The Statistician (1991) 40, pp. 235–244

Multilevel modelling of survey data*

HARVEY GOLDSTEIN

Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL, UK

1 Introduction

Real populations have hierarchical structures. Thus, student populations are grouped within schools or other institutions which themselves may be further grouped, for example within education authorities. Offspring are grouped within families and families or households are grouped geographically.

Procedures for the analysis of data from complex sample surveys recognise such groupings or clusterings and a large literature exists which shows how to obtain valid inferences from such data. Most of this literature (see, for example, Kish, 1965) is concerned with the statistical properties of simple summary statistics such as means and proportions, although there has also been interest in drawing inferences using more complex models such as regression (Kish & Frankel, 1974). The need to acknowledge the data structure stems from a recognition that the clustering generally induces non-independence between population units so that statistical models based upon independence assumptions become invalid. What matters here is the population structure rather than the sample structure. For example, if we were to draw a simple random sample from a clustered population the non independence of units would remain and affect any inferences based upon statistical models.

A major potential advantage of a model-based approach to survey data analysis is that the hierarchical population structure itself can be modelled and is often of intrinsic interest. Likewise, stratification factors can be incorporated directly into a model as group effects. Previous attempts to apply models such as regression to clustered data have calculated the usual coefficient estimates, for example using ordinary least squares (OLS), and then computed unbiased estimates of standard errors, etc. under appropriate assumptions about the clustering structure. This is the approach taken by Kish & Frankel (1974) and Fuller (1984).

There are two principal drawbacks to such an approach. First, it is inefficient because OLS parameter estimates are less efficient than generalised least squares (GLS) estimates based upon the true structure of the residual covariance matrix. Secondly, this approach does not allow us to explore the clustering structure itself. More recently (Skinner *et al.*, 1990), new kinds of models have been employed to overcome these drawbacks and in the next section I shall develop such an approach based upon multilevel modelling. Following this I will give some examples to show how new kinds of insights can be obtained using these models.

2. The basic multilevel model

To introduce a simple multilevel model I shall use a data set which consists of public examination results for some 31000 16 year olds in what was formerly the Inner London Education Authority (ILEA).

The data were collected over three years—1985, 1986 and 1987—and to start with I consider one year's data. Prior to entry to secondary school, each child in the ILEA was assigned to one of three academic achievement bands, largely on the basis of a verbal reasoning (VR) test. Band 1 contains the highest 25%, band 2 the next 50% and band 3 the bottom 25%. In addition, each child was assigned to an ethnic group by their secondary school. School level variables such as the school denomination, gender composition, etc. are available. A full description of the data is given by Nuttall *et al.* (1989). The response variable is a score derived from 'O' level and CSE grades whereby an 'O' level A grade is given a score of 7, a B grade a score of 6, etc. These scores are then summed over all exams for each student.

In the basic model the exam score is regressed on explanatory variables such as ethnic group and the VR test (using two dummy variables) so as to make adjustment for selective intake differences between schools. We can write such a model as

$$y_{ij}=\alpha+\Sigma_m\beta_m x_{mij}+\sum_{l=1}^{l=2}\gamma_l z_{lij}+u_j+e_{ij} \tag{1}$$

Here x refers to ethnic group and z to VR band. We assume that each school has an 'effect' u_j and e_{ij} is the residual for the ith student in the jth school.

We could view equation (1) as a standard linear model in which a dummy variable is defined for each school and estimates obtained for the u_j. In some situations this may be appropriate, for example where there are only a few schools and we are interested in those schools only. More generally, however, we would choose to regard the sample schools as representative of the population of all schools and make inferences about that population. This immediately leads us to regard the u_j as random rather than fixed variables. In particular we suppose that they have a common distribution, most conveniently a normal one characterised by a variance, σ_u^2 and with a zero mean. In general this variance is unknown and one of the aims of the analysis is to estimate it.

