MULTILEVEL MODELLING NEWSLETTER

The Multilevel Models Project:

Dept. of Mathematics, Statistics & Computing
Institute of Education, University of London

20 Bedford Way, London WC1H 0AL, ENGLAND
e-mail: temsmya@uk.ac.lon.ioe

Telephone: 071-612-6682 Fax: 071-612-6686

Vol. 4 No. 1 January, 1992

IN THIS ISSUE

A article by *Lindsay Paterson* summarizes an investigation into using multilevel methods to model social segregation among Scottish Secondary Schools from surveys in 1978, 1980, 1984 and 1988. Some interesting points are raised by the author. Readers' comments are invited.

A article on three level growth models for educational achievement by *Huub van den Bergh* and *Hans Kuhlemeier* show that how long summer holidays can affect pupils' learning at the beginning of a school year with the social economic background of students and the absenteeism of respondents being taken account into the modelling.

News on:

Multilevel models workshop in April and May

New dates for multilevel analysis clinic

Data library and macros for multilevel analysis

ESRC seminar series

CONTRIBUTORS

Thanks very much to the people who provided articles for this issue.

Lindsay Paterson
Centre for Education Sociology
Edinburgh University
Edinburgh EH8 9LW
Scotland

Huub van den Bergh
State University of Utrecht
Faculty of Humanities
Trans 10 3512 JK Utrecht
The Netherlands

Hans Kuhlemeier
Cito
postbox 1034
6801 MG Arnhem
The Netherlands

The Multilevel Models Project is supported by a grant from the Economic and Social Research Council, U.K.

APPLICATIONS

Multilevel Modelling and Segregation Indices

Lindsay Paterson

1. Introduction

Segregation indices have been used for over 50 years to measure the spread of social groups across, for example, schools or neighbourhoods. However, they suffer from the usual problems of analysis based on aggregated data. In particular, they provide no means of understanding the processes by which segregation comes about (*Duncan and Duncan, 1955*). This article is a summary of an investigation into using multilevel methods to model social segregation among Scottish secondary schools. The topic is of sociological interest because open enrolment was introduced in Scotland in 1981, and there was disagreement among proponents and opponents of the policy as to whether it would decrease or increase segregation (*Adler et al, 1989*).

A fuller account of the work is provided by *(Paterson 1991)*.

2. Data

The data are taken from the Scottish School Leavers Surveys of 1978, 1980, 1984 and 1988. These were postal-questionnaire surveys, with samples selected systematically from a frame ordered by birth-date within school. Response rates were around 80%, and post-stratification weights defined by gender and examination attainment are used to compensate for the remaining non-response. Sample sizes are shown in part (1) of the Table.

3. Measures of segregation

We investigate the extent to which schools differ in the proportion of their pupils who are not middle class. Let pi be the proportion of school leavers in a given year from school i whose father's occupation is not classified into the UK Registrar Generals' categories I or II *(OPCS, 1970, 1980)*; this group of occupations is referred to as the designated group. The proportion in the group in each year is shown in part (2) of the Table.

Segregation indices are attempts to summarise the extent to which the p_i differ. *(James and Taeuber 1985*) provide a review. An example is the Variance Ratio Index, which can be defined in terms of the binary indicator of membership of the designated group: it is the fraction of the variance of that variable that lies at the school level.

There are two problems with the conventional indices. The first is technical. When they are calculated from samples within units (here, within school) they are biased. For further details of this problem, see *(Paterson 1991)*.

The second problem is the one we are concerned with here: they are too crude to provide much of an understanding of the process of segregation. Some of the indices do give some insight. *Atkinson* has developed an index which, by means of a variable parameter, can be made particularly sensitive to schools with p_i in a particular range. This could be used to investigate whether, for example, segregation was changing more rapidly in predominantly middle-class schools than elsewhere. The segregation curve can be used for the same purpose (*James and Taeuber, 1985*). However, even *Atkinson*'s index and the segregation curve cannot be straightforwardly adapted to examine groupings of schools defined in other ways, nor to assess the role of parental or pupil characteristics in explaining segregation.

A more fundamental version of this second problem is that the conventional indices use the observed proportions p_i, rather than the underlying social probabilities that generated them. Yet, to understand social processes,

we have to be able to distinguish between mechanisms and outcomes. The distinction between proportions and probabilities parallels the distinction between finite-population and super-population inference which is common in the literature on sample surveys (*Skinner et al, 1989*).

