopes. The complex criss-crossing of Fig. 1(e) is the result of a lack of pattern between the slopes and intercepts as shown in Fig. 3(e). While Fig. 1(f) shares the same positive association between slopes and intercepts as Fig. 1(c), this time, as shown in Fig. 3(f), the slopes vary about zero since in some communities the slopes are positive but in others they are negative.

The distributions of the intercepts and slopes given in Figs 2 and 3 refer specifically to places not people and their variability is summarised by only three extra terms in the random part of the model. Importantly, this situation prevails whether there are 20 contexts or 200. Thus, unlike traditional ANOVA/ANCOVA approaches, it is these higher level random terms that summarise the extent to which places differ and not a multitude of parameters in the fixed part of the model. It should be noted, however, that predictions of place-specific intercepts and slopes can be obtained once the overall between-place variation has been estimated. Since the contexts identified are treated as coming from a population distribution the estimation procedure can "pool" all the information in the data thus allowing the predictions of place-specific relationships to be based on precision-weighted estimators (Morris, 1983) which take account of sample sizes. Imprecisely estimated, context-specific relations are "shrunk" towards the overall fixed relationship, while reliably estimated, within-context relations are largely immune to this shrinkage. Hence, multilevel models have the potential to avoid the mis-estimation problems caused by small numbers and sampling fluctuations in traditional methods based on single-level separate-regressions*.

The higher level distributions of slopes and intercepts arise through the specification of models at each level and then their combination into an overall model (Jones, 1991). More specifically, there is an individual-level, micro-model which represents the within-place equation, and an ecological, macro-model in which the parameters of the within-place model are the responses in the between-places models. This simultaneous specification allows for the separation, in a quantitative sense, of the compositional from the contextual (Mason *et al.*, 1984). The central empirical question concerning contextual effects is whether the higher-level variation remains significant when a range of appropriate and relevant individual variables (eg: age, gender, income, class, employment status, housing tenure, educational background) are included in an overall model to allow for the population composition of particular places†.

Multilevel modelling does not, however, simply provide a means of assessing the relative contribution of compositional and contextual effects. Importantly, the technique provides a way of showing how, and for which types of people, contextual effects matter. This is emphasised by the graphs in Fig. 4 in which between-context variation (vertical axis) has been plotted against the predictor variable age (horizontal axis) for each of the models shown in Fig. 1. In Fig. 4(a) all places are the same and so there is no between-place variation. In Fig. 4(b) there are differences between places but the same degree of difference exists at all ages. In Fig. 4(c), (d) and (f) the degree of variation between places changes according to age and does so either in an increasing fashion ((c) and (f)) or in a decreasing fashion (d). These graphs reveal that what we are actually doing is modelling the total between-context variation as a function of an individual predictor variable. Consequently, not only can we can separate contextual and compositional effects in a general sense but we can also investigate how contextual variation in the population changes according to individual characteristics.

*Since the multilevel approach involves estimating more than one random term, OLS regression strategies will not provide reliable estimates of these parameters. Consequently, various specialised software packages are now available whilst appropriate facilities have also been included in the general packages *SAS*, *GENSTAT* and *BMDP*. Of the four original specialised packages, *GENMOD* is no longer generally available. *VARCL* handles varying slopes and intercepts models up to three levels and varying intercepts-only models up to nine. This program can also deal with binomial and Poisson variation at level-1 as outlined later in this paper but only simplified non-linear procedures that are known to underestimate the coefficients are implemented. *VARCL* is no longer being developed. *HLM* (Bryk *et al.*, 1988) is undergoing development and a Windows version has just been released; it is generally regarded as a package that is particularly straightforward to use. *MLn* (Rasbash and Woodhouse, 1995) is a complete package containing a high-level macro programming language for the implementation of new procedures as well as general data manipulation and graphical facilities; given this broader range the package has a steeper learning curve but is more flexible. More recently, newer packages have been developed. *BUGS* (Gilks *et al.*, 1994), which is mentioned briefly again later in the paper, is one of these and, unlike the original programs, is based on a Markov Chain Monte Carlo approach, especially one variant thereof, Gibbs Sampling. More interactive packages based on *Xlisp-Stat* are also becoming available. For a detailed comparative review of the four original packages including details of the different algorithms that they use see Kreft *et al.* (1994). For more information about recent software developments, as well as a general critical discussion of issues surrounding the estimation of multilevel models see de Leeuw and Kreft (1995). Contact addresses for some of the packages together with useful Internet information are given in an Appendix.

†It should also be noted that the degree of higher-level variation may increase after taking compositional factors into account. This possibility is considered in more detail later in the paper.

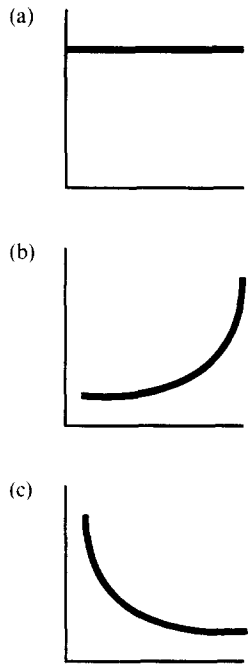

Fig. 5. Between-individual variation.

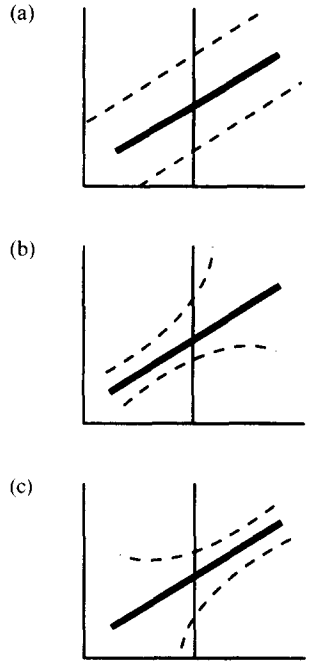

Fig. 6. Between-individual population heterogeneity.

Between-individual heterogeneity

So far, the emphasis has been on heterogeneity in terms of differences between contexts. In technical terms, this has involved elaborating the higher-level random part of the model. There is no reason, however, why we should not also anticipate complex heterogeneity between individuals. Unfortunately, single-level regression techniques are unable to take this into account as only a single random term can be specified to summarise between-people differences. Thus, such techniques assume that whilst people may differ by a "fixed" amount, they have the same degree of variability. Within a multilevel modelling framework it is possible to circumvent this restriction and investigate whether different types of people display different degrees of variability.

To illustrate this possibility we can continue with the limiting long term illness/age example used in the first graphical typology and three possible cases are outlined in Figs 5 and 6. Figure 5 is similar to Fig. 4 but rather than showing how between-context variation changes according to age, it shows how between-individual variation changes with age. Figure 6, meanwhile, shows the overall general limiting long term illness/age relationship which as before is a single thick line with a positive slope. The dashed lines in Fig. 6 show the varying distribution of individuals in the population around the general relationship.

The first case shown corresponds to the standard single-level regression model. There is variation between individuals (Fig. 5(a)) but it is unrelated to age giving the same degree of between-individual variation at all ages in the population around the general relationship (Fig. 6(a)). In the two other cases, however, the degree of variation between individuals changes with age. In Fig. 5(b) we see that between-individual variation increases as age increases so that in the population, Fig. 6(b), there is greater variability in sickness experience amongst older people. In Fig. 5(c) we see the opposite situation whereby between-individual variation decreases as age increases so that in the population, Fig. 6(c), the highest degree of variability in sickness experience is amongst younger people.