With this assumption equation (1) is no longer a standard regression or any ordinary generalised linear model since it contains two random variables rather than a single random 'residual'. Efficient estimates, for example maximum likelihood or generalised least squares ones, can be obtained by using one of a number of recently developed algorithms and corresponding software packages. The present analyses are carried out using a software system for three-level analysis, ML3, developed as part of a research project supported by the Economic and Social Research Council at the Institute of Education (Prosser *et al.*, 1990). A general introduction to specifying and analysing such multilevel models is given by Goldstein (1987). The term 'multilevel' refers to the random variables in the model which are defined as varying between units at different levels of the hierarchy. Thus equation (1) is a two-level model because there is random variation between students (level 1) and between schools (level 2). To illustrate this further consider the following cases.

3 Levels of variation

Figure 1 is a representation of a simple relationship between a continuous test score obtained at intake when children enter a school and an output exam score at some appropriate end point. The straight line represents the average relationship and would be used to predict the output score for a given input score. The term 'level 1 variation' refers to variation in children's output scores. The educational system forms a hierarchy of levels where children (level 1) are grouped within schools (level 2) which are grouped with LEA's (level 3) etc.

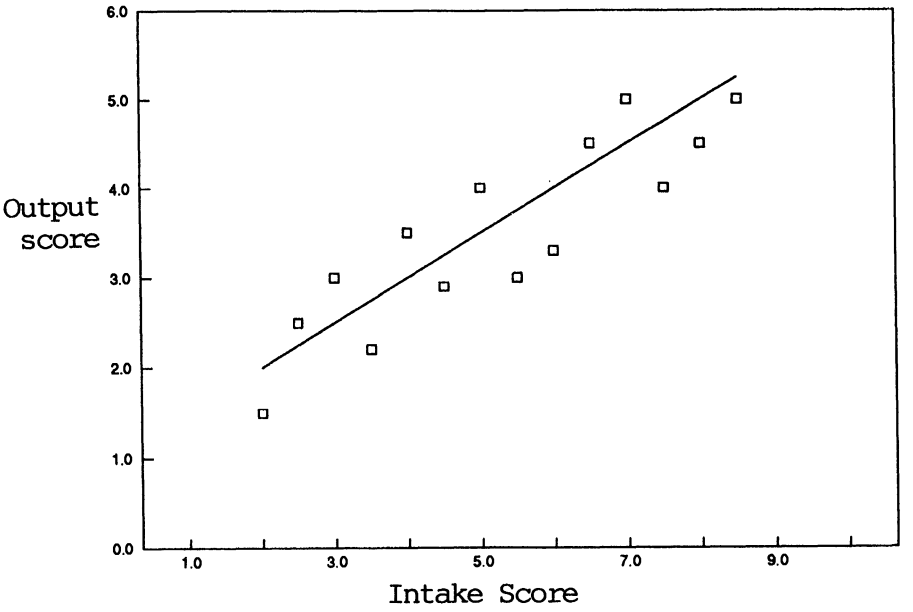

Fig. 1. Level 1 variation.

