MULTILEVEL MODELLING NEWSLETTER

Produced through the Multilevel Models Project:
c/o Math, Statistics & Computing Department
Institute of Education, University of London,
20 Bedford Way, London WC1H 0AL
e-mail: rprosser%uk.ac.ion.ioe
Telephone: 071-636-1500 ext 473

Vol. 2 No. 3 November 1990

NEW MULTILEVEL SOFTWARE OUT NOW

The Multilevel Models Project team is pleased to announce the release of two upgrades for *ML3–Software for Three-level Analysis*. These are Version 2 of *ML3*, and an extended memory implementation called *ML3-E*.

Both products feature a new high-resolution (HR) plotting component that facilitates data exploration and creation of publication-quality scatterplots, line graphs, barcharts, and histograms. Graphs can be printed on PostScript and HP LaserJet laser printers and 9- or 24-pin Epson-compatible dot-matrix printers, and plot information can be stored for future use.

Titles, axis labels, and other text are added with a menu-driven editing facility. This can also be used to display the case numbers of particular points, for example, in identifying outliers. A pair of commands is available for adding horizontal and vertical error bars. A zoom feature enlarges sections of a plot.

The HR plotting commands operate cumulatively, permitting the superimposing of plots. So, for example, a normal curve can be displayed on top of a histogram. (*ML3*'s macro facility can be used to produce a new "procedure" that generates normal curve images.)

A user can examine the distribution of a continuous variable in each of the groups in a sample. A new command provides boxplot-like displays for all level 2 or level 3 units on a single graph. The new PLGL instruction shows Y-on-X plots for all groups simultaneously.

A revised *Users' Guide* explains the operation of these features and contains new information on a multilevel extension of the generalized linear model for proportions. An example of the use of the macro facility to fit such models is provided. The manual may be ordered separately from the program.

ML3-E is designed for users who work with large datasets. It is compiled to run on 386-based microcomputers (under DOS) and uses all the extended memory available. A user with N Mb of RAM, for example, could have a worksheet containing up to $(N - 1.2)/4$ spaces. The extra space facilitates the use of more sophisticated models. A set of two-level models, each with a single outcome, can be combined into a three-level multivariate model, for example. Although a coprocessor is not essential, use of one is highly recommended. *ML3-V* for Digital Equipment Corp. VAX computers (running VMS) is also available for the analysis of large problems.

The *ML3 / ML3-E* package includes a program disk (5.25 in or 3.5 in), a *Users' Guide*, four datasets for examples in the manual, and one year of basic support. The current single copy price of *ML3* V2 is £250 (US$475); *ML3-E* is priced at £300 (US$570). A 40% discount is offered to academic users, and quantity discounts are available.

IN THIS ISSUE

PROJECT NEWS

MONTHLY DATA ANALYSIS CLINIC TO START IN LONDON

Part of the Multilevel Models Project's mandate from its funding source, the Economic and Social Research Council of the U.K., is to disseminate information about multilevel modelling. To that end, the Project is starting a new service as an adjunct to its biannual workshop series—a monthly data analysis clinic. The intended clientele is current users of *ML3: Software for Three-level Analysis*. Persons new to multilevel modelling are encouraged to register for the next three-day workshop. This will be held in the spring of 1991, and the exact dates will be announced in the next issue of the *Newsletter*.

The clinic, to be held at the Institute of Education (the address is on p.1), will provide an opportunity for statisticians, social scientists, and educational researchers to receive individual help with their data analyses from members of the Project team. There will be no charge for the consultation. Participants should bring data in the form of ASCII files or *ML3* worksheets on diskettes (preferably 3.5 inch ones).

Times and Dates

Hours of operation will be 9:30 am to 5:00 pm. The first four dates are as follows:

- December 11
- January 29, 1991
- February 26
- March 19

Persons wishing to participate should telephone the MSC Department secretary (071-636-1500 ext 390) to arrange a time.

WORK BEGUN ON HYPERTEXT HELP

Exploratory work has begun towards development of hypermedia instructional materials on multilevel modelling. The core of a hypertext "book" is a set of linked computer screen "pages" or frames containing bits of text and / or graphics. Each frame can be linked to many other frames permitting a user to follow a variety of paths through the material.

Several tools help in finding desired information quickly. A keyword search facility can assemble a set of frames about a chosen topic, for example, and maps show potential paths. Supplementary details—definitions of terms, say—can be obtained easily by clicking a mouse on a "hotword." Users can place "bookmarks" on frames they want to return to.

Hypermedia applications are elaborate hypertextbooks which could contain animation sequences, points of access to (internal/ external) mathematical and statistical facilities and databases, and/ or an expert system inference engine.

Jon Rasbash is currently testing different hypermedia authoring systems and so far has examined *The Knowledge Engine* by Software Artistry and *ToolBook* by Asymetrix. He is experimenting with Microsoft *Windows 3.0* as a development environment. One difficulty at the moment is generating / screen capturing the equations and graphs needed for explaining multilevel modelling concepts. Another challenge is the linking of the information system with *ML3*.