4. Multilevel modelling

These considerations suggest that a modelling approach is needed; and the nature of the problem suggests that the modelling should be multilevel. The standard multilevel theory provides an obvious way of generalising the Variance Ratio.

The variable to be modelled is the binary indicator of membership of the designated group. In a given year, define $y_{qi} = 1$ if the q^{th} sample member from school i is in that group, and $y_{qi} = 0$ if not. Then we can arrive at a direct generalisation of the Variance Ratio by means of a linear multilevel model:

$$y_{qi} = \underline{\beta}^T \underline{x}_{qi} + u_{1qi} + u_{2i} \qquad (1)$$

where $\underline{x}_{qi}$ is a vector of explanatory variables, $\underline{\beta}$ is a vector of parameters to be estimated, and u_1 and u_2 are error terms respectively at the individual and at the school level, with expectations 0 and variances v_1 and v_2. The distributions of the errors are Normal, and are independent of each other and of errors at the same level for different values of q or i. The expectation term $\underline{\beta}^T \underline{x}_{qi}$ is the probability of belonging to the designated group. Non-linear models for y_{qi} are mentioned briefly in Section 5.

Then the multilevel variance ratio is the proportion of the variance in y that lies among schools:

$V^{ML} = v_2/(v_1 + v_2)$.

The simplest analogue of the Variance Ratio is defined when $\underline{x}_{qi}$ is just a constant term. But the advantage of modelling is that $\underline{x}_{qi}$ can be used to represent the social processes that might have generated the segregation. If we introduce a variable x in this way, and the value of V^{ML} falls as a result, then we can conclude that x has contributed to an explanation of the segregation. The variable x could be at the pupil or at the school level. An example of a pupil-level variable is the educational background of the pupil's parents. An example of a school-level variable is an indicator of whether the school is predominantly middle-class.

Having shown that a particular x does change the value of V^{ML}, the next step is to decompose V^{ML} into different values for different levels of x. For example, if we define x_i to be 1 for middle-class schools and 0 for others, then, by modelling the error terms u_1 and u_2 as linear functions of x_i, we can get separate values of V^{ML} for middle-class schools and other schools.

Some examples of modelling are in the Table. All the estimation was done by the programme *ML3 (Prosser et al, 1990)*; the standard errors of the variance ratios were calculated using the information matrix of the estimates and a standard formula for the variance of a ratio *(Kish, 1965, pp.206-208)*. The school-level residuals from the models were unimodal, but also slightly skew, and definitely flatter than Normal; however, this shape is probably reasonably satisfactory because of the large number of schools.

Part (3) of the Table shows the values of V^{ML} with no explanatory variables in the model. Part (4) shows V^{ML} when parental education is added to the expectation component of (1), with still a simple decomposition of the random component. (This $\underline{x}_{qi}$ is defined to equal 1 if the pupil's parents stayed on in education beyond the minimum leaving age, and 0 if not.) Comparing parts (3) and (4), we can see that parental education has explained about one third of the occupational-class segregation.

Parts (5) and (6) of the Table show that modelling can replicate and extend *Atkinson*'s index, because it, too, can assess the contribution to overall segregation that came from schools with p_i in particular ranges. Define a new indicator x_i to be 1 if the school is predominantly middle class, and to be 0 otherwise. ("Predominantly middle class" has been defined to be having an average of at least

APPLICATIONS

Table

Description of samples, and multilevel segregation indices

(standard errors in brackets)

		1978	1980	1984	1988
Sample description					
(1) sample sizes	pupils	8846	21946	6364	4913
	schools	436	450	446	440
(2) proportion in designated group		0.83	0.79	0.78	0.71
Segregation					
(3) overall, unadjusted		0.17	0.18	0.14	0.14
		(0.013)	(0.012)	(0.012)	(0.014)
(4) overall, unadjusted for parental education		0.10	0.12	0.084	0.10
		(0.010)	(0.009)	(0.010)	(0.012)
(5) separately for middle-class schools, unadjusted	middle-class schools	0.099	0.066	0.049	0.022
		(0.039)	(0.020)	(0.029)	(0.023)
	other schools	0.063	0.068	0.051	0.071
		(0.008)	(0.006)	(0.008)	(0.011)
(6) separately for middle-class schools, adjusted for parental education	middle-class schools	0.72	0.037	0.011	0.001
		(0.035)	(0.015)	(0.021)	(0.019)
	other schools	0.044	0.047	0.032	0.053
		(0.006)	(0.005)	(0.007)	(0.010)

50% of leavers over the decade with fathers in classes I or II; this gave 72 middle - class schools, containing about 10% of the pupils.) Then, modelling the variances separately in these two types, we get part (5) of the Table. It appears that segregation was falling among the middle-class schools, but not among the others. This conclusion was the same as that which could be drawn from the same data by *Atkinson*'s index with suitable values of the variable parameter (not shown in the Table).