Whilst the final two cases violate one of the standard assumptions of traditional regression (the assumption of homoscedasticity), they can be explicitly handled within a multilevel approach by including age in the random part of the micro-model. This approach also applies for categorical predictors. Thus, if we have a variable indicating the respondents' gender not only can we see whether men and women differ on average but also whether they are differentially variable. Besides being of substantive interest in its own right, complex between-individual heterogeneity can have important implications for estimates of between-context heterogeneity as there may be confounding across levels. What may appear to be higher-level, contextual variability may in fact be between-individual, within-context heterogeneity (Bullen *et al.*, 1997). In practical terms, models with complex heterogeneity at both level 2 and level 1 should be fitted followed by an empirical evaluation of all the estimated parameters. In conceptual terms, it can

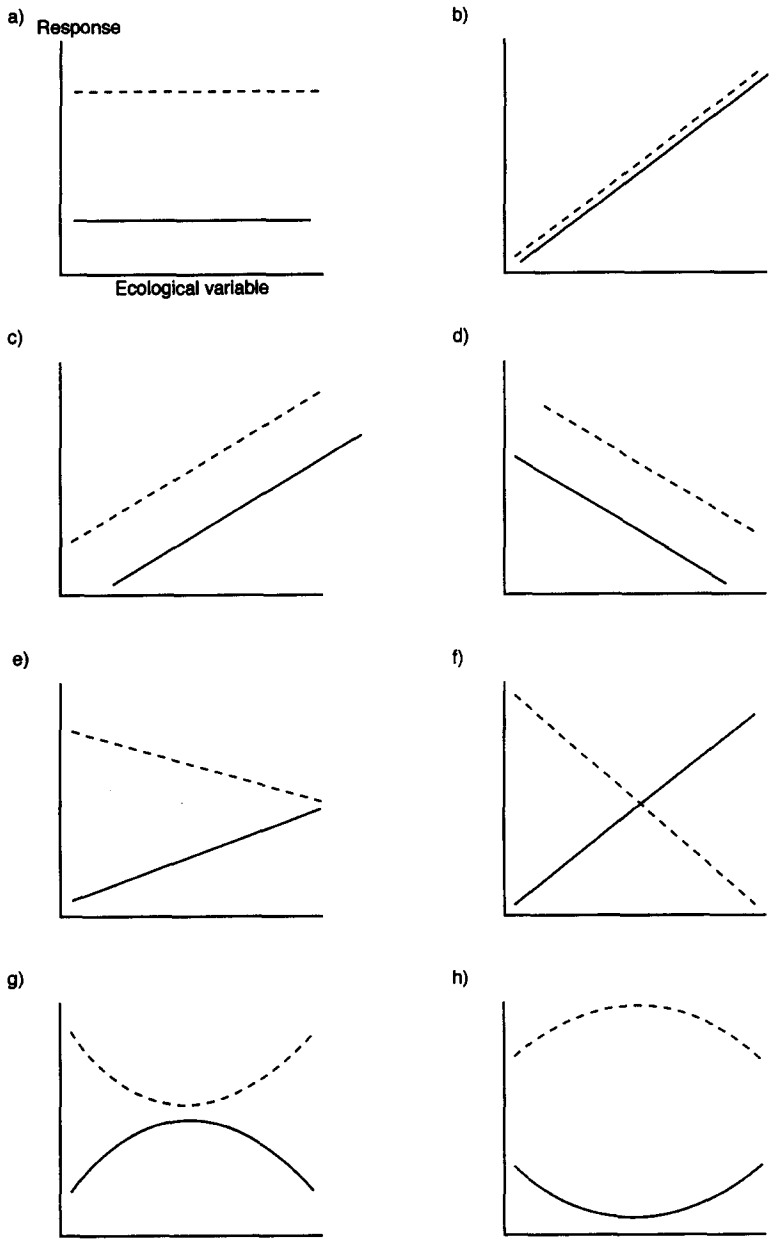

Fig. 7. Individual and ecological cross-level relationships.

be seen that within a multilevel approach heterogeneity and difference—both between people and contexts—are seen as the norm, not an aberration.

Finally, it should be noted that a range of different types of micro models are now applicable thus allowing a variety of different response variables to be handled (Yang *et al.*, 1996). As well as the standard Gaussian model for continuous responses, logit, log–log and probit models can be specified to model proportions and binary outcomes, Poisson and Negative Binomial Distribution models to model counts, and multinomial and ordered multinomial models to model multiple categories. In effect, such models work by assuming a specific distribution for the random part in the micro-model, while maintaining Normality assumptions for higher level random parts.

Cross-level interactions

The distinctive feature of the third graphical typology is that an additional predictor is included in the model that refers not to individual characteristics but to the nature of the places. Again there is a two-level model (individuals in places) with the response being the probability of an individual reporting limiting long term illness. This time, however, the individual predictor variable identifies low social class as opposed to high social class and the

place-level predictor is the ecological variable, the percentage of high social class in a place. A very wide range of differing results based on this model are possible of which a selection are shown in Fig. 7. In this figure, the vertical axes represent the response, the horizontal axes the ecological variable, while the lines on the graphs represent different types of individual (the dotted lines and the solid lines represent low social class and high social class individuals respectively).

Thus, Fig. 7(a) shows that there are marked differences between low and high social class individuals but no ecological effect; Fig. 7(b) represents the converse situation—little difference between low and high social class people but a large ecological effect. The parallel lines of Fig. 7(c) and (d) represent the cases when both the individual and ecological effects are marked and the ecological effect is the same for both social class categories. Figure 7(e) represents the case where the ecological effect is negative for low social class but positive for high social class so that in areas with large percentages of high social class people, the reporting of limiting long term illness by low and high social class individuals converges. Figure 7(f) represents the case where the ecological effects are so marked that the individual effects are actually reversed. Figure 7(g) and (h) show models in which non-linear interaction terms are of importance so that either the smallest or largest ecological effects are found at "middling" levels of the contextual variable. While such concepts of environmental influences have long been of interest to empirical researchers (Davies *et al.*, 1961; Lazarsfeld and Menzel, 1961; Van den Eeden and Hüttner, 1982) and are the subject of considerable conceptual interest, it is only with the development of the multilevel model that the full complexities of such relationships can be effectively analyzed.

A multitude of multilevel models

The fourth and final graphical typology, Fig. 8, reveals the considerable generality of a multilevel approach by showing how the basic two-level structure can be extended to reflect a number of other more complex, yet frequently occurring and substantively interesting, data structures. The two level structure of Fig. 8(a) can be readily extended to the three level structure of Fig. 8(b) with individuals at level 1 nested within local neighbourhoods at level 2 and regions at level 3. Variables can be included at each of the three levels making it possible, for example, to examine the illness/age relation in the context of both local economic prosperity and regional economic prosperity. The extension of the framework to many levels is important as it ensures that any contextual effects are apportioned to the relevant level. As Blaxter has pointed out:

> there is evidence that there is greater heterogeneity within Standard Regions than there is between them. Research has suggested that it is material deprivation within regions that is associated with health (Blaxter, 1990, p. 75)

Health researchers are also obviously interested in how outcome measures change over time and the multilevel framework can be used to reflect context in terms of temporal settings. Figure 8(c) shows how a repeated cross-sectional design can be represented as a multilevel structure. Here level 3 are places, level 2 are years and level 1 are individuals. Thus, level 2 represents repeated measurements of places. Such a structure can be used to examine "outcome trends" within higher level units taking account of their changing compositional make-up. Figure 8(d) shows a repeated measures or panel design in which level 1 is the measurement occasion, level 2 is the individual and level 3 are places. Thus, level 1 represents repeated measurements of individuals. Such a structure allows the assessment of individual change within contextual setting. Unlike conventional repeated measures methods which require a fixed set of repeated observations for all persons (Ware, 1985), both the number of observations per person and the spacing among the observations may vary in a multilevel approach.

Conceptually, the same response measured at different times is no different from many responses measured at one time. Consequently, the multilevel framework can also be used to represent several different, though related, response variables (Duncan *et al.*, 1995b). In the case of health research, this would enable researchers to examine several different measures/dimensions of health status simultaneously, for example a subjective assessment of physical health, a physiological measurement of respiratory functioning and a self-completed psychiatric questionnaire. These different measures would form a set of response variables at level 1, which would nest within individuals at level 2, who would nest within communities at level 3. This form of multilevel structure is shown in Fig. 8(e). In substantive terms, two main benefits arise from a multilevel, multivariate approach. First, the measures are directly comparable in terms of how each is related to individual-level characteristics. Answers to complex questions can be given: for example, is subjective assessment of physical health related to age and socio-economic status in the same way as self-reported assessments of psychiatric well-being? Second, the residual covariance matrix between the set of responses can be estimated at any level so that it is possible to assess the "correlation" of health status measures between individuals and between places, conditional on other variables. A technical advantage is that it is not necessary for measurements to be made on all individuals for all responses.