It is intended that a user of *ML3* should be able to learn about particular types of models while he/ she is analyzing data. The hypermedia component's function would be, in part, to assist in the selection and setup of appropriate models and interpretation of the output. Events during the fitting of a model–parameter estimates dropping to zero, for example–could trigger the display of suggestions about possible causes and remedies. Conversely, a hypertextbook about multilevel modelling could give learners hands-on experience with *ML3* by walking them through stored examples.

A simple module about variance components models is in preparation. The frames provide a brief review of some aspects of OLS regression modelling and introduce ideas such as nested structure in data, multiple random terms associated with the intercept, clustering effects, and partitioning variance. Notation and specialized vocabulary are given links to a glossary, and key features of graphs and equations are assigned pop-up footnotes. Users will be able to add their own notes to the material provided.

Prof. David Hand of the Open University has agreed to serve as a consultant for this component of the Multilevel Models Project. He has a background in statistical computing and experience with statistical expert systems.

THEORY & APPLICATIONS

MULTILEVEL COVARIANCE STRUCTURE WORK

Bengt Muthén

Muthén and Satorra (1989) outlined several possible covariance structure models and their maximum likelihood estimation. However, they did not discuss how such estimation should be carried out in practice nor did they provide examples of such modelling. Muthén (1989) discussed the relationships between multilevel and conventional structural equation modelling and showed that standard software for the latter could be used for the former in the special case of balanced data. Recent related work is that of Goldstein and McDonald (1988), McDonald and Goldstein (1990), and Longford and Muthén (1990). The McDonald and Goldstein paper outlines maximum likelihood estimation of general multilevel structural models with the aim of developing specially designed software to carry out the complex computational tasks. The Longford and Muthén paper focuses on efficient computation for hierarchical factor analysis models. It is interesting to note, however, that a good part of these tasks was already carried out more than 20 years ago in the dissertation of Schmidt (1969)–see also Schmidt and Wisenbaker (1986)–using specially designed software.

While the development of special software is of great value for these situations, Muthén (1990) shows that quite general multilevel models can be fitted by maximum likelihood using relatively minor modifications of already existing structural equation modelling software. A simpler ad hoc estimator which can be carried out by existing multiple-group structural equation software is also proposed. This is a maximum-likelihood-based consistent estimator which gives practically useful approximations to the maximum likelihood χ^2 test of model fit and standard errors of estimates. The author has written a simple FORTRAN program for computing the necessary sample statistics, and the program is available to anyone interested. The implication is that this type of multilevel factor analysis and covariance structure modelling can be done already today by anyone having access to programs such as LISREL, LISCOMP, EQS, etc. which should stimulate the use of this important methodology.

To exemplify the opportunities of multilevel structural equation modelling, Muthén (1990) analyzes data on U.S. eighth grade student mathematics achievement related to teacher-reported opportunity to learn (OTL) measures. To this end the LISCOMP program (Muthén, 1987) and modifications thereof are used. A two-level model is considered for about 800 students in about 180 classrooms. The model involves a single-factor model at both the student and class level. The class-level factor is interpreted as being related to effects of tracking and variations in instruction. Apart from a significant class-level factor variance, variable-specific residual variation across classes was found for algebra and geometery topics. These topics are not sufficiently taught in many eighth grade classes. An expanded multilevel model relates the class-level components of the achievement scores further to the class-level OTL variables through a MIMIC (multiple causes, multiple indicators) structure. The OTL variables are found to have rather little influence on the class-level factor and more direct influence on each achievement variable's class component.

References

Goldstein, H., & McDonald, R. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455-467.

Longford, N., & Muthén, B. (1990). *Factor analysis for clustered observations.* Los Angeles: UCLA.

McDonald, R., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42,* 215-232.

Muthén, B. (1987). *LISCOMP. Analysis of linear structural equations with a comprehensive measurement model: Theoretical integration and user's guide.* Mooresville, IN: Scientific Software.

Muthén, B. (1989). Latent variable modelling heterogeneous populations. *Psychometrika, 54,* 557-585.

Muthén, B. (1990). *Mean and covariance structure analysis of hierarchical data* (UCLA Statistics Series No. 62). Los Angeles: University of California.

cont'd on p. 11

THEORY &

EXAMINING MISSINGNESS IN HIERARCHICALLY STRUCTURED DATA WITH A MULTILEVEL LOGISTIC REGRESSION MODEL

Bob Prosser

The purpose of this note is two-fold:

- to describe one consideration regarding incompleteness in multilevel data—*response mechanisms*; and
- to illustrate one method for examining these mechanisms—multilevel logistic regression (MLR).

MLR models, of course, have many applications; situations involving binary outcome variables are legion. The present example can be adapted easily. For further information related to MLR, see Goldstein (in press).