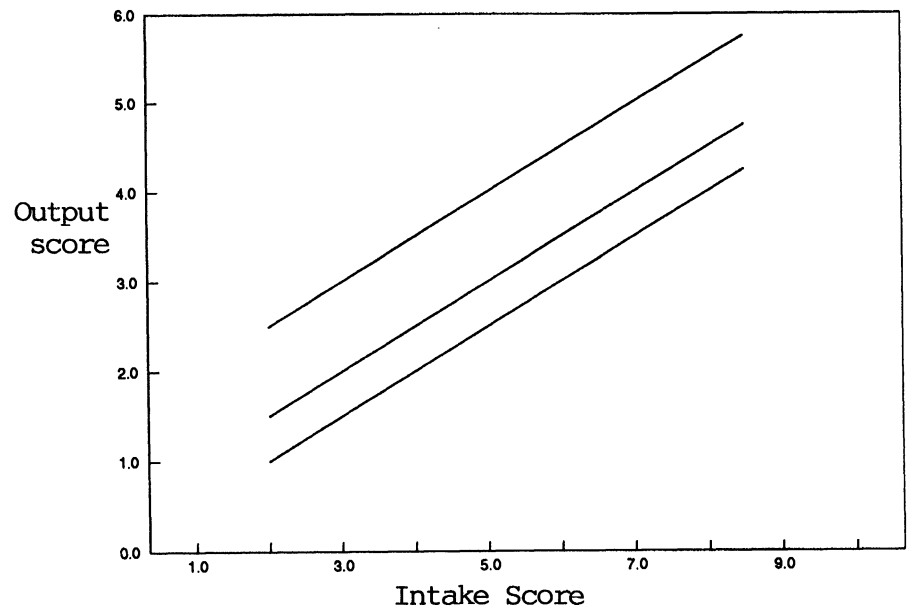

Fig. 2. Level 2 variation.

In Fig. 2 the individual points representing students have been omitted and just the average relationship retained. This is given for each of three schools. The inference which can be drawn from this diagram is that for any given intake score the predicted output scores differ between the schools (level 2 variation). It is in this sense that those researchers concerned with 'school effectiveness' talk of school differences. This 'intercept' variation is assumed to be random and is that described by the u_j in equation (1).

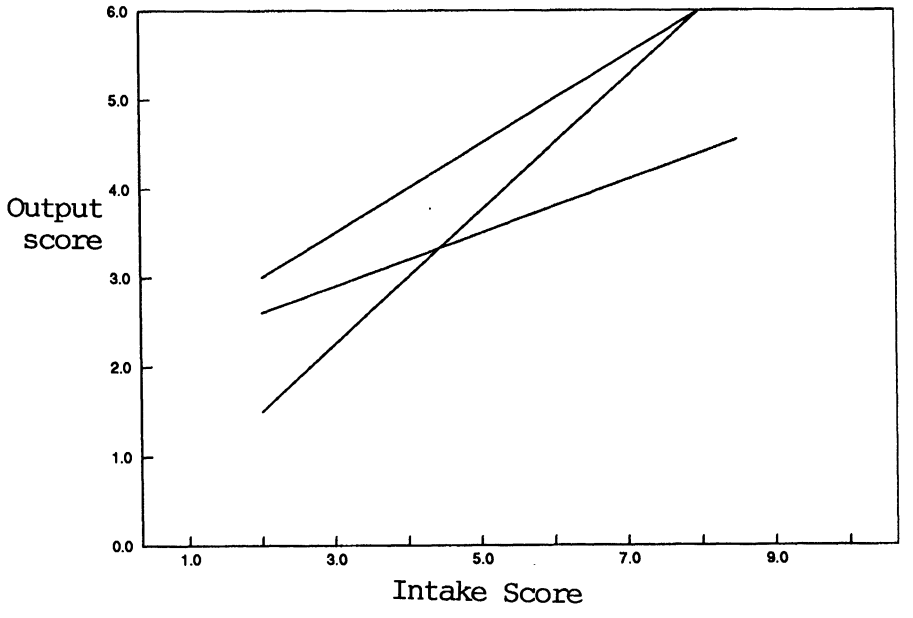

Fig. 3. Complex level 2 variation.