Since multilevel models can use both continuous and categorical data, the different responses can

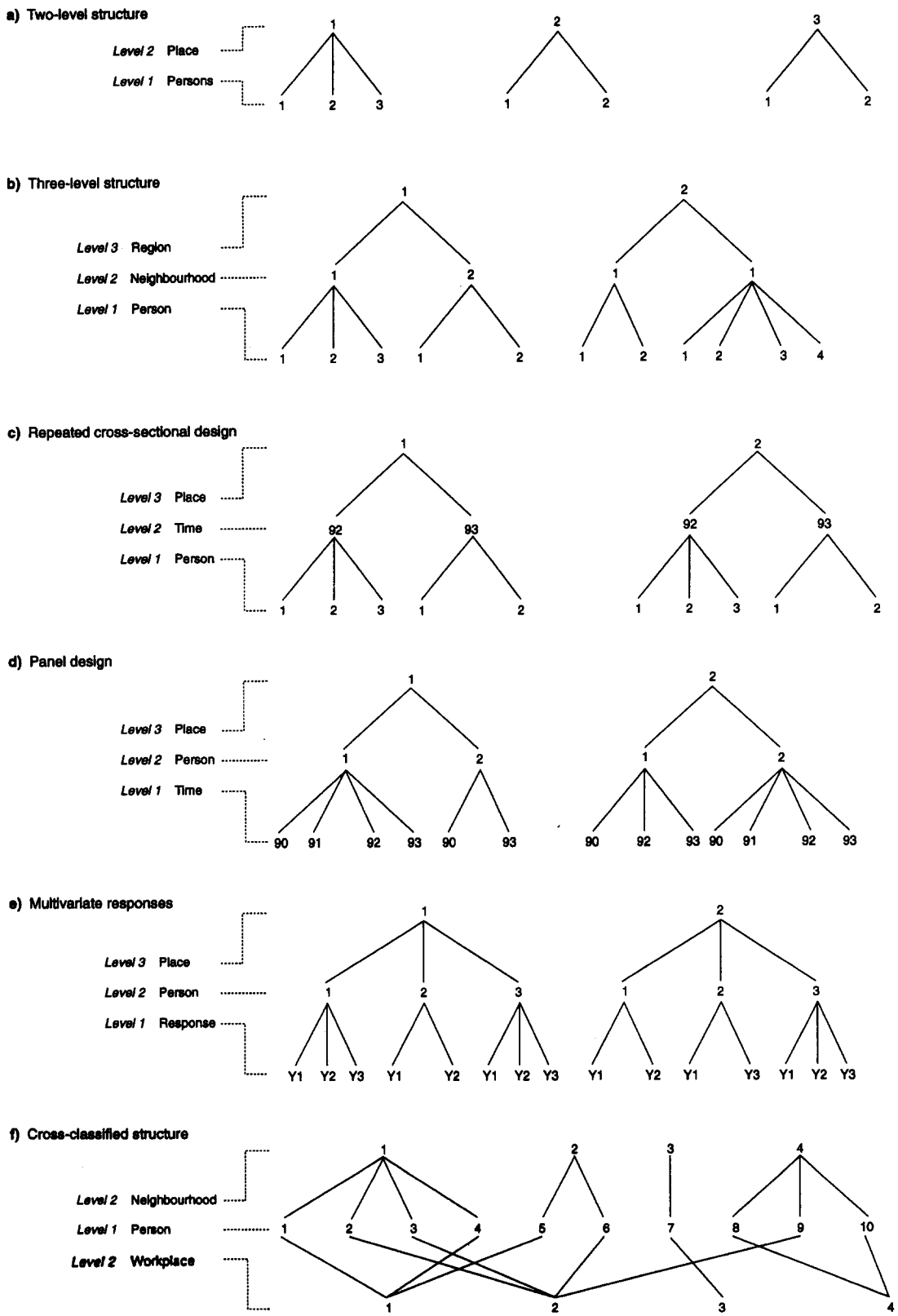

Fig. 8. A range of multilevel data structures.

take the form of continuous values or they can represent systems of classification (eg: low score/high score). Moreover, multivariate multilevel models can accommodate sets of responses which are a mixture of both categorical and continuous variables. This variant of the multivariate multilevel model—known as the mixed, multivariate multilevel model—has great potential as researchers are often interested in two different, though related, dimensions of behaviour. First, whether it actually occurs or not (occurrence), and second, if it does occur, to what degree (quantity). For example, in terms of health-related behaviour, researchers are interested in knowing both who smokes and who does not, and the number of cigarettes consumed by those who are smokers (Duncan *et al.*, 1996).

All the models considered so far have been strictly hierarchical, and contextual effects have nested within each other. It is likely, however, that such a conception does not cover all the necessary possibilities. Contextual sources of variation overlap and more than one may exist at each level. The resulting structure need not be hierarchical. Each lower-level unit may belong to more than one unit at the next higher level. In the case of health outcomes, individuals' health status may be influenced both by where they live and where they work. This can be modelled using a cross-classified structure with individuals at level-1 and both neighbourhood and workplace at level-2. This structure is shown in Fig. 8(f). Explanatory variables can be included for individual-level characteristics and for both level-2 units. Substantively, this allows different contexts to be simultaneously modelled making it possible to identify contextual settings which are having a confounding influence. In the example given here, it may be that what appears as between-workplace variation is in fact really between-neighbourhood variation (Goldstein, 1994).

One other situation where a multilevel approach can be used are survival models where attention is focused on the time to an event. An example would be the time a respondent stays alive after the beginning of a study and we would could relate this to both individual and contextual characteristics. Such an approach does, however, require special methods as the complete survival time is often unknown for many respondents. Further details on work in this area can be found in Goldstein (1995) and Jones *et al.* (1997).

Before illustrating the use of multilevel models by drawing on some existing research one final point concerning their generality should be noted. If we make our models realistically complex (Best *et al.*, 1996) then we may anticipate using a combination of different structures in one analysis. One area where this is particularly likely to occur is in the analysis of panel studies. Many people do not stay in the same context but move from one to other and this could be reflected by combining the repeated measures and cross-classified structures.

EXEMPLIFICATIONS

Geographies of health-related behaviour

A series of studies in both Britain and America have identified significant area variations in the health-related behavioral practices of individuals (Duncan, 1995). All of this work has suggested, either implicitly or explicitly, that these differences cannot be explained in terms of varying population compositions between different areas. In other words, research has overwhelmingly suggested that similar types of people are behaving differently in different types of places. As Blaxter writes, following an analysis based on data from the *Health and Lifestyle Survey* (Cox *et al.*, 1987) conducted in mainland Britain in 1984/5:

> The geographical question, in particular, is obviously quite complicated. Patterns of behaviour by social class may be quite different in different types of area (Blaxter, 1990, p. 203)

The over-riding message from Blaxter and other's work is that geography in terms of contextual effects plays an important role in health-related behaviour. It is a message, however, that is based upon highly simplistic and problematic methods of analysis.

Methodologically, these studies of area variations in health-related behaviour can be characterised as consisting of two different forms of approach, both of which only work at a single level thereby losing the inherent multilevel nature of both the problem and the data. First, the majority of studies are based upon the calculation of simple areal rates which control for population composition by disaggregating or standardising on the basis of personal characteristics (Blaxter, 1990; Balarajan and Yuen, 1986; Cummins *et al.*, 1981; Dunbar and Morgan, 1987). Since this work only considers a very narrow range of personal characteristics simultaneously, (most usually age, gender and social class) it does not provide a rigorous evaluation of the relative importance of contextual effects and compositional effects. Second, a smaller number of studies have adopted a more sophisticated regression modelling strategy to allow for varying population compositions (Braddon *et al.*, 1988; Whichelow *et al.*, 1991). This work is, however, compromised by the technical problems highlighted in the earlier section as only single-level ANOVA procedures have been adopted.

Besides sharing the same crude methodological bases, these studies are all characterised by a simplistic operationalisation of geographical context. To avoid problems arising from small numbers so as to ensure statistically reliable rates, the researchers have been forced to work with extremely coarse

Table 1. Data structure used by Duncan *et al.* (1993)

Level 1:	9003 individuals in the analysis of smoking 6211 individuals in the analysis of drinking Responses: regular smoker or not (binary) : units of alcohol consumed in a week (continuous) Predictors: age, gender, social class, employment status, housing status, marital status, age on leaving school
Level 2:	396 electoral wards
Level 3:	22 regions (The Economist classification)

geographical aggregations. For example, most of the British work has only considered Standard Regions and some has involved analyses based on aggregations of these into larger spatial units (Balarajan and Yuen, 1986). Consequently, this previous research also displays little sensitivity to a more complex and refined regional geography. Furthermore, these studies have not considered a number of different geographical scales simultaneously and so remain open to the charge that area effects may not have been apportioned to the appropriate level. Consequently, the support of this existing research for a contextualised interpretation of health-related behaviour needs to be treated with caution. While the research problem and the data around which it is framed are inherently multilevel, the empirical research that has been conducted has almost exclusively worked only at a single level*.

