MULTILEVEL LINEAR MODELS FOR THE EDUCATIONAL SYSTEM

MURRAY AITKIN

1. INTRODUCTION

During the last eight years, interest in assessing the comparative performance of teachers, schools and educational policies has greatly increased in the USA, the UK and other countries. Particularly in the context of declining budgets for education, a serious concern for accountability and effectiveness of the school system has led to attempts to rate teachers, schools, education authorities and even states using simple summaries of administrative data and aggregate scores on standardized tests. These attempts at ratings have aroused anxiety, anger and even legal challenges. Some of the simple methods of comparison or rating used in the past have been seriously defective methodologically, and unwarranted conclusions have been drawn from them. In this paper I describe a model for a hierarchically structured educational system, the analysis of which provides a methodologically sound description of differences among teachers, schools, authorities and states. The implications of this model for educational policy issues are then discussed, and two general conclusions are drawn:

i) in general, it is impossible to assess in a methodologically sound way the effect of a change or "intervention" in educational policy on outcome variables from this model, unless the change is embedded in some form of randomized experiment;

ii) the assessment of such effects is necessary over an extended period of time, requiring longitudinal analysis of data from the model.

2. HIERARCHICAL MODEL FOR AN EDUCATIONAL SYSTEM

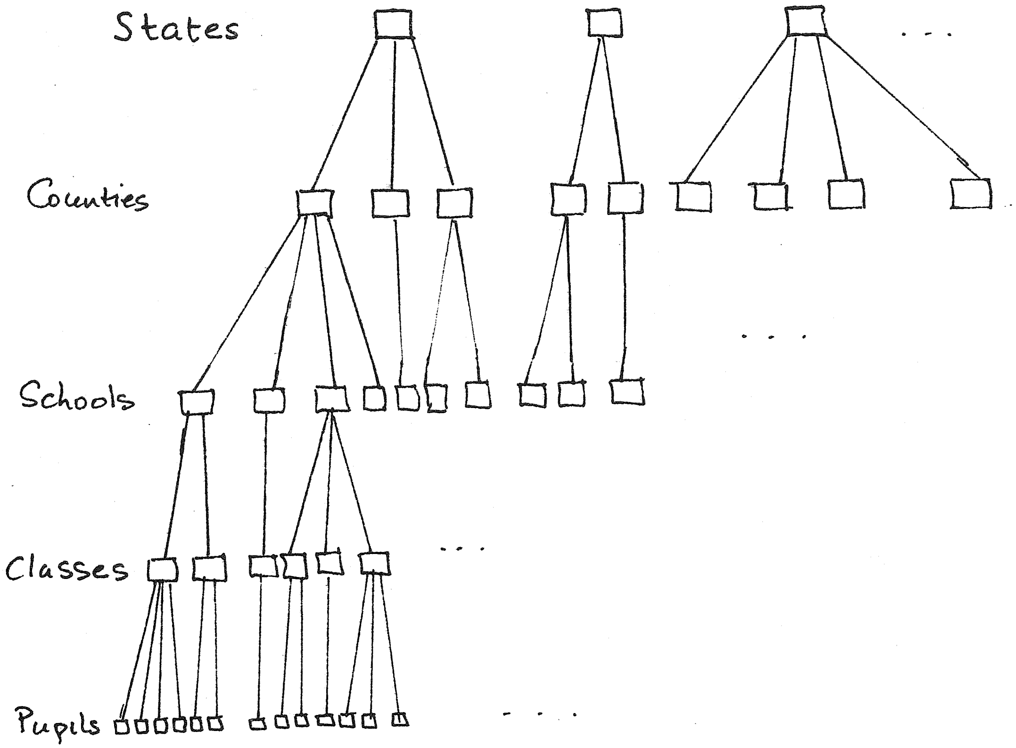
An educational system is not homogeneous: there are distinct differences in educational policies, teacher training, curricula, per-student spending, parents' attitude and many other variables across states, educational authorities, schools and classes. A natural representation of the system is a *hierarchy*, with states at the highest level, education authorities or counties grouped or *nested* within states, schools nested within education authorities, classes nested within schools, and students nested within classes. At each level of the hierarchy, variables are measured on each "unit", and each unit at the next level down within this unit shares the same values of these variables. Thus each county within a state shares the same values of the state variables, each school within a county shares the same values of the county variables, and so on. Counties within the same state are thus more homogeneous than counties in different states, schools within the same county are more homogeneous than schools in different counties, and so on.