The present approach allows the analysis to be taken further than does *Atkinson*'s index, because we can try to explain the difference between the two groups of school in terms of further explanatory variables. Part (6) of the Table shows the result when the expectation component of the model contains the indicator of parental education as well as the indicator of middle-class schools. We can conclude that for the second half of the 1980s, the segregation among middle-class schools is explicable by the pattern of parental education, but that such an explanation does not work so well for the other schools.

5. Conclusions, gaps, and further work

Thus multilevel modelling offers a unified framework for segregation indices which can allow us to look for statistical explanations of the patterns that the indices show.

Two methodological extensions of this work are worth mentioning. The first is to use logistic regression rather than the linear model (1). In fact, when this was tried for the analogue of part (3) of the Table, using the theory of (*Goldstein 1991*), very similar results were obtained, and so a linear model may be adequate.

The second extension is to look for multilevel-modelling versions of all the conventional indices, just as we have been doing for the Variance Ratio. For example, the dissimilarity index defined by *(James and Taeuber 1985)* can be interpreted as being based on the same principles as the Variance Ratio but using the absolute norm to measure distance instead of the squared norm; a similar comment can be made about the *Gini* index. To develop multilevel versions of these, we would have to develop multilevel software that could handle minimum-absolute-norm regression instead of least-squares regression.

The work reported here is still at an early stage. Some comments which I have had are as follows; more comments and criticisms would be welcome.

1) Does the modelling approach genuinely get at social processes, as opposed to merely more elaborate descriptions? The special force of this point here is that a segregation index is itself a statistical artefact (not a real social fact like, say, an examination pass).

2) For similar reasons, the "process" summarised by equation (1) is not like some other statistical models. In the meaning intended here, the probability of a pupil's belonging to classes I and II has to be interpreted in an epistemological sense, in contrast to, for example, the probability of a pupil's attaining three examination passes, which could be understood as describing a stochastic process. A possible resolution of the two ideas might be available through the argument that probability judgements are analogies of chance set-ups (*Shafer, 1981, pp.16-18*).

3) The approach based on observed proportions, rather than underlying probabilities, could continue to be important from the point of view of social rights: knowing that a social process could have turned out otherwise is no consolation to those who have suffered from a particular random outcome.

4) Are we justified in assuming that the school-level errors are independent among schools? For example, spatial correlation among the school errors could arise if a particular community with several schools had a very low proportion of middle-class families; then the residuals at the school level from model (1) for these schools would all be positive. By including a measure of the socio-economic composition of communities, this phenomenon could be modelled explicitly.

5) The separation of "middle class schools" is no more than a first step, because most of these 72 schools were geographically isolated from each other, being surrounded by other schools which had a much lower proportion of middle-class pupils. Again, an explicit representation of community in the model is needed.

6) It would useful to be able to take account of measurement error in, for example, the definition of "middle class school".

Acknowledgements

The work was supported by the UK Economic and Social Research Council (grant number C00280004). The surveys were supported by the Scottish Education Department and other departments of Scottish and UK central government. I am grateful for comments on the fuller report (*Paterson, 1991*) by *Michael Adler, Frank Bechhofer, Geoff Cohen, Harvey Goldstein, David McCrone, Andrew McPherson, David Salmond, and T.M.F.Smith.*

PROJECT NEWS

References

Adler, M., Petch, A. and Tweedie, J. (1989). *Parental Choice and Educational Policy*. Edinburgh: Edinburgh University Press.

Duncan, O.T. and Duncan, B. (1955). *A methodological analysis of segregation indexes*. American Sociological Review, 20, 210-217.

Goldstein, H. (1991). *Nonlinear multilevel models, with an application to discrete response data*. Biometrika, 78, 45-51.

James, D.R. and Taeuber, K.E. (1985). *Measures of segregation*. In: *Sociological Methodology*, ed N.B.Tuma, 1-31. San Francisco: Jossey-Bass.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Office of Population Censuses and Surveys (1970). *Classification of Occupations*, 1970. London: HMSO.

Office of Population Censuses and Surveys (1980). *Classification of Occupations*, 1980. London: HMSO.