Since Blaxter's work has such an important position in academic analyses of area variations in health-related behaviour in Britain, a re-analysis of the *Health and Lifestyle Survey (HALS)* data covering smoking and drinking behaviour using multilevel modelling procedures was carried out (Duncan *et al.*, 1993). *HALS* was based on a multi-stage sampling design with individuals, within electoral wards, within constituencies, within regions being randomly sampled. By using a multilevel modelling approach, this hierarchical structure can be explicitly recognised and maintained. It should be emphasised, however, that this structure is not merely a consequence of multistage sampling but, following a contextualised interpretation of individual health-related behaviour, the population itself is conceptualised to have a complex hierarchical structure. Individuals, wards, constituencies and regions are seen as distinct structures, or levels, in the population which may be associated with important processes influencing health-related behaviour.

In the multilevel re-analysis only three of the four sampling levels adopted in the *Health and Lifestyle Survey* were used and the regional classification employed was more refined than that used by the survey team. Rather than using the standard United Kingdom statistical region classification, the wards sampled in *HALS* were grouped according to the Economist's twenty-two region classification (Johnston *et al.*, 1988). Essentially, this grouping represents a sub-division of the standard regions into metropolitan and non-metropolitan areas and is, therefore, considered to be more finely attuned to the regional, "macro" geography of the country. Data collected on two health-related behaviours—cigarette smoking and alcohol consumption—were analyzed and the actual multilevel structure employed is shown in Table 1.

For each behaviour, two models were estimated. First, a "null" model which estimated between-ward and between-region variability without taking into account the different types of people in particular places (no predictor variables were included besides that representing the intercept). Second, a random-intercepts model (like Fig. 1(b)) which included the predictor variables given in Table 1 so that the variability between places was estimated taking into account the population composition of each place. The basic finding from this work was that place-based contextual factors were much less important for health-related behaviour than had previously been stated. Whilst the null models identified some degree of significant ward and regional variation in both smoking and drinking behaviour, this collapsed after population composition was taken into account in the random intercepts model. This result is illustrated at the regional level in Fig. 9 for smoking behaviour. As can be seen, going from the null model to the random intercepts model led to the convergence of most lines around the value 0 which represents the overall national estimate. Hence, once population composition was taken into account, regional differences were only very small. This same finding was found at the ward level for smoking and at ward level and region level for alcohol consumption. Thus, on the basis of this work it seemed that geographical variations in smoking and drinking behaviour in Britain were, to a large degree, an artefact of varying population compositions. The results showed there was much greater variation in behaviour between individuals than between wards and regions.

This conclusion was also somewhat premature, however, as all the models were based on the assumption that there was the same single overall place-effects (or contextual differences) for all types of people. Returning to the section outlining the first graphical typology, only models of the form of Fig. 1(b) were estimated, not Fig. 1(c)–(f). Thus, compositional and contextual effects were only separated in crude terms and no attempts were made to identify complex, person-specific forms of contextuality. This limitation was corrected in later

*One exception to this is some recent work in the USA by Diehr *et al.* (1993) which has adopted a random-effects multilevel approach. While this study is better in technical terms it is still rather limited as it is based on only 15 communities.

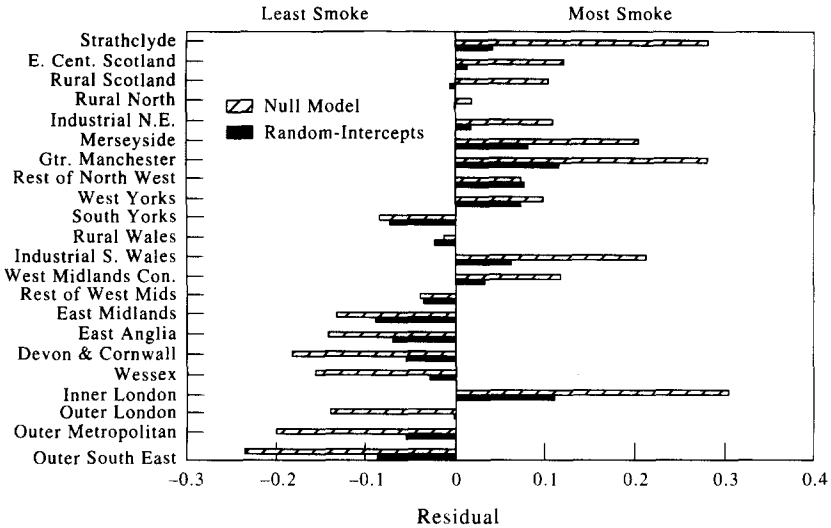

Fig. 9. Model comparison—smoking behaviour between regions (Source: Duncan *et al.*, 1993).

work which applied models like those in Fig. 1(c)–(f) to examine whether particular types of people display different behaviours in different types of places (Duncan, 1995). The results obtained suggested that the earlier multilevel analysis had produced misleading results by only considering overall place-effects rather than people-specific ones. For example, models analyzing between-place variability in alcohol consumption for specific social class categories found substantial amounts of behavioral variability at the ward level for lower social class categories.

As pointed out earlier, however, it is advisable that researchers do not consider complex heterogeneity between contexts without also considering complex heterogeneity between individuals as it is possible for the two to be confounded. Consequently, further models were also estimated which examined whether different types of people displayed different degrees of variability in terms of alcohol consumption. Thus, in these models population composition was taken into account both in terms of how particular types of people behaved on average and in terms of how their behaviour varied. The results obtained suggested that what mattered was between-individual heterogeneity, not between-place heterogeneity. Of particular importance was a gender effect whereby men were not only drinking more on average (a "fixed" effect) but were also substantially more variable amongst themselves in terms of their drinking practices. When this dimension of population composition was taken into account, the contextual effects identified previously for the lower social class categories were considerably reduced. Thus, for the particular geographies used, it appeared that contextual or place effects do play only a minor role and what is most significant are compositional effects. In terms of health-related behaviour, it seems that area variations do not mean that areas *per se* have significance.

This work on health-related behaviour shows that multilevel techniques can provide quite different results and findings to those obtained using traditional single-level techniques. The example is interesting because it reveals that, while the technique represents a powerful means of investigating complex forms of contextuality, it carries no built-in assumptions regarding the importance of context (Mason, 1991). Multilevel models are just as capable of showing that context does not matter when theoretical and empirical work has suggested it might, as they are of revealing the opposite situation. It should also be noted, however, that the reverse situation to that given by the present example is also possible. Failing to take account of population composition can lead to genuine contextual effects being hidden or masked. This could occur, for example, when a place with genuinely low illness rates has relatively high numbers of people who, on the basis of their individual characteristics, have a high probability of being ill. Exactly this situation has been found in a recent multilevel analysis of health outcomes in Britain with the authors reporting results suggesting "areal differences are greater when the compositional effect of their populations is allowed for" (Shouls *et al.*, 1996, p. 372).

