RESEARCH

Brief No: RB791
August 2006
ISBN 1 84478 797 2

CROSS-CLASSIFIED AND MULTIPLE MEMBERSHIP STRUCTURES IN MULTILEVEL MODELS: AN INTRODUCTION AND REVIEW

Antony Fielding, University of Birmingham
Harvey Goldstein, University of Bristol

Context of the review

This review provides an introduction to the role of statistical modelling of complex social situations, in particular considering the application of advanced methodologies to an educational context. By building an overview of techniques such as cross classified multi-level modelling and multiple membership structures and through a comprehensive review of the application of these techniques the review should inform advances in educational research. The application of these advanced statistical techniques which allow for hierarchically structured data will assist the Department in accurately identifying the individual and independent effects of various factors on attainment and in particular the simultaneous effects of area of residence effects and school effects in educational production functions.

Explanatory models using standard multiple regression

The aim of many statistical models is to try and account for the variation in some response variable by a set of one or more explanatory variables or effects. The most basic of these is the familiar multiple regression model. Explanatory variables are introduced both in their own right, since there is substantive interest in them, but also as control variables. The latter often adjust for indirect effects so that one can look at the individual and independent effects of variables of interest. Thus, for example, in studies of educational outcomes of children one might be interested in the predictive influence of prior measures of ability. For the reasons stated, also included in a regression model might be other individual characteristics such as gender or measures of social circumstances. The use of such regression models, assumptions required and limitations in many contexts are widely reported in the literature.

Hierarchical structures and data and multilevel modelling

Multilevel models extend regression methods to cope with more complex situations and data structures which are not explicitly recognised by standard multiple regression assumptions and methods. The simplest hierarchical model, which can be extended to more levels, is two level where it is recognised, for example, that children are nested within schools. Schools exert an effect on children's performance in various ways and this must be recognised. A three level model might recognise that schools in turn are nested within local education authorities each with its own effect. Multilevel modelling methodology which is a form of random effects model is necessary because standard regression which does not respect the clustered structure may result in inappropriate statistical estimation of effects. However and perhaps more importantly, it allows the fuller detailed study of effects and explanations operating at various levels in the structure. In studies of Key Stage 1 performance, for example, one may wish to investigate explanatory factors which may include individual child characteristics and in particular baseline measures taken at reception into primary schools. A basic multilevel model would recognise that children are clustered in schools and would incorporate school random effects in the model which might then allow average performance, net of child effects, to vary between schools. However, since appropriate modelling frameworks and estimation procedures have been set up more extensive questions can be investigated.

department for
education and skills
creating opportunity, releasing potential, achieving excellence

For instance, it may turn out that the effect of baseline ability may vary between schools. A school which performs well with low ability children might not perform so well with those of higher ability. Multilevel modelling has become the accepted way of handling such structures and its approaches can address a range of interesting and sometimes complex questions.

Cross–classified data structures and their modelling

As work on multilevel methodology, software, and range of applications developed it was discovered that there was often much more complexity in data structures than was permitted by basic hierarchical formulations. This led to developments of the methodology to handle this complexity and associated separation of effects. Children for instance are nested not only in schools but may form the lowest level in an area of residence hierarchy. Children from the same school may not necessarily come from the same neighbourhood. Further, children living in the same area may not necessarily attend the same school. We still have a two level hierarchy but the units in which children are nested are now combinations of particular schools and particular areas.

A model which recognises this and allows the separation of school and area effects and their possible elaboration is known as a cross-classified multilevel model. A feature of such structures which also means special statistical procedures have to be developed to handle them is their lack of balance. It is not uncommon to use our example to find concentrations of children in particular cell combinations of school and area. Many combinations exhibit sparsity in that they contain very few children and a majority of combinations may be empty. Other examples of cross-classified structures of varying degrees of further complexity may be found in repeated measures studies where there may be complex crossed effects of such factors as occasion, assessors (raters), and measuring instruments.

A further good example was reported on the continuity of school effects on secondary school performance. Students are not only nested in their secondary school but also in the primary school they previously attended. The cell combinations of primary school and secondary schools form a level 2 cross classification. Applications of cross-classified models have often revealed that primary school effects and factors associated with them may be just as influential on later performance as factors associated with current secondary school. These examples give a flavour of the nature of cross-classified situations. However, very complex situations can arise with crossings by many factors at each level and also at several levels. Methodology for cross-classified multilevel models has proved quite flexible in handling a variety of such situations.