Paterson, L. (1991). *Segregation indices and multilevel modelling. Edinburgh:* Centre for Educational Sociology.

Prosser, R., Rasbash, J. And Goldstein, H. (1990). *ML3 Software for Three-level Analysis*. London: Institute of Education.

Shafer, G. (1981). *Constructive Probability*. Synthese, 48, 1-60.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *General Introduction*. In: *Analysis of Complex Surveys*, eds Skinner, C.J., Holt, D. and Smith, T.M.F., 1-20. New York: Wiley.

Multilevel Workshop in May

The regular multilevel modeling workshop offered by the Multilevel Models Project will take place in the Institute of Education, University of London, on *19-21 May 1992*. As usual, the three-day session emphasises on data analytic practice rather than mathematical statistics. The course contains a mixture of hands-on-sessions using the ***ML3*** program, lectures and group discussions aiming to enable participants to do multilevel analysis and interpret the results. Participants should have a knowledge of ordinary multiple regression, and are strongly encouraged to bring their own dataset for the workshop. There is no charge for attending it. The following subjects will be covered during the course:

Basic 2 level variance component models and those with random coefficients; Analysis of categorical explanatory variables; Residual calculation and analysis; Repeated measures data analysis; Multivariate data analysis; Logit models and use of macros;

Further information and application form are available from *Min Yang* at the project address. Reservation for following workshops can also be taken.

Multilevel Data Library

A data library has been designed for purposes of teaching and training in multilevel modelling by the Multilevel Models Project. Twelve data sets, mostly from real social surveys, are available. Each data set consists of an ASCII file, an ML3 worksheet and a text file describing the data source and coding.

Any one who intends to use or to contribute to the library is welcome to contact us at the address on the front page. No fee is charged for using them.

Workshop In April

One week seminar on multilevel models will be given to a group of Dutch psychometrcians and sociometricians on 6th - 10th April 1992 at the Institute of Education, University of London. It is organized by the Interuniversitair Onderzoeksinstituut voor Psychometrie en Sociometrie (IOPS) in Amsterdam, and funded by the Netherlands Organization of Scientific Research.

Macro Library

The multilevel models project is continuing to expand its library of ***ML3*** macros for specialist analyses. There are now available for logit linear models, survival two models and time series models among others.

APPLICATIONS

A Three Level Growth Model for Educational Achievement: Effects of Holidays, Socio-Economic Background and Absenteeism

Huub van den Bergh & *Kans Kuhlemeier*

Introduction

In general, growth can be modelled as a function of time or measurement occasions. In that case a polynomial is used, for instance, with a linear component, a cubic component etc. until the next term does not contribute to the explanation of the scores. One could do the same in modelling academic achievements by means of a polynomial.

Modelling individual growth data in an educational context comes down to the specification of a three level polynomial model: measurement occasions are nested within students and students are nested within schools. At the first level the individual change in achievement is specified, at the second level the deviation of individuals around their trajectory, and at third level the deviation between schools can be specified. One could extend model in order to estimate the effects of specific contextual characteristics. We would like to mention three: the summer holidays, the socio-economic background of the students and absenteeism of students.

A student does not receive education throughout the year; there are some gaps, usually coinciding with the term holidays. In particular the long summer holidays might slow down the growth in learning. At this point we can follow one of two options. We can specify whether a measurement took place at the beginning of the year, and therefore model a possible summer holiday effect explicitly, or we can fit a more complex polynomial than in the first case and infer a summer holidays effect from the results. We have chosen the first option.

Secondly, schools differ with respect to the socio-economic background of their students. If we wish to compare schools with respect to their achievement or effectiveness, it makes sense to take into account these differences in background. In general, the explanation of the effect of social background is in terms of opportunities to learn at home, parental encouragement etc. For instance, students from a higher socio-economic background have more books at home and therefore more opportunities to read, and they might be more encouraged to read at home etc. As a result these students may have higher scores on a reading test. Following this line of reasoning we might expect that the effect of socio-economic status is higher in the holidays. During the summer holidays much more time is spent at home than during the school year. Hence, especially during the holidays there might be larger differences in time allocation between students from different social economic background. Therefore, we expect a larger effect of social background on achievement scores if the student takes a test at the beginning of the school year then in the middle or at the end of the school year.

Thirdly, modelling growth data implies longitudinal research. One of the major problems in longitudinal research is the absenteeism of respondents at measurement occasions. Especially in an educational context, students who are absent might be absent with reason. For instance, because they go to special classes when the tests are taken.