Comparisons of institutional performance

As our second example, we consider work concerned with notions of context and composition in relation to health service performance. Over the last ten years, quantitative measures of performance have become an established means of assessing the efficiency and effectiveness of health care delivery systems in both Britain and America. In Britain, a

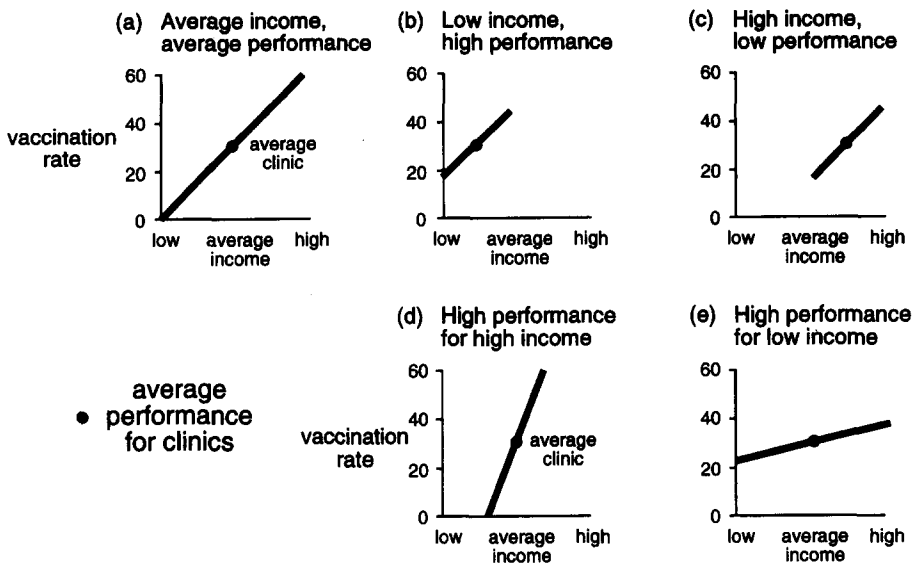

Fig. 10. Working at a single level: the problems of averages.

whole series of performance indicators are now routinely available covering both primary and secondary health care provision although it is only in Scotland, like America, that there are currently "death tables"*.

Measuring health service performance is a highly contentious issue. For one thing, there is the fundamental issue of data quality. Whether it is due to non-recording, simple human error or wilful manipulation, many critics argue that the data available is not of sufficient quality to be able to make reasonable judgements concerning the performance of different service providers (Goldstein and Spiegelhalter, 1996). Even if the data is good enough, there are a series of important statistical issues that have to be appreciated and recognised. Some of these issues will now be considered in light of work carried out on variations in vaccination uptake between clinics, although the discussion applies to performance measurement in general.

Traditionally, analyses of variations in vaccination uptake between clinics have been based on simple crude aggregate rates. As work by Jones and Moon (1990, 1991) and Jones *et al.* (1991) demonstrates, however, such rates are seriously deficient for they mean that the analysis is being carried out at the wrong level. If we assume for the moment that mothers from high income families are much more likely to have their children immunised, then as Fig. 10 shows, exactly the same average clinic performance can be achieved in a number of different ways depending on the financial status of the mothers that particular clinics serve. Clinic (a) is truly "average" in performance and mother's income; (b) is a "good" clinic bought down to the average by having many low income mothers; (c) is a "poor" clinic brought up to the average by having many high income mothers; (d) is an "elitist" clinic which achieves its overall average performance by being a "good" clinic for children of high income mothers and being a "poor" clinic for children of low income mothers; (e), meanwhile, is an "egalitarian" clinic which performs "well" for children of low income mothers at the expense of "good" performance for children of high income mothers.

Figure 10 emphasises that crude clinic averages are composed of three distinct sources of variation:

$$\begin{array}{ccc} \text{AVERAGE} & & \text{COMPOSITION} \\ \text{CLINIC} & = & \text{OF THE} \\ \text{PERFORMANCE} & & \text{CLINIC} \end{array}$$

$$\begin{array}{ccccc} \text{CONTEXTUAL} & & \text{COMPOSITION/} \\ + \quad \text{CLINIC} & + & \text{CONTEXTUAL} \\ \text{DIFFERENCE} & & \text{INTERACTION} \end{array}$$

In this case "composition" refers to the clinic make-up in terms of the financial status of the family that children come from to be immunised; the "contextual" is the overall difference a clinic makes irrespective of client make-up; the interaction term represents the differential effect of a clinic in relation to the make-up of its client list. Crude averages in conflating these distinct sources of variation are un-interpretable and meaningless. The contextual differences of "good", "poor", "elitist" and "egalitarian" are rendered "average" by differing client compositions in clinics (b) and (c), and by

*As has been reported recently, it is likely that such tables will be published in England and Wales within the near future (*The Guardian*, 'Dorrell tests death leagues', July 3, p. 2, 1996). Such tables are considered in more detail in McKee and Hunter (1995).

Table 2. Data structure for multilevel assessment of immunisation uptake performance (Jones and Moon, 1991)

Level 1:	2048 infants aged under 2 years Response: completed pertussis immunisation schedule or not (binary) Predictors: mother's age, employment status, family stability, housing tenure, mother's smoking behaviour, previous death of infant
Level 2:	126 general practices in a city in Southern England

substantial interactions in clinics (d) and (e). While Fig. 10 is based on a single compositional variable, the financial status of the child's mother, the same problems could apply with equal force to other child characteristics.

Multilevel models provide an important way of circumventing these problems since they are able to provide measures of performance which take into account the characteristics of clients while simultaneously recognising uptake may be affected by the characteristics of particular service providers. In this instance, the degree of higher level variation which remains after including level-1 predictor variables reflects the degree of variability between clinics in uptake performance independent of client composition. As outlined earlier, predictions can also be obtained for specific clinics and these could then be ranked to construct tables indicating comparative performance.

The value of this approach has been demonstrated in a paper by Jones and Moon (1991). The data structure on which their work is based is shown in Table 2. Using two types of multilevel model, they establish a number of important findings. First, there are substantial differences between crude aggregate rates and the multilevel predictions of performance based on the residual variation in a null model even though it takes no account of clinic composition. These differences arise because of the shrinkage procedure outlined earlier in this paper with rates for practices with unreliably small target numbers being shrunk towards the overall performance*. Second, when account is taken of client make-up, there is substantial re-ordering of clinics in terms of performance. What becomes apparent is that while the best practices remain the same (those serving suburban areas characterised by middle class professional households), a number of large clinics based in health centres servicing local authority estates with stable childrearing populations improve their position by over 20 places. The results for this are given in Table 3. As the fixed part of the model shows that, on average, local authority tenants are less likely to have their children immunised, it seems several clinics are working hard to overcome the general tendency for their clients to have low uptakes. There is, therefore, an important degree of contextuality in vaccination uptake performance and the multilevel approach is capable of revealing its socio-spatial dimensions.

By way of conclusion to this second illustration, we would like to flag some warnings concerning the application of multilevel models in comparing the performance of different institutions. While adjusted rankings based on such models certainly constitute an improvement over simple league tables based on crude aggregate rates, they are far from being completely adequate and considerable caution still needs to be exercised. There is, of course, the general problem referred to at the beginning of this section. No matter how sophisticated the model, if the data is the poor, the results will be poor. More specifically, there is some disjuncture between using multilevel models to generate performance measures and their general design. As was emphasised in the discussion of the first graphical typology, a multilevel analysis treats the clinics explicitly identified as a sample from a population so that the main focus is on the *general* "effects" of clinics for *any* group of clients in the future. Immediately, therefore, there is some conceptual distance between multilevel techniques and their use in identifying particular institutions. In more practical terms, what needs to be remembered, yet is often forgotten, is that while it is possible to make predictions for specific institutions based on shrinkage estimators, they are not simply point estimates and degrees of uncertainty are associated both with them and the rankings that derive from them (Goldstein and Spiegelhalter, 1996). As much research is now showing, even after careful modelling which adjusts for client composition and takes into account sample sizes, institution rankings carry such large uncertainty bands that it is extremely difficult to isolate particular institutions. For example, Draper (1995), commenting on work done by Goldstein and Thomas (1993) on school differences in pupil attainment, emphasises that when uncertainty is taken into account the resulting categorisation of performance is necessarily so broad that the majority of schools (70%) can only be accurately located "somewhere in the middle of a large gray area" (Draper, 1995, p. 133). Thus, analyses of institutional performance based on multilevel models need to carry a "health warning" themselves and as Goldstein and Spiegelhalter put it, their results can be used as "screening instruments, but not as definitive judgments on individual institutions" (Goldstein and Spiegelhalter, 1996, p. 397).

*It should be noted that shrinkage estimators do not find favour with everyone and it is important to appreciate how they work for as de Leeuw and Kreft note in relation to work on school performance, it helps to understand 'the frustration of the principal of an excellent school who sees the predictions of success of her students shrunken towards the mean' (de Leeuw and Kreft, 1995, p. 184). The potential for such 'mis-representation' is greatest when sample sizes amongst the higher-level units are extremely unbalanced. This issue is considered in more detail later in this paper.

Table 3. Practices whose rank changed by more than 20 places when account is taken of client make-up. (Source: Jones and Moon, 1991).