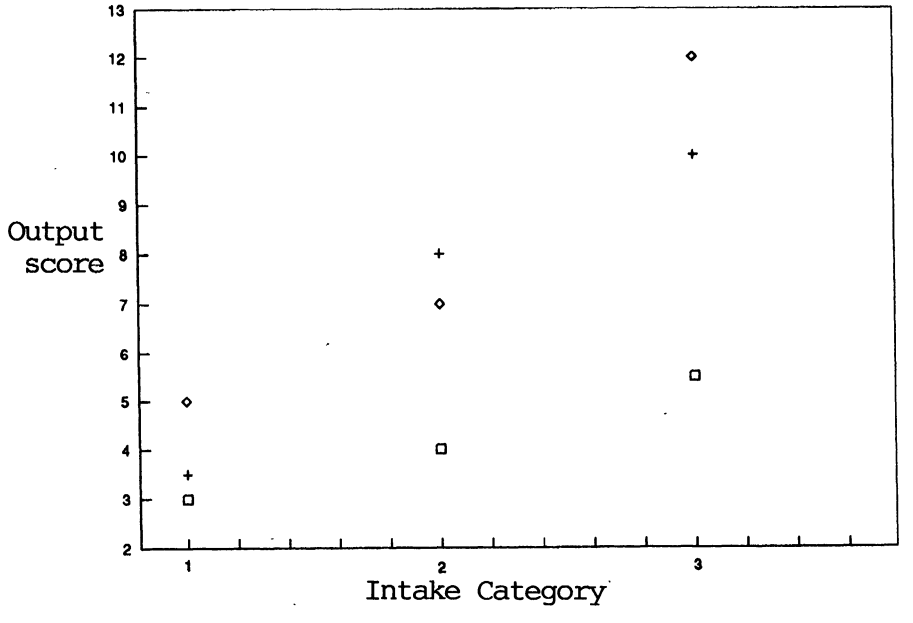

Fig. 4. Complex level 2 variation.

In Fig. 3 the lines are no longer parallel, so that the rank order of the predicted values depends on the actual value of the intake score. This is the pattern which tends to occur in real life and leads to some complexity in describing school differences. In this case the school slopes are assumed to vary randomly in addition to the intercepts.

Figure 4 is similar to Fig. 3 except that the students are now grouped into three ordered categories on intake, for example VR bands, rather than using a continuously distributed

test score. Again, the rank ordering of the three schools depends on the intake category. In the analysis to be described in the next section we will allow school differences to vary by such intake categories.

4 A three-level analysis

Returning to the full data set covering three year cohorts of students we can define a three-level model as follows. Level 1 is that of the student, level 2 is that of cohort or year and level 3 is that of school. Thus the total variation in exam scores can be decomposed into that between students within each cohort, that between cohorts within schools and that between schools. Our new model can therefore be written as

$$y_{ijk} = \alpha + \Sigma_m \beta_m x_{mijk} + \sum_{l=1}^{l=2} \gamma_l z_{lijk} + v_k + u_{jk} + e_{ijk} \qquad (2)$$

In this model the subscript i refers to the student, as before, the subscript j refers to the year or cohort and the subscript k refers to the school. To complete the specification we define the random variable distributions

$$v_k \approx N(0, \sigma_v^2) \qquad u_{jk} \approx N(0, \sigma_u^2) \qquad e_{ijk} \approx N(0, \sigma_e^2)$$

We also assume that the random variables are independent across levels.

A further elaboration is desirable. According to model (2) the ethnic group effects or differences and those between students from different VR groups are constant, that is the same for every school and for each cohort. In reality, however, we might expect that, say, the average difference in exam scores between those in VR group 1 and VR group 3 varied across schools perhaps because of curriculum policies, organisation, etc. To accommodate such possibilities we need to make γ_1 which estimates the VR1 – VR3 difference, random at level 3, so we denote it by γ_{1k}. This leads to two further 'random' parameters, namely the variance of this difference and its covariance with v_k.

4.1 Results

We look first at the 'fixed' coefficient estimates for the elaborated version of model (2) with further fixed explanatory variables and random coefficients added.

When the analysis is carried out a whole series of interesting findings emerge. First, the progress of many of the ethnic minority groups such as Pakistanis, Indians, Greeks, South East Asians and Bangladeshis is high compared with pupils of English, Scottish, Welsh or Irish (ESWI) background. Those from ethnic minority groups are up to a whole 'A' grade better on average than those in the ESWI grouping. Secondly, as would be expected, there is a very large difference associated with the pupils' VR band on entry to school. Those in the top band are, on average, nearly 3 'A' grades ahead of those in the bottom band.