The variables measured at the state and county level represent state and county educational policy and budget "input" variables, and those at the school level represent school policies, quality of physical plant, socioeconomic indicators for the school district and so on. Variables may be measured at the teacher level, particularly in studies of elementary school processes; such variables might be age, qualifications, experience, attitude and actual classroom behaviour and teaching philosophy. At the student level both input and outcome variables might be measured: ability, socioeconomic status of the family, ethnic origin, number of siblings, attitude to school, days absent, number of disciplinary events, and achievement test results overall or in specific curriculum areas. At each level above the student level, we may also use *aggregate* variables, averages or other summaries of lower level variables. At the class level, we can have class mean student ability scores; at the school level, mean or median days absent from school for all students; at the county level, mean and standard deviation of the number of years experience of all teachers in the county's schools, and so on.

The hierarchical structure is conveniently represented diagrammatically, as shown in Figure 1.

In assessing the relative educational performance of counties, schools, classes and students, we need a statistical model which relates appropriate outcome or performance variables to input variables at each level of the hierarchy, and which recognizes the varying homogeneity of the population among and within the population units at each level of the hierarchy. The first requirement is met by *regression* or *linear models*, and the second by *variance component, multilevel* or *random effect* models. The two requirements are met jointly by *multilevel linear models*, which are now in widespread use following the development of efficient software for their analysis.

Figure 1



3. MULTILEVEL LINEAR MODELS

Multilevel linear models for hierarchically structured educational systems were discussed in a British context by AITKIN, ANDERSON and HINDE [1], AITKIN, BENNETT and HESKETH [2] and AITKIN and LONGFORD [3]. A short booklength treatment of these models is given by GOLDSTEIN [4]. A recent conference proceedings volume (WILLMS and RAUDENBUSH [5]) gives an extensive discussion of applications of the model in many educational contexts, and references to software.

The essential feature of these models is the representation of *random* *variation* at each level of the hierarchy. For example, for a three-level model of county/school/pupil results on an achievement test, the simplest possible model represents the achievement test score Y_{csp} of pupil p in school s of county c by

$$Y_{csp} = \mu + a_c + b_{cs} + e_{csp},$$

where μ is an overall (population) mean score for all pupils, a_c is a "county deviation", the deviation of the mean score for county c from the population mean, b_{cs} is a "school deviation", the deviation of the mean score for school s in county c from the mean for this county, and e_{csp} is a "pupil deviation", the deviation of the score for pupil p in school s in county c from the mean for this school.

In conventional regression or linear models, the pupil-level deviations e_{csp} would be modelled as independent and normally distributed $N(0, \sigma_e^2)$. In the multilevel model, we model in addition the b_{cs} as $N(0, \sigma_s^2)$ and the a_c as $N(0, \sigma_c^2)$, with additional assumptions of independence of a_c,b_{cs} and e_{csp}. The variance of a score Y_{csp} is then $\sigma_Y^2 = \sigma_c^2 + \sigma_s^2 + \sigma_e^2$; the variances σ_c^2, σ_s^2 and σ_e^2 are called *variance components*. A consequence of these model assumptions is that the achievement scores Y_{csp} are in general correlated, rather than independent. Scores Y_{csp} and $Y_{csp'}$ of two pupils in the same school are correlated

$$\rho_s = \frac{\sigma_c^2 + \sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_e^2},$$

and scores Y_{csp} and $Y_{cs'p'}$ of two pupils in different schools in the same county are correlated

$$\rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_s^2 + \sigma_e^2}.$$

Scores Y_{csp} and $Y_{c's'p'}$ of two pupils in different counties are independent. Thus this model represents analytically the varying homogeneity which we observe in practice in educational systems: pupils in closely related "units" (same school) are more homogeneous (have higher correlation between their results) than those in more distantly related units (different schools in the same county).

This model is specified by the values of the four parameters $\mu, \sigma_c^2, \sigma_s^2, \sigma_e^2$. From the available data for a random sample or a complete enumeration of the population, we can estimate efficiently (by maximum likelihood, ML, or restricted maximum likelihood, REML) the values of the parameters, giving estimates $\hat{\mu}, \hat{\sigma}_c^2, \hat{\sigma}_s^2, \hat{\sigma}_e^2$. Several packages are available for this purpose; the VARCL program of LONGFORD [6] is particularly useful as it can handle both large regression models with many variables (up to 300 in one implementation) and also large data sets (with no constraint on the number of lower-level units) with up to three levels in the hierarchy.