Practice No	Uptake (%)	Change in ranking	Type of patients/location
115	66.7	+47	Stable peripheral local authority estate
21*	53.8	+35	Stable peripheral local authority estate
107	50.0	+34	Stable peripheral local authority estate
30*	64.8	+28	Mixed tenures; suburb
105	57.1	+23	Stable peripheral local authority estate
16*	46.1	+23	Stable peripheral local authority estate
106	75.0	+22	Stable inner city local authority estate
70*	71.4	+21	Stable inner city local authority estate
29*	59.5	+21	Mixed tenures; suburb
17*	66.6	+20	Stable peripheral local authority estate
104*	66.7	−24	Inner city working class people; elderly people
65	66.7	−25	Middle class elderly people; urban
102*	66.6	−32	Mature professionals; urban
37	75.6	−36	Mature professionals and elderly people; suburb

* Practice based health centres.

Health outcomes, social class and deprivation

The third and final illustration goes beyond separating context and composition effects to consider the interaction between the two in terms of health outcomes. Differences in health outcomes according to social class and deprivation are well-known both at the level of individuals and at the level of geographical areas (Shouls *et al.*, 1996). What is much less well-known is whether there are any health differences arising from interactions between individual social class position and the social class composition of areas. Or, put another way, does the relationship between social class and health take different forms in different types of places. Work by Jones and Duncan (1995) was conducted to help answer this specific question.

Using data from the *Health and Lifestyle Survey*, a series of models with the underlying structure given in Table 4 was developed based on a response measure consisting of a general subjective assessment made by individuals concerning their own health. A two fold modelling strategy was then followed. To begin, the set of individual level predictors given in Table 4 were included to take account of varying population compositions. After this was done, significant differences between the wards remained. Thus, a distinctive "ecology" of self-assessed health status was apparent.

Attempts were then made to account for these between-ward differences by including higher-level ecological and interaction variables relating to an deprivation index. This index was derived from four Census variables and can be taken to represent the broad socio-economic ecology of the surrounding area in which respondents live. Interaction terms were fitted based on social class, housing tenure, lifestyle and income groups and both linear and quadratic equations were attempted. Figure 11 shows the results obtained for the cross-level interactions between individual social class position and ward deprivation. In this figure, the horizontal axis is the ward deprivation index with negative scores reflecting more affluent wards and positive values less affluent ones whilst the vertical axis is the probability that an individual will assess their health as being poor.

As can be seen, the ecological effects are substantial and they are in the same "direction" for all social groups. As ward deprivation increases, the probability that an individual will assess their health as being poor increases. Since most of the interaction terms are not significant, these relations take the same form for individuals in different social class categories with the exception of social class III manual who are more likely to report poor health as ward deprivation increases. Thus, the analysis reveals a clear relationship between the health measure under consideration and ward deprivation. In terms of self-assessed health, it seems as though differences between individuals cannot simply be reduced to differences in personal characteristics and these results challenge the conclusions reached by Slogget and Joshi discussed at the start of this paper.

SOME IMPORTANT CAVEATS

Mason has impressionistically characterised the introduction of new statistical developments as following a process which begins with initial formulation, welcome reaction and enthusiastic use (Mason, 1995). After this highly successful begin-

Table 4. Data structure used in the work of Jones and Duncan (1995)

Level 1: 9003 individuals	
Response:	self-assessment of health as being poor/fair as opposed to good/excellent;
Predictors:	age, height, gender, social class, employment status, housing tenure, income, behavioral habits
Level 2: 396 wards	
Predictor:	a deprivation index formed from combining four census variables;
Level 3. 198 constituencies	
Predictor:	average annualized household weekly income, 1990–1991

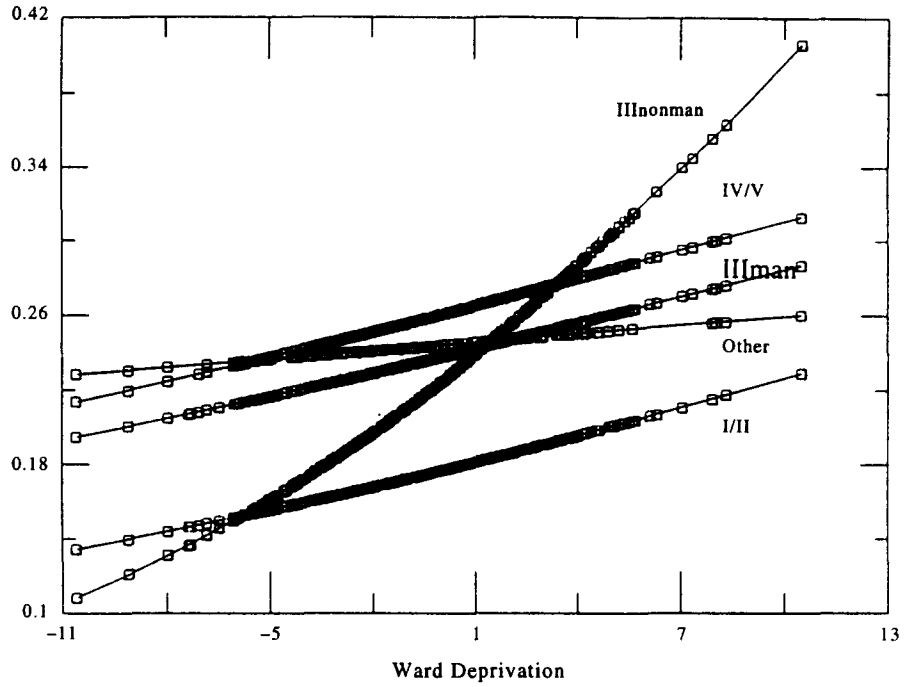

Fig. 11. Individual social class and ward deprivation interactions for self-assessed health (Source: Jones and Duncan, 1995).

ning, there follows a period of critical reflection and close appraisal leading to an accurate and realistic assessment of the technique's value so that, finally, the technique is incorporated within the "social science armamentarium" (Mason, 1995, p. 221) in its appropriate place. In terms of this schema, it is fair to say that multilevel analysis is presently at the stage of critical evaluation which, for many, has heralded a welcome end to an initial period characterised by over-enthusiastic and uncritical application. A series of papers in a special edition of the *Journal of Educational and Behavioral Statistics* provides a valuable outline of the debates surrounding the evaluation of multilevel models (*Journal of Educational and Behavioral Statistics*, 1995, ed. I. G. G. Kreft). In this section, we wish to point out three areas of concern arising from these debates.

The first relates to operationalising context in terms of the hierarchical structure of particular data sets (Mason, 1995). This problem is most formidable in terms of analyses focusing on indeterminate spatial structures rather than more clearly defined institutional ones (though, as Mason (1995) points out, even these should not be accepted uncritically). As Massey notes, "localities are not simple areas you can easily draw a line around" (Massey, 1991) yet many of the health-based applications carried out so far pay little attention to this warning and use the structure of the data set to define higher-level units. These structures often derive from administrative boundaries and whilst they may capture some notion of context at different spatial scales, they are more likely to represent practical conveniences which have no explicit theoretical justification in terms of the outcomes being studied.

The second area of concern relates to sample sizes. To obtain reliable estimates of both within and between-context variation, we require *many* individuals from *many* places. The former allows a precise assessment of within-context relations, the latter a precise assessment of between-context relations, and the two together allow one to be distinguished from the other. Of course, this compromise between the number of higher and lower level units is best achieved by multistage sampling designs. While it is difficult to be specific about the required sample size at each level, some guidance has been offered from educational research, the area in which multilevel modelling has been most applied. Paterson and Goldstein suggest a minimum of 25 individuals in each of 25 groups to do useful work (Paterson and Goldstein, 1992); Bryk and Raudenbush find that with 60 pupils in each of 160 schools it is possible to have a total of four coefficients random at the school level (Bryk and Raudenbush, 1992, p. 203). Obviously, these data requirements are quite substantial and either due to unawareness or lack of financial resources many studies will not be able to satisfy these criteria, particularly with regards to the number of contexts. It is important to note that there are substantial technical implications if multilevel analyses are carried out when the number of higher level units are small. As Draper (1995) emphasises, the established multilevel modelling software packages all rely on estimation strategies involving maximum likelihood

and other large sample techniques. Such techniques are known, however, to produce downwardly biased variance estimates and confidence intervals that are too "short" when the number of level-2 units is small. One way around this problem is to use a fully Bayesian approach and one software package, *BUGS*, is now available for doing this (Gilks *et al.*, 1994).