Table 1 shows a selection of the average group differences estimated in the analysis. The 'unadjusted' differences are simply the 'raw' mean group differences before any modelling is done, while the 'adjusted' ones are those when account has been taken of the intake VR band and other explanatory variables in the model. Thus, whereas the Bangladeshi children actually have a lower mean exam score than the ESWI children, after adjustment they are much higher than the ESWI children. In other words they make more progress. It is this measure of progress which is the basis for comparisons between schools.

The most interesting findings, however, relate to the way in which differences vary across schools. Gender differences, ethnic group differences and verbal reasoning band

Table 1. ILEA exam results 1985–1987: three-level analysis

Selected mean differences: (s.e.)	
(*a*) *Unadjusted*	
Girls – Boys	2·5
Bangladesh – ESWI	−0·2
(*b*) *Adjusted*	
Girls – Boys	2·5(0·2)
VR1 band – VR3 band	19·0(0·3)
Bangladesh – ESWI	4·7(0·7)
Roman Catholic-maintained	2·4(0·3)

	Covariance matrix (correlation)		
	Intercept	VR1 – VR3	Carribean – ESWI
Intercept	2·9		
VR1 – VR3	−1·9(0·0)	17·4	
Carribean – ESWI	−0·4(−0·2)	−1·8(−0·4)	1·1

Mean exam score = 19. S.D. = 10.

differences themselves vary from school to school and from a study of these we can gain insights into school differences.

As can be seen from Table 1 the between-school variance of the VR1 – VR3 difference is 17·4, that is a standard deviation of just over 4 points. Given that the mean difference for all students is 18 points, this implies that there are some schools with an estimate of γ_{1k} that is a difference of as much as 27 points and others with a difference as low as about 11 (using ± 2 s.d.). For the kth school the model can be made to provide an estimate $\hat{\gamma}_{1k}$ of its own 'effect' or 'residual' together with an estimated standard error for this residual. These estimates are often known as 'shrunken' estimates (Goldstein, 1987 provides a description and rationale). Thus a school with a large estimated difference might be said to have 'reduced' the initial difference and one with a small difference to have increased it. This raises the interesting speculation that certain school curriculum or staffing policies have a 'homogenising' effect while others have the effect of emphasising initial differences. Since the intercept variance is small, the VR band 3 mean does not vary greatly across schools. This implies that school differences arise from differences in the means for the VR band 1 students.

Table 1 also shows that the gender difference on average is the equivalent of about a CSE grade 3 (2·5 points) in favour of girls, but there are schools where the difference is negligible and other where it is nearly an 'A' grade.

4.2 Implications

By modelling the hierarchical data structure we not only obtain valid inferences for the coefficient means in the model, we also uncover interesting variation among schools in the values of these coefficients. This particular analysis demonstrates that the average effects can provide quite misleading inferences given the large amounts of interschool variation.

The next stage in such an analysis would be to explore reasons for such interschool differences. This might be done by carrying out new studies which measured further factors such as curriculum policy or school organisation, and perhaps the social background of the students. We could then see whether including such factors in the model reduced the between-school variation. Another possibility would be to identify those schools with extremely large or small 'effects' and to carry out case-studies of those

schools to look for possible explanatory factors. Following the demise of the ILEA such studies will be more difficult but other Local Education Authorities have shown interest.

5 Discrete response data

In many surveys the response variable is a count or a proportion rather than a continuous variable, and analogous multilevel models can be formulated. I shall consider the case of proportions as responses where for each level 2 unit or cluster there is a set of level 1 units or categories each of which has an associated response proportion. An example is where the cluster is a school; the level 1 unit is a classification of students by social class, and the response proportion is the number of students with high achievement scores divided by the total number of students in each social class category. A simple model can be written as

$$\pi_{ij} = \alpha + \Sigma_l \beta_l x_{lij} + u_{ij} \tag{3}$$

Here π_{ij} is the mean proportion for the ith category (level 1 unit) for the jth cluster (level 2 unit). The x_{lij} are dummy variables describing the structure of the level 1 units or categories. A specific example is given in the next section. The u_{ij} are random variables at level 2, one for each category, with associated variances and covariances. In practice we would normally wish to summarise these and in the extreme case simply consider a single term u_j for each cluster. The observed proportion can now be modelled, for example as a binomial variable with mean π_{ij} and variance $\pi_{ij}(1-\pi_{ij})/n_{ij}$ where n_{ij} is the number in that category. The aim of the analysis, as before, is to estimate the fixed and random parameters.