From these estimates, we can determine the relative (estimated) variability of test scores among counties, among schools within counties and among pupils within schools. If, for example, $\hat{\sigma}_c^2 = 0$ then counties are homogeneous, and all the variability in test score outcomes can be assigned to schools and pupils. If $\hat{\sigma}_c^2$ is large but $\hat{\sigma}_s^2 = 0$ then counties differ considerably but schools within counties are homogeneous. In general, pupil (individual) variability $\hat{\sigma}_e^2$ will be a large proportion of the total variability $\hat{\sigma}_c^2 + \hat{\sigma}_s^2 + \hat{\sigma}_e^2$.

This simplest possible "null" model will not in general be of much interest, except to provide estimates of the variance parameters, because it contains no policy or other "explanatory" variables related to outcome. In general we will have a set (vector) of county level variables x_c, a set of school level variables x_s and a set of pupil level variables x_p which may be related to the test score or other outcome variable, and the aim of the analysis is to identify the important variables in each set. Recall that x_c may contain county level aggregate or composite variables from x_s and x_p, and x_s may contain school level aggregate or composite variables from x_p. The variables in each set may themselves be powers or products (interactions) of other variables in the set, or products of variables in the set and in sets at higher levels in the hierarchy.

The model is then extended by replacing the simple mean score μ for the p-th pupil in the s-th school in the c-th county by

$$\mu_{csp} = \beta' x_c + \gamma' x_s + \delta' x_p,$$

where β, γ and δ are vectors of regression coefficients for the county, school and pupil variables. Fitting the model gives ML or REML estimates for β, γ and δ and the variance components σ_c^2, σ_s^2 and σ_e^2. A large (and statistically significant) coefficient for a variable shows an association between this variable and the outcome variable, after allowing for variations due to other variables included in the model. Whether particular variables are needed in the model is assessed by standard statistical methods of hypothesis testing, using the likelihood ratio test. Elimination of redundant variables gives a final reduced ("parsimonious") model which contains only those variables necessary to describe the systematic variation present in the data.

A critical issue in correct specification of the model is the careful modelling of variation in mean outcome at the individual level: considerable information is available at this "bottom" level of the hierarchy, and misspecification of the model at this level may substantially bias parameter estimates of variables at higher levels.

A further extension of the model allows the regression coefficient vector for school level variables to be different in different counties, and that for pupil level variables to be different in both different schools *and* different counties. The most general model would have

$$\mu_{csp} = \beta' x_c + (\gamma + g_c)' x_s + (\delta + d_{cs})' x_p$$

where the random coefficient vectors g_c and d_{cs} are themselves modelled as $N(0, \Sigma_c)$ and $N(0, \Sigma_{cs})$. This model is formally equivalent to an interaction model, since for example the term $g_c' x_s$ means that the "effect" of school variables in x_s interacts with county through the multiplier g_c: some school variables have a larger effect in some counties than in others.

4. INTERPRETATIONS OF THE MODEL, AND THEIR LIMITATIONS

We give below a set of questions which can be answered from the model.

i) How much variability is there in outcome score across the levels of the hierarchy?

The estimated variance components $\hat{\sigma}_c^2, \hat{\sigma}_s^2$ and $\hat{\sigma}_e^2$ from the "null" model with constant mean give the breakdown of the total variance $\hat{\sigma}_Y^2$ across the levels.

ii) Which variables at each level are important, in being strongly related to individual student outcomes?

The standardized regression coefficients for the model (i.e. $\hat{\beta}_j / S.E.(\hat{\beta}_j)$) give the importance of the corresponding variables; if a standardized regression coefficient is zero, or close to zero, the corresponding variable does not contribute to the variation in student outcomes beyond the other variables in the model and can be omitted from the model. Interpretation of the model is assisted by *model reduction* procedures: starting from the most complex plausible model, we progressively simplify the model by the omission of variables which do not contribute to the variation in outcome. This produces one or more final *parsimonious* models which retain a "minimal set" of important variables; only these need to be interpreted.

iii) Does the importance of school level variables differ from county to county, or the importance of pupil level variables differ from school to school, or from county to county?

The interaction components in Σ_c and Σ_{cs} provide this information. Zero or very small ML estimates for the variance components (the variances of the slope distributions) show that no important variation in the regression model occurs across schools or counties; a large variance component for one or more slopes shows systematic variation across schools or counties in the importance of the corresponding variables. In practice we do not fit simultaneously large numbers of random slopes in the model, because there is generally little information to support the estimation of multiple random parameters, and the convergence of the ML algorithm may become extremely slow. If a large slope variance component is found for a variable, we would usually look for interactions of this variable with other variables at the higher level to explain this slope variation.

iv) Which schools are "doing well", and which are "doing badly"?