APPLICATIONS

In this contribution we have modelled scores on tests for mechanical reading, which are administered during the first four years of primary education. 527 students from 24 schools for primary education took parallel versions of a Dutch test called the 'one minute reading test' at six occasions: the end of the first year, the beginning of the second year, the end of the second year, the beginning of the third year, the end of the third year and the beginning of the fourth year. This test in essence comes down to reading aloud as many words as possible from a given list in one minute.

The model was constructed corresponding to the considerations mentioned above. The reading score y_{ijk} of student j ($j = 1, 2, ..., N_k$) at school k ($k = 1, 2, ..., N$) on occasion i ($i = 0, 1, 2, ..., 5$) can be written as:

$$
\begin{aligned}
Y_{ijk} = \quad & \pi_{000} + && \text{(mean score at the end of the first grade)} \\
& \pi_{100} t_{ijk} + && \text{(mean linear change)} \\
& \pi_{200} t_{ijk}^2 + && \text{(mean cubic change)} \\
& \pi_{300} Holiday_{ijk} + && \text{(mean effect of summer holidays)} \\
& \beta_{400} Ses_{jk} + && \text{(general effect of social economic background)} \\
& \beta_{500} Ses_{jk} * Holiday_{ijk} + && \text{(additive effect of ses at the beginning of a year)} \\
& \beta_{600} Absent_{jk} + && \text{(mean effect of absenteeism)} \\
& (e_{ijk} + && \text{(intra individual residual)} \\
& \mu_{0jk} + \mu_{1jk} t_{ijk} + \mu_{2jk} t_{ijk}^2 + \mu_{3jk} Holiday_{ijk} + && \text{(individual residuals)} \\
& \upsilon_{00k} + \upsilon_{10k} t_{ijk} + \upsilon_{20k} Holiday_{ijk}) && \text{(school residuals)}
\end{aligned}
$$

Note that the between school variance as well as the between student variance is a function of measurement occasion t. For instance, the total between school variance at occasion t can be written as:

$\text{var}(\upsilon_0) + 2t * \text{covar}(\upsilon_{0,1}) + t^2 * \text{var}(\upsilon_1) + 2 * Holiday * \text{covar}(\upsilon_{0,3}) + 2t * Holiday * \text{covar}(\upsilon_{1,3}) + Holiday^2 * \text{var}(\upsilon)$.

Results

All effects specified in the model above appeared to be significant (i.e. larger than twice the estimated standard error). Especially the effect of absenteeism proved to be interesting. Firstly, because it is a relative large effect (-.35 on overall standardized scores) and secondly, because the estimation of the effect of absenteeism has a large effect on the other parameters in the model. For instance, if absenteeism is taken into account, the between school variance is relatively small at the end of the summer holidays, and relatively large at the end of the school year (see *Figure 1*). If absenteeism is not modelled, the picture is reversed: the differences between schools are relatively large at the end of the summer holidays, and relatively small after a period of education at the end of the school year. Low achievers tend to be absent at two or more occasions. We have summarized the results in two figures.

In *Figure 1* the students' mean growth in reading mechanics is presented as a solid line. It can easily be seen that the growth in reading mechanics mainly takes place during the school year (i.e. from occasions two to three, and from four to five). The growth levels off during the summer holidays (i.e. from occasions one to two, from three to four, and from five to six). The dotted lines indicate the between student standard deviation (i.e. the mean score at occasion t ± one between student standard

APPLICATIONS

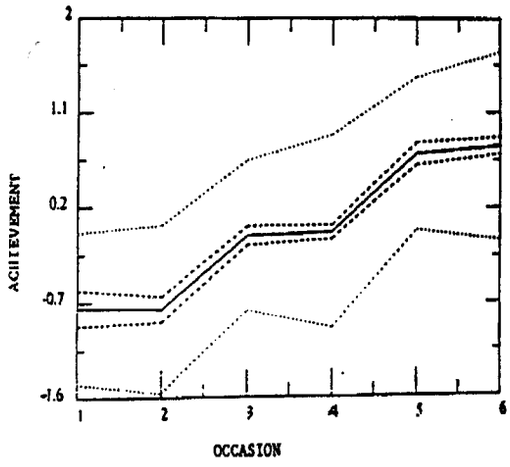

Figure 1: The mean achievement scores (solid line), mean achievement scores ± one between student standard deviation (dotted lines) and the mean achievement scores ± one between school standard deviation (bold dotted lines).