Multiple membership structures, data and models

Extending cross classifications to multiple membership structures allows for a greater degree of complexity and assist in disentangling the independent effects of each factor. Such structures arise in situations where lower level units in a nested hierarchy can be members of more than one higher level unit simultaneously. Education examples may illustrate such situations. Suppose one is studying progress in secondary schools from age 11 to Key Stage 3 and it might be supposed that the whole school experience over that period is influential. For a student who stayed in the same school the nesting in the school and its effect is unique. However, what of a student who moves schools in the period? That student then 'belongs' to several schools. Statistical models and procedures to disentangle estimated effects of various schools are thus required. If in addition the level 2 in the hierarchy was a cross-classification of area of residence and school, some students may have moved areas too and are also in a multiple membership relation to areas. Such combinations of cross classified effects in multiple membership situations are not uncommon, particularly in education. Combined effects from units for which individuals have multiple memberships are represented in statistical models by weighting them in some sensible and convenient way. Teacher effects have been handled in this way in some applications in the literature. Thus students doing a 2 year A level course in a subject are nested within a teaching group. If that group was taught solely by one teacher and teachers taught several groups a cross-classified model of teacher and group could be developed and estimated. However, this situation is not as common as might be supposed. A group may be taught 50 % of the time by one teacher and 25% of the time by two others. One way of weighting the effect contribution of the three teachers is in proportion to this time involvement. A variety of fairly complex situations which have been handled in the applications literature may be formulated using such multiple membership ideas.

Estimation methodology and software issues

The algebraic and diagrammatic representation of these structures is often a useful aid to the understanding of statistical approaches to fitting the complex effects involved. They are explored in the full report. There are two broad approaches to estimating what are referred to as linear mixed models for multilevel and other complex structures. Up until fairly recently it was mostly done through the maximum likelihood framework though the technicalities of its implementation vary across software. There is now a wide variety of software which can handle the complex models involved with varying degrees of ease of user implementation and restrictivity. Reviews of such software which compare and contrast these implementations may be found on the Centre for Multilevel Modelling webpage (http://www.mlwin.com/softrev/revr.html).

Recently there have been developments using the alternative types of estimation procedure known as Monte Carlo Markov Chain (MCMC). For many situations, though not all, this has advantages from statistical, interpretational and computational perspectives. Its main platforms are the packages MLwiN and WinBUGS. In the main the report concentrated on linear models for continuous variable responses. However, procedures and software are also available for generalised linear mixed models which focus on binary responses (e.g. succeed or not), count data (disease incidence), ordered categories (grades in an examination) or categorised responses.

Review of applications of models for complex structures in the literature

The search of the literature has revealed a wide variety of disciplines and research areas apart from education where the complex modelling under discussion has been successfully applied. Many have design and structural features which parallel those in education which are not yet exploited and point the way to their greater use in this context. Areas and applications reviewed include health research, interviewer effects in surveys, social networks, epidemiology, veterinary science, animal ecology, genetics, transportation, psychometrics, test reliability, and forestry.

Conclusion and additional comments

A number of current issues and developments may be mentioned in conclusion and which often may complement the structural issues discussed. For instance, models with multivariate responses (e.g. several GCE Advanced Level scores) may now be satisfactorily used and embedded in complex structures. Event history and the modelling of time in various states (e.g. employment status) is also a recent development in this context. Multilevel factor analysis methods are being investigated as is the study of measurement error and the incorporation within complex structures of methods for handling missing data.

The application of complex statistical modelling has now begun to mirror the complex social reality as a context in which data is collected. This often yield new insights. Complex modelling may even become routine as a corpus of statistical methodology. However, there are caveats. Statistical modelling is not a substitute for grounded substantive theory, though it may help develop and evaluate such theory. There is a danger that routine implementation becomes a substitute for thinking. Multilevel methodology is becoming an invaluable tool in empirical research but that empiricism must be embedded in a full understanding of the complexity it involves.

Additional Information

Copies of the full report (RR791) - priced £4.95 - are available by writing to DfES Publications, PO Box 5050, Sherwood Park, Annesley, Nottingham NG15 0DJ.

Cheques should be made payable to "DfES Priced Publications".

Copies of this Research Brief (RB791) are available free of charge from the above address (tel: 0845 60 222 60). Research Briefs and Research Reports can also be accessed at www.dfes.gov.uk/research/

Further information about this research can be obtained from Jen Halmshaw, Room 5A18, DfES, Sanctuary Buildings, Great Smith Street, London SW1P 3BT.

Email: jen.halmshaw@dfes.gsi.gov.uk