This question can be approached in two different ways. On the one hand, if, for example, per-pupil spending in the school and average years of experience of the teacher are important variables, with large positive standardized regression coefficients for outcome, then schools with high values of these variables are "doing well", and those with low values are "doing badly".

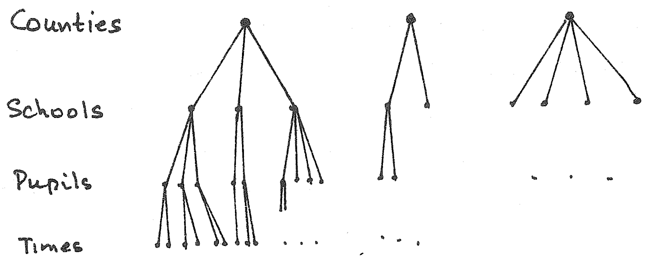
On the other hand we might argue that such schools would be *expected* to do well, given their values of these variables, and the real question is whether the schools are doing *better or worse than expected, given the values of the variables.* If the (residual) school variance component is still large after fitting these explanatory variables, then there are large variations in outcome across schools which are *not* related to these variables, and we may try to identify these through large positive and negative values of the school "random effects" b_{cs}. For technical reasons we identify these values through their "posterior means"; these give more appropriate and stable estimates than simple school means.

A serious difficulty with the second approach is that we are unable, in a one-off cross-sectional study, to say whether the "unexplained " variation at the school level represents real and stable differences between schools due to relevant but unmeasured variables, or whether they are simply random variation which might be completely different in another study. To assess this question a longitudinal study is necessary. If the estimated school variance component $\hat{\sigma}_s^2$ is zero, then there *is* no unexplained variation at this level, and any attempt to rank schools based on school deviations or school means is meaningless.

v) How are "standards" changing over time?

When assessment or testing is carried out repeatedly (say yearly), an additional hierarchical level of "time" is added to the model. How this level appears in the model and in the analysis depends on the level of the hierarchy at which units are repeatedly assessed.

At one extreme, the same students might be assessed at each time period. In a three-level model of county/school/pupil with two time points, this would give the following hierarchy.



At the bottom level of times, we have one explanatory variable, an identifier for time 1 or time 2. The greater homogeneity of repeated measures within the same pupil is again represented in an extended variance component model:

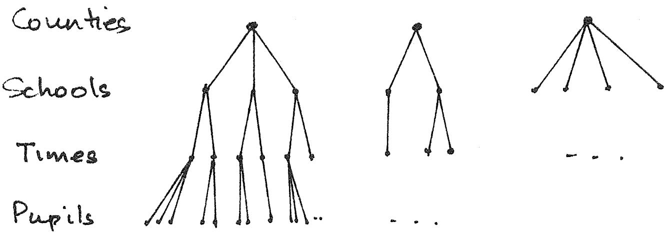
$$Y_{cspt} = \mu_{cspt} + a_c + b_{cs} + e_{csp} + f_{cspt}$$

where the "time deviation" f_{cspt} is $N(0, \sigma_t^2)$, that is the deviation of the score at time t for pupil p in school s in county c from the mean for this pupil. One model for the mean would be

$$\mu_{cspt} = \beta' x_c + \gamma' x_s + \delta' x_p + \theta_t$$

where θ_t is a "fixed effect" parameter representing the overall change of mean score ("standard") with time; the model might also contain interactions of the county, school, or pupil variables with time, representing changes of the importance of these variables with time, and therefore differentially changing "standards" in different counties and schools.

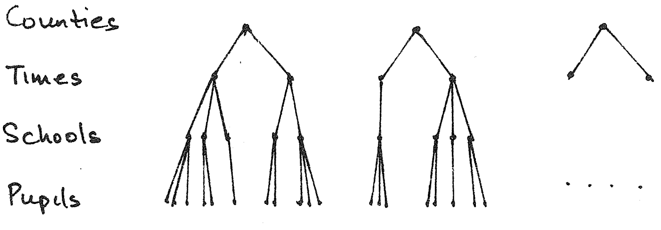
More commonly, repeated testing would be carried out in the same schools, but with different students, for example in successive years. The hierarchy would then be as follows.



Now pupils are nested in times rather than times in pupils, and the model would be

$$Y_{cstp} = \mu_{cstp} + a_c + b_{cs} + f_{cst} + e_{cstp}.$$

In a national longitudinal survey of counties in which different schools were sampled in each wave of the survey, the hierarchy would be that shown below.