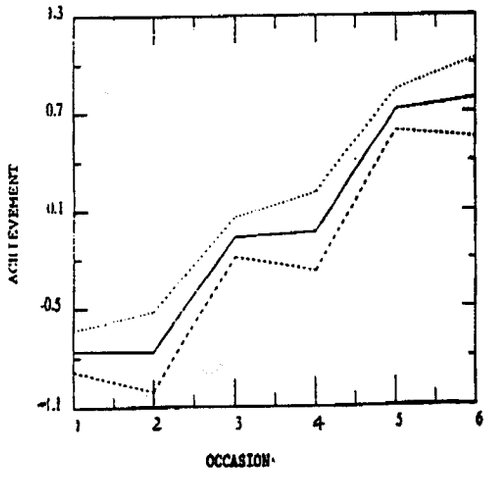

Figure 2: The mean achievement for a student from high social economic background (dotted line), and from low social economic background (dashed line).

deviation). It can be seen that the between student variance increases during the summer holidays (for instance, from occasions three to four) and stays more or less constant during the school year (for instance, from occasions two to three).

In *Figure 1* the between school standard deviations are plotted as the mean student achievement ± one between school standard deviation. It can be seen that the between school variance is relatively large at the end of a school year and small at the beginning of the year. The between school variance seems to behave opposite to the between student variance. In *Figure 2* the effect of socio-economic background is expressed. The solid growth curve represents the growth of a student from an average social background. There appears a general effect of social background, students with a high social background have a somewhat higher reading mechanics score then students with a low social background. Furthermore, there appears to be an additive effect of background after the summer holiday. Hence, there is a relatively large influence from a students' background at the beginning of a year which diminishes during the school year.

Discussion

The growth in reading achievement seems to be related to holiday. The mean growth rate is relatively small during the summer holidays and large during the school year. Education seems to have an effect on achievement in reading. The between student and the between school variance in reading both increase over time. The increase in between student variance is considerably larger then the increase in between school variance. The increase in both type of variance does not seem to be linear, but the student variance increases during the summer holidays, whereas the between school variance decreases during the holidays and increases during the school year. Hence, traditional coefficients like the intraclass correlation (i.e. the between school variance as a function of the total variance) show a peculiar pattern. In general, there is a slow decrease in the intraclass correlation for reading over time, but this coefficient is relatively small at the beginning of the year

and large at the end. Intuitively this seems to make sense. Differences in achievement scores between schools are (at least partly) a result of education, and hence are largest at the end of the school year.

The effect of social background seems to be larger at the beginning of the school year then at the end. Although these effects are in accordance with our hypothesis mentioned in the introduction both effects are a bit hard to interpret. On the one hand our line of reasoning might be right, or we could reason that teachers are doing a relatively good job in counteracting the effects of social background. On the other hand both effects might be an artefact. Students from low social background during the summer holidays actually forget how to do it; there achievement scores are lower at the beginning of a year than at the end of the year before. Hence, they have to relearn, and because relearning is easier and goes faster than learning new skills, students from lower social backgrounds will learn faster during the school year. In terms of the model this is estimated as a lower general effect of social background, and the height of this parameter is an artefact of the effects of forgetting and relearning.

We do not know to what extent we can generalize the conclusions mentioned above to other subject areas. The data analyzed here only concern 24 schools and one subject area, although we do have similar results for spelling.

* *

New Dates for Clinics

The monthly clinic on multilevel data analysis, free for users of *ML3*, is continuing. It takes place in Room 683 at the institute from 10:00 am to 5:00 pm. Future dates are as follows:

February 11 1992   **March 10 1992**

May 12 1992   **June 9 1992**

Persons wishing to participate please call *Min Yang* on *071 612 6682* first to make an appointment.

ESRC Seminar Series

The Economic and Social Research Council has founded a seminar series on Longitudinal and Multilevel Data Analysis will be held from 11 am - 5 pm on Fridays between March 1992 and November 1993. It aims to provide an introduction to methods of analysing longitudinal and multilevel data. Topics of seminars are as follows. There is no charge for participants.

Introduction to Longitudinal Data: Design and Analysis

Introduction to Multilevel Modelling

Longitudinal and Multilevel Analysis

Generalized Linear Models for Longitudinal Data

Multilevel Modelling of Educational Data

Event History and Survival Analysis

For joining the mauling list and receiving details of each seminar, please contact LAMDA Seminars at the following address:

SSRU, City University, Northampton Square, London EC1V 0HB