The above model will often be satisfactory for values of π_{ij} which are not extreme, but in accordance with standard statistical practice a more realistic model for many data sets is that which uses a logit link, namely

$$\text{logit}(\pi_{ij}) = \alpha + \Sigma_i \beta_l x_{lij} + u_{ij} \tag{4}$$

Full details of how to set up and analyse such models are given in Goldstein (1991), and an example of the ML3 commands is given in Prosser *et al.* (1990).

5.1 A survey of unemployment

This example is concerned with the analysis of youth unemployment data in Scotland. The response variable is the proportion of individuals employed and is categorised by gender and qualification level (unqualified, qualified). A total of 122 geographical areas were sampled and so we have a two-level model with four (2×2) level 1 units per level 2 unit (but often with some level 1 units missing). The number of cells (m_j) per level 2 unit has a maximum value of four.

We write a 'main effects' model, in the exponential form

$$\pi_{ij} = \exp(\alpha + \beta_1 x_{ij1} + \beta_2 x_{ij2} + u_j)\{1 + \exp(\alpha + \beta_1 x_{ij1} + \beta_2 x_{ij2} + u_j)\}^{-1} \quad i = 1, \ldots, 4 \tag{5}$$

where x_{ij1} is a dummy variable for gender, x_{ij2} is a dummy variable for qualification level and π_{ij} is the expected proportion in the ith cell of the (2×2) classification for the jth area. We use a single random 'intercept' term for each cluster u_j.

Table 2 shows the results of fitting equation (5) together with an explanatory variable defined at level 2, the proportion of one-parent families in the area.

Looking first at the fixed coefficients we see that the proportion employed is higher for men and for those with a qualification. Based on the quadratic relationship with the percentage of one-parent families, the highest percentage employment occurs in areas with

Table 2. Proportion employed by gender, qualification and percentage of one-parent families (logit model)

Parameter	Estimate	S.E.
Fixed		
Intercept	1·47	
Gender	0·158	0·11
Qualification	0·968	0·11
Percent 1-parent	−0·252	0·09
(Percent 1-parent)^2	0·012	0·005
Random		
Level 2		
σ_u^2	0·209	0·07
Level 1		
σ_{e1}^2	1·14	0·19
σ_{e2}^2	1·04	0·17
σ_{e3}^2	0·94	0·14
σ_{e4}^2	1·13	0·17

The level 1 parameters are ordered by gender within qualification. Gender is coded 0 = female, 1 = male. Qualification is coded 0 = unqualified, 1 = qualified. Percentage 1-parent is the percentage of one-parent families in the area. This percentage varies from 1·4% to 16·2%. Number of level 1 units (cells) = 401. Number of level 2 units (areas) = 122.

about 10% one-parent families, although it is not clear why areas with both high and low percentages of one-parent families should have relatively low percentages in employment.

The between-area variance of 0·21 is fairly substantial in comparison with the effects of the fixed explanatory variables. Further analysis of these data would look at other explanatory variables, especially those measured at the area level, to see how much of this between-area variability could be explained. As in the examination data we can also estimate an 'effect' for each area and study the characteristics of the areas with extreme estimates.

The assumption of a single common level 2 variance is made in this model, but we could allow any of the fixed coefficients to be random also. In the present case, for example, if we allow the gender coefficient to vary across areas we obtain a positive estimate of the variance, but with a relatively large standard error. It does indicate, however, that the between-area variance for males is larger than that for females. We could make further coefficients random, allowing different variances for the chosen categories.