The model would now be

$$Y_{ctsp} = \mu_{ctsp} + a_c + f_{ct} + b_{cts} + e_{ctsp}.$$

In each case, time appears as an explicit variable in the μ-model, with its possible interactions with higher-level variables, and the formal analysis of the model in each case is the same, though the estimated variance components for time may be very different since they refer to time changes at different levels of the hierarchy.

In addition to assessing the direct changes over time in mean outcome, we can assess the stability of county or school deviations over time, by examining the need for (random) interactions of these deviations with the time indicator, that is for random slopes of the time indicator over schools or counties. Large estimated variance components for time-by-school or time-by-county interactions mean that the random school or county deviations are "unstable", and changing from time to time. Attempts to rank schools or counties on the basis of their posterior mean random effects will have little point or value if these posterior means change substantially from year to year.

5. CAUSAL INFERENCE AND POLICY EVALUATION FROM MULTILEVEL MODELS

Policy makers at all levels are concerned to evaluate the effects of educational policy changes on outcomes. Models, including multilevel models, have an important role in the analysis of the effects of policy changes, but they are not a substitute for the proper design of studies to assess these changes. Model-based inference about policy changes has been applied in the following design circumstances, with decreasing validity from i) to iii).

i) The policy change is evaluated in a *randomized experiment*, in which a randomly selected experimental group of areas is assigned the new policy, and a control group of areas is assigned the present or an alternative policy; randomization and careful design of the study allow unequivocal analysis of the data and assessment of the effect of the new policy compared with the present one. Such effects may become evident only over a period of several years, requiring longitudinal studies for their evaluation.

ii) The policy change is carried out in some areas but not in others, with no randomization in the choice of areas, and a comparison of new and present policy areas is made, with an attempt to "adjust" the difference in outcome for other variables which differ systematically between new and present policy areas and which might themselves have caused any change in mean outcome.

Since no randomization of policies to areas is used, we cannot be certain that any covariance adjustment of the outcome difference by regression on other variables will remove all differences between areas. Careful matching of areas on relevant variables can help, but the inference that the policy difference caused the outcome difference is always suspect.

iii) The new policy is already in use in some areas, and the present policy in other areas. The fitted model has a large coefficient for the (dummy) policy variable, after regression on all other variables thought relevant. The coefficient is often interpreted as the change in mean outcome that *would be* produced if the present policy *were to be* changed to the new policy in those areas using the present policy.

Inferences of this kind are completely speculative since no actual change of policy has occurred in any area, and there is no way of knowing whether the same model, with the same coefficient values, would apply in a randomized experiment.

REFERENCES

[1] AITKIN, M., ANDERSON, D.A. and HINDE, J.P. (1981). Statistical modelling of data on teaching styles (with Discussion). *J. Roy. Statist. Soc. A* **144**, 419-461.

[2] AITKIN. M., BENNETT, N.S. and HESKETH, J. (1981). Teaching styles and pupil progress: a reanalysis. *Brit. J. Educ. Psych.* **51**, 170-186.

[3] AITKIN, M. and LONGFORD, N.T. (1986). Statistical modelling issues in school effectiveness studies. *J. Roy. Statist. Soc. A* **149**, 1-43.

[4] GOLDSTEIN, H. (1987) *Multilevel Models in Educational and Social Research.* Griffin: Bucks.

[5] WILLMS, D. and RAUDENBUSH, S. (1991). *Pupils, Classrooms and Schools : International Studies of Schooling from a Multilevel Perspective.* San Diego: Academic Press.

[6] LONGFORD, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817-827.

DISCUSSION

In the presentation I described an analysis of the 1984 IEA Science Study in Israel to illustrate the importance of careful modelling of the individual-level model and the effect on interpretations of ignoring interactions between pupil-level variables. The 1984 survey is being replicated in the same schools in 1992 with the support of the Israel Research Foundation (the Ford Foundation in Israel), and a longitudinal analysis of the two surveys will allow the separate identification of class- and school-level variance components, which are not identifiable from the 1984 survey because only one class was sampled from each school. Support for further waves of the survey is being sought, to establish and validate a simple indicator system based on the multilevel model.

In the discussion questions were raised about the relative merits of ML and REML, the latter being thought theoretically preferable. In my response I noted that ML is computationally easier to implement than REML and for the large-scale studies typical in educational research the differences in regression parameter estimates from the two approaches were generally negligible, though differences in variance component estimates could be larger.

Department of Statistics
School of Mathematical Sciences
Raymond and Beverley Sackler Faculty of Exact Sciences
Tel Aviv University