Finally, it is important to appreciate the implications of how higher level units and their effects are treated in a multilevel analysis. As we have emphasised throughout, the distinguishing feature between multilevel methods and traditional forms of contextual analysis is that the higher level units are seen as a sample drawn from a population and inferences are made about this population. It must be stressed, however, that just because multilevel models operate in this way, it does not guarantee that it is appropriate in any particular instance. As Morris writes, "there are crucial exchangeability judgements embraced (in multilevel models) that sometimes are not carefully considered in applications" (Morris, 1995, p. 198). Thus before using the technique, researchers need to be sure that, first, the sample of contexts they are working with does, in general, come from (can be exchanged with/is similar to) the population that they wish to make inferences about and, second, that this holds true for each specific context they have data for. If, for any reason, a researcher believes that certain contexts are truly unique or that they come from different populations they should not be regarded as exchangeable with the remaining random sample of contexts and they need to be treated as fixed effects. Whilst this can be done empirically after fitting a model and looking for outliers, there are dangers to this approach as the shrinkage estimation procedure may give a misleading picture of similarity. As Draper points out (Draper, 1995), the only way to guarantee exchangeability is at the research design stage. Studies need to be based on data from a representative sample taken from a clearly defined target population. Draper also makes another design-related point. He suggests that when models are based on data from observational designs, nothing can really be said about causality. While these ideas have been challenged by Raudenbush (1995), they do emphasise an important general point which may well have been lost in the enthusiastic rush to use this new technique. Multilevel models are only as good as the data they fit.

CONCLUSIONS

Multilevel models certainly have a number of features that make them attractive in quantitative health research. Technically, by taking into account the complexity of the data, they provide a routine and accurate adjustment for the difficulties associated with autocorrelation. In substantive terms, they provide a coherent framework for framing and testing ideas about contextuality and variability. Importantly, as we have demonstrated, this framework has considerable generality and can be used to tackle a number of different and important questions of interest to health researchers. As writers are increasingly recognising, however, such models are undoubtedly more complex, and, in the words of Draper, this opens up "the possibility of interpretive confusion and overstatement of what may be validly concluded from a given body of evidence" (Draper, 1995, p. 139). Multilevel models undoubtedly have great potential but like all statistical techniques they need to be used carefully and cautiously.

Acknowledgements—The authors would like to acknowledge the useful comments of two anonymous referees and thank the ESRC Data Archive at the University of Essex for providing the Health and Lifestyle survey data (Cox, 1988).

REFERENCES

Alker, H. S. (1969) A typology of ecological fallacies. In *Quantitative Ecological Analysis*, eds. M. Dogan and S. Rokkan. MIT Press, MA, U.S.A.

Balarajan, R. and Yuen, P. (1986) British smoking and drinking habits: regional variations. *Community Medicine* **8,** 131–137.

Best, N. G., Spiegelhalter, D. J., Thomas, A. and Brayne, C. E. G. (1996) Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society Series A* **159,** 323–342.

Blaxter, M. (1990) *Health and Lifestyles.* Tavistock/Routledge, London.

Braddon, F. E. M., Wadsworth, M. E. J., Davies, J. M. C. and Cripps, H. A. (1988) Social and regional differences in food and alcohol consumption and their measurement in a national birth cohort. *Journal of Epidemiology and Community Health* **42,** 341–349.

Britton, M. (1990) Geographical variation in mortality since 1920 for selected causes. In *Mortality and Geography: A Review in the mid-1980's for England and Wales*, ed. M. Britton. HMSO, London.

Bryk, A. S., Raudenbush, S. W., Seltzer, M. and Congdon, R. (1988) *An Introduction to HLM: Computer Program and User's Guide*, 2nd edn. University of Chicago, Department of Education, Chicago.

Bryk, A. S. and Raudenbush, S. W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage, Newbury Park.

Bullen, N., Jones, K. and Duncan, C. (1997) Modelling complexity: analyzing between-individual and between-place variation. *Environment and Planning, Series A* **29,** 585–609.

Carr-Hill, R., Smith, P., Martin, S., Peacock, S. and Hardman, G. (1994) Allocating resources to health authorities: development of methods for small area analysis of use of inpatient services. *British Medical Journal* **309,** 1046–1049.

Carr-Hill, R., Rice, N. and Roland, M. (1996) Socio-economic determinants of rates of consultations in general practice based on fourth national morbidity survey of general practices. *British Medical Journal* **312,** 1008–1013.

Congdon, P. (1995) The impact of area context on long term illness and premature mortality: An illustration of multi-level analysis. *Regional Studies* **29,** 327–344.

Cox, B. D. (1988) *Health and Lifestyle Survey, 1984–5*, (computer file). ESRC Data Archive, Colchester.

Cox, B. D., Blaxter, M., Buckle, A. L. J. *et al.* (1987) *The Health and Lifestyle Survey: A Preliminary Report.* Health Promotion Research Trust, London.

Cummins, R. O., Shaper, A. G., Walker, M. and Wale, C. J. (1981) Smoking and drinking by middle-aged British men: effects of social class and town of residence. *British Medical Journal* **283,** 1497–1502.

Davies, J. A., Spaeth, J. L. and Huson, C. (1961) A technique for analyzing the effects of group composition. *American Sociological Review* **26,** 215–226.

Davies, P. and Gribben, B. (1995) Rational prescribing and interpractitioner variation: a multilevel approach. *International Journal of Technology Assessment in Health Care* **11,** 428–442.

de Leeuw, J. and Kreft, I. (1995) Questioning multilevel models. *Journal of Educational and Behavioral Statistics* **20,** 171–189.

Diehr, P., Koepsell, T., Cheadle, A., Psaty, B. M., Wagner, E. and Curry, S. (1993) Do communities differ in health behaviours? *Journal of Clinical Epidemiology* **46,** 1141–1149.

DiPrete, T. and Forristal, J. (1994) Multilevel models: methods and substance. *Annual Review of Sociology* **20,** 331–357.

Draper, D. (1995) Inference and hierarchical modelling in the social sciences. *Journal of Educational and Behavioral Statistics* **20,** 115–147.

Dunbar, G. C. and Morgan, D. D. V. (1987) The changing pattern of alcohol consumption in England and Wales 1978–1985. *British Medical Journal* **295,** 807–810.

Duncan, C. (1995) Health-related behaviour in context: a multilevel approach. Unpublished Ph.D. thesis. University of Portsmouth.

Duncan, C., Jones, K. and Moon, G. (1993) Do places matter? A multilevel analysis of regional variations in health-related behaviour in Britain *Social Science and Medicine* **37,** 725–733.

Duncan, C., Jones, K. and Moon, G. (1995a) Psychiatric morbidity: A multilevel approach to regional variations in the UK. *Journal of Epidemiology and Community Health* **49,** 290–295.

Duncan, C., Jones, K., and Moon, G. (1995), Blood pressure, age and gender. In *A Guide to MLn for New Users*, ed. G. Woodhouse. Multilevel Models Project, Institute of Education, University of London, London.

Duncan, C., Jones, K. and Moon, G. (1996) Health-related behaviour in context: A multilevel modelling approach. *Social Science and Medicine* **42,** 817–830.

Ecob, R. (1996) A multilevel modelling approach to examining the effects of area of residence on health and functioning. *Journal of the Royal Statistical Society, Series A* **159,** 61–75.

Gatsonis, C., Normand, S.-L., Liu, C. and Morris, C. (1993) Geographical variation of procedure utilisation: a hierarchical model approach. *Medical Care* **31,** YS54–59.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. (1994) A language and program for complex Bayesian modelling. *The Statistician* **43,** 169–178.

Goldstein, H. (1991) Multilevel modelling of survey data. *The Statistician* **40,** 235–244.

Goldstein, H. (1994) Multilevel cross-classified models. *Sociological Methods and Research* **22,** 364–375.

Goldstein, H. (1995) *Multilevel statistical models.* Edward Arnold, London.

Goldstein, H. and Spiegelhalter, D. (1996) League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A* **159,** 385–443(with discussion).

Goldstein, H. and Thomas, S. (1993) *Guardian A-Level analysis 1993: Technical Report.* Institute of Education, University of London, London.

Gould, M. and Jones, K. (1996) Analyzing perceived limiting long-term illness using U.K. Census Microdata. *Social Science and Medicine* **42,** 857–869.