At level 1, rather than assume binomial variation, a 'scale factor' has been estimated which will be 1·0 when the variation is binomial. The scale factors are obtained by defining dummy level 1 explanatory variables

$$z_{ij} = \{\hat{\pi}_i(1-\hat{\pi}_i)/n_{ij}^{-1}\}^{1/2}$$

which are uncorrelated. The $\hat{\pi}_{ij}$ are the predicted means for the categories, and are reestimated at each iteration. The scale factors are then the estimated level 1 variance estimates of these explanatory variables, as shown in Table 2. As can be seen from the estimates these scale factors are fairly close to unity, although this will not always be so. A particular case would be where there are substantial clustering effects within the designated level 2 units, and departures from unity will indicate that these may be present.

6 Discussion

The advantages of a model based approach to survey data analysis can be realised by the application of multilevel models. These models not only provide more efficient estimates than traditional approaches, they also allow the exploration of variation between clusters which may be of interest in its own right. This is especially so in the analysis of educational data, but will also be so in other fields, for example epidemiology, where the clusters are institutions or operational units. Likewise stratification factors can also be incorporated into the model, as fixed explanatory variables, and their additive and interactive effects studied.

While just two relatively straightforward analyses have been described in this paper, more complex models can be constructed readily. Thus, multivariate responses can be modelled, and for discrete data this allows us to handle multinomial responses as well as binomial responses. It is also worth mentioning that a binary (0, 1) response variable creates no difficulties. This situation is simply the extreme case of modelling proportions in which each level 1 record consists of a single individual.

An important issue with some kinds of data is that of measurement error. In the first example there may be misclassification probabilities associated with assignment to verbal reasoning bands, and if these are substantial they can affect our inferences about both fixed and random parameters. Likewise at level 2 we can have measurement errors where we only have estimates of a level 2 explanatory variable rather than the true values. This would occur, for example, where instead of the known proportion of one-parent families we only had an estimate based upon a sample survey. In such a case we would normally have an estimate also of the measurement error variance and this information allows us to construct a modified analysis. Further work is currently being undertaken on this and procedures for handling measurement errors will be incorporated in future versions of the software.

The software which has been developed for these analyses, ML3, is a complete data analysis package, with general data editing and display facilities together with specially tailored data manipulation and plotting for multilevel data structures. It is designed for 286- or 386-based PC machine, and for the latter is able to utilise extended memory where present so that very large data sets can be analysed. A VAX version is also available.

7 Summarising remarks

Real populations have hierarchical structures. Measurements on population units typically reflect these structures. Recent developments in multilevel modelling can produce analyses which are more efficient and flexible than traditional techniques for survey data analysis. This paper give examples from the area of educational achievement and surveys of employment.

Acknowledgements
My thanks are due to Cathy Garner for supplying the employment data and to Desmond Nuttall and the Inner London Education Authority for the examination data. Bob Prosser, Jon Rasbash and Tim Holt supplied valuable comments and the research was supported by the Economic and Social Research Council.

References

FULLER, W. A. (1984) Least squares and related analyses for complex survey designs, *Survey Methodology*, 10, pp. 97–118.

GOLDSTEIN, H. (1987) *Multilevel Models in Educational and Social Research* (London, Griffin/New York, Oxford University Press).

GOLDSTEIN, H. (1991) Nonlinear multilevel models, with an application to discrete response data, *Biometrika*, 78, pp. 45–52.

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

KISH, L. & FRANKEL, M. R. (1974) Inference from complex samples (with discussion, *Journal of the Royal Statistical Society, Series B*, 36, pp. 1–37.

NUTTALL, D. L., GOLDSTEIN, H., PROSSER, R. & RASBASH, J. (1989) Differential school effectiveness, *International Journal of Educational Research*, 13, pp. 769–776.

PROSSER, R., RASBASH, J. & GOLDSTEIN, H. (1990) ML3: software for 3-level analysis. Users Guide (London, Institute of Education).

SKINNER, C. J., HOLT, D. & SMITH, D. & SMITH, T. M. F. (1990) (Eds) *Analysis of Complex Surveys* (Chichester, Wiley).