Hedeker, D., McMahon, S., Jason, L. and Salina, D. (1994) Analysis of clustered data in community psychology: With an example from a worksite smoking cessation project. *American Journal of Community Psychology* **22,** 595–615.

Hox, J. and Kreft, I. (1994) Multilevel analysis methods. *Sociological Methods and Research* **22,** 283–299.

Johnston, R. J., Pattie, C. J. and Allsopp, J. G. (1988) *A Nation Dividing*? Longman, London.

Jones, K. (1991) Specifying and estimating multilevel models for geographical research. *Transactions of the Institute of British Geographers* **16,** 148–159.

Jones, K. and Bullen, N. (1994) Contextual models of house prices: a comparison of fixed- and random-coefficient models developed by expansion. *Economic Geography* **70,** 252–272.

Jones, K. and Duncan, C. (1995) Individuals and their ecologies: analyzing the geography of chronic illness within a multilevel modelling framework. *Health and Place* **1,** 27–40.

Jones, K., Gould, M. I. and Duncan, C. (1997) Death and deprivation: an exploratory survival analysis of deaths in the Health and Lifestyle Survey. *Paper presented at the Annual Conference of the Royal Geographical Society/Institute of British Geographers*, Exeter, January 1997.

Jones, K. and Moon, G. (1990) A multilevel model approach to immunisation uptake. *Area* **22,** 264–271.

Jones, K. and Moon, G. (1991) Re-assessing immunization uptake as a performance measure in general practice. *British Medical Journal* **303,** 28–31.

Jones, K., Moon, G. and Clegg, A. (1991) Ecological and individual effects in childhood immunisation uptake: a multilevel approach. *Social Science and Medicine* **33,** 501–508.

Kreft, I. G. G. (Ed.) (1995) Hierarchical linear models: Problems and prospects (1995) *Journal of Educational and Behavioral Statistics* **20** (2), Special Issue.

Kreft, I. G. G., De Leeuw, J. and Van der Leeden, R. (1994) Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *The American Statistician* **48,** 324–335.

Langford, I. and Bentham, G. (1996) Regional variations in mortality-rates in England and Wales: An analysis using multilevel modelling. *Social Science and Medicine* **42,** 897–908.

Lazarsfeld, P. F. and Menzel, H. (1961) On the relation between individual and collective properties. In *Complex Organisations*, ed. A. Etzioni. Holt, Reinhart and Winston, New York.

Leyland, A. (1995) Examining the relationship between length of stay and readmission rates for selected diagnoses in Scottish hospitals. *IMA Journal of Mathematics Applied in Medicine and Biology* **12,** 175–184.

Leyland, A. and Boddy, F. A. (1997) Measuring performance in hospital care: length of stay in gynaecology. *European Journal of Public Health* **7,** 136–143.

Longford, N. T. (1993) *Random Coefficient Models.* Clarendon Press, Oxford.

Macintyre, S. (1986) The patterning of health by social position in contemporary Britain: directions for sociological research. *Social Science and Medicine* **23,** 393–415.

Mason, W. M. (1991) Problems in quantitative comparitive analysis: ugly ducklings are to swans as ugly scatter plots are to . . . ? In *Macro-Micro Linkages in Sociology*, ed. J. Huber. Sage, Newbury Park, CA.

Mason, W. M. (1995) Comment. *Journal of Educational and Behavioral Statistics* **20,** 221–227.
Mason, W. M., Wong, G. Y. and Entwistle, B. (1984) The multilevel model: a better way to do contextual analysis. In *Sociological Methodology*, ed. S. Leinhardt. Jossey-Bass, San Francisco.
Massey, D. (1991) The political place of locality studies. *Environment and Planning Series A* **23,** 267–281.
McKee, M. and Hunter, D. (1995) Mortality league tables: do they inform or mislead? *Quality in Health Care* **4,** 5–12.
Morris, C. (1983) Parametric empirical Bayes. *Journal of the American Statistical Association* **78,** 47–65.
Morris, C. (1995) Hierarchical models for educational data: an overview. *Journal of Educational and Behavioral Statistics* **20,** 190–199.
Paterson, L. and Goldstein, H. (1992) New statistical methods for analyzing social structures: an introduction to multilevel models. *British Educational Research Journal* **17,** 387–393.
Rasbash, J. and Woodhouse, G. (1995) *MLn Command Reference*. Multilevel Models Project, Institute of Education, University of London, London.
Raudenbush, S. W. (1995) Reexamining, reaffirming and improving the application of hierarchical models. *Journal of Educational and Behavioral Statistics* **20,** 210–220.
Robinson, W. S. (1950) Ecological correlations and the behaviour of individuals. *American Sociological Review* **15,** 351–357.
Shouls, S., Congdon, P. and Curtis, S. (1996) Modelling inequality in reported long term illness in the UK: combining individual and area characteristics. *Journal of Epidemiology and Community Health* **50,** 366–376.
Silk, J. (1977) *Analysis of Covariance and Comparison of Regression Lines. Concepts and techniques in modern geography*, p. 20. Geobooks, Norwich.
Skinner, C., Holt, D., and Smith, T. F., (eds) (1989) *The Analysis of Complex Surveys*. Wiley, New York.
Sloggett, A. and Joshi, H. (1994) Higher mortality in deprived areas: community or personal disadvantage? *British Medical Journal* **309,** 1470–1474.
Van den Eeden, P. and Hüttner, H. J. M. (1982) Multilevel research. *Current Sociology* **30,** 1–181.
Von Korff, M., Koepsell, T., Curry, S. and Diehr, P. (1992) Multilevel analysis in epidemiologic research on health behaviours and outcomes. *American Journal of Epidemiology* **132,** 1077–1082.
Ware, J. H. (1985) Linear models for the analysis of longitudinal studies. *American Statistician* **39,** 95–101.
Whichelow, M. J., Erzinclioglu, S. W. and Cox, B. D. (1991) Some regional variations in dietary patterns in a random sample of British adults. *European Journal of Clinical Nutrition* **45,** 253–262.
Wu, Y.-W. B. (1995) Hierarchical linear models: a multilevel data analysis technique. *Nursing Research* **44,** 123–126.
Yang, M., Goldstein, H. and Rasbash, J. (1996) *MLn macros for advanced multilevel modelling*. Multilevel Models Project, Institute of Education, University of London, London.

APPENDIX

Contact addresses for the following software packages are as follows:

- ***BMDP*** is available from:

BMDP Statistical Software Inc

1440 Sepulveda Blvd. Suite 316

Los Angeles, CA 90025

USA

- *BUGS* (Bayesian inference Using Gibbs Sampling) (http://weinberger.mrc-bsu.cam.ac.uk/bugs/) is available from:

MRC Biostatistics Unit

Institute of Public Health

Robinson Way

Cambridge CB2 2SR

UK

e-mail: bugs@mrc-bsu.cam.ac.uk

- *GENSTAT* is available from:

NAG Ltd

Wilkinson House

Jordan Hill Road

Oxford OX2 8DR

UK

- *HLM* is distributed by:

ProGamma, (http://www.gamma.rug.nl/)

Iec ProGAMMA

P.O. Box 841

9700 AV Groningen

The Netherlands

e-mail: gamma.post@gamma.rug.nl

and by

Scientific Software Inc.

1525 East 53rd St.,

Suite 906,

Chicago, IL. 60615

- *MLn* is distributed by Pro Gamma and by the Institute of Education (http://www.ioe.ac.uk/multilevel):

Multilevel Models Project

Mathematical Sciences Centre

Institute of Education,

20 Bedford Way

London WC1H 0AL, UK

e-mail:temsmya@ioe.ac.uk

- *SAS* is available from:

SAS Institute Inc

SAS Campus Drive

Cary, NC 27513

USA

- *VARCL* is distributed by ProGamma (see above)

There is an e-mail discussion list devoted to multilevelmodels. To join the list, send a message to:

mailbase@mailbase.ac.uk

The message should contain a single line, with a command of the form:

join multilevel ⟨firstname(s)⟩ ⟨lastname⟩

The following home pages on the Internet also provide a valuable resource:

- University of London Institute of Education's Multilevel Models Project:

http://www.ioe.ac.uk/multilevel/index.html

or its mirror site at the University of Montreal

http://www.medent.umontreal.ca/multilevel

- Michigan State University's Longitudinal and Multilevel Methods Project:

http://www.edcu.msu.edu/units/LAMMP