Response Mechanisms for Hierarchically Structured Data

It is a common occurence in research that desired information is missing for some of the units being studied. Students may be ill the day a test is administered, for example. Information about macro-units in a multilevel investigation may also be only partially complete. Teachers asked to describe classroom practices in a study of student achievement, for instance, may omit details on some topics.

Statistical analysis is more problematic with an incomplete dataset than with a complete one. Failure during statistical analysis to come to terms appropriately with the absence of the measurements results in bias and inefficiency in parameter estimation.

A key consideration in describing data incompleteness is the nature of the process causing the losses. The elements of a matrix of data from population members can be viewed as being governed by a *response mechanism*—a stochastic process determining which sample units provide measurements on a given variable (Little & Rubin, 1987). Suppose that in a population, all people will state their age but only some will disclose their income. If the probability that a sampled person tells his/ her income does not depend on either age or income, the unrecorded incomes are, in Little and Rubin's terminology, *missing completely at random* (MCAR). The people who supply both age and income values are, in effect, a random subsample of the selected sample. If the response probability depends on age (but not income) the data are missing at random (MAR). When the probability of recording income for a person depends on the income, a nonrandom response mechanism is at work.

In many situations, one cannot assume that missing data are missing randomly, but the causes of missingness may sometimes be quite understandable: sample members with values below a particular threshold may be less likely than others to supply information on a given variable, for example. When data are missing nonrandomly, it may be possible in an analysis to incorporate a model of the response mechanism—a distribution function, with unknown parameter values, for an indicator of nonresponse—and improve on the parameter estimates that would be obtained if the mechanism were ignored.

In a study of a hierarchical social system, data may be missing at any or all levels, and it may be reasonable in some instances to assume that response mechanisms at the different levels operate independently. For example, in a study of achievement in which students supply test scores and teachers answer questionnaires concerning classroom practices, there may be no relationship between a teacher's propensity to omit a question and a student's chances of missing a test. Different mechanisms may be at work for different variables at a given level. Further, different groups may have different missingness processes, possibly related to group characteristics.

APPLICATIONS

When missing data are MCAR, simple incomplete data strategies such as listwise deletion and various forms of imputation may be useful under different conditions (Prosser, 1991). It is important to examine whether or not an assumption of MCAR missingness is tenable. One way to do this with hierarchically structured data is with MLR. We will focus here on incompleteness at level 1.

Modelling Missingness Using a Simple MLR Model

We want to determine whether a person's chances of lacking a score on some variable Y are related to other variables that will be used in a multilevel analysis. Suppose P_{ij} is an indicator of whether or not person i ($i = 1, \ldots, n_j$) in group j ($j = 1, \ldots, J$) has a score on Y: P_{ij} takes the value 1 if Y_{ij} is absent and 0 otherwise. Let Π_{ij} be the true propensity of person i in group j to lack a Y score.

Model the missingness indicator as a sum of the true propensity and a random deviation:

$$P_{ij} = \Pi_{ij} + e_{ij} \tag{1}$$

If P_{ij} has a binomial distribution about Π_{ij}, its variance for person i of group j is given by $\Pi_{ij}(1 - \Pi_{ij})/n_{ij}$ with $n_{ij} = 1$. Binomial variation is tied to the assumption that if multiple Y measurements were to be obtained for individual i, the presence/ absence of a value on one occasion would be independent of the outcome on another. Such a response mechanism may be unrealistic. The general approach illustrated here can accommodate "extra-binomial" variation.

A two-level random intercept model for the true propensity (involving, say, two variables from the main model, X_1 and X_2) gives

$$\Pi_{ij} = \frac{\exp(\gamma_0 X_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + u_{0j})}{1 + \exp(\gamma_0 X_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + u_{0j})}$$

(The intercept variable X_0 is a vector of 1s.) If the symbol L_{ij} is used for the *linear predictor* $\gamma_0 X_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij}$, the above expression for Π_{ij} is equivalent to saying that $\text{logit}(\Pi_{ij}) = L_{ij} + u_{0j}$. The u_{0j} contributes an overall elevation or depression specific to group j in the logits of the true propensities, and $\text{Var}(u_{0j})$ is a measure of *between-group* variation of these logits.

A Taylor expansion for Π_{ij} in terms of L_{ij} gives the basis for an estimation approach:

$$\Pi \approx Q(L^*) + Q'(L^*)(L - L^*) \tag{2}$$

where $Q(\ell) = \exp(\ell)/(1 + \exp(\ell))$, $Q'(\ell) = \exp(\ell)/(1 + \exp(\ell))^2$, and L^*_{ij} is a currently available estimate of L_{ij}.

Using a rearrangement of equation (2) gives the model for the units in the jth group:

$$\begin{aligned} P_{1j} &\approx [\gamma_0(X_0 Q'_{1j}) + \gamma_1(X_{11j} Q'_{1j}) + \gamma_2(X_{21j} Q'_{1j}) + \alpha(Q_{1j} - L^*_{1j} Q'_{1j})] + u_{0j} Q'_{1j} \\ &\vdots \\ P_{n_j j} &\approx [\gamma_0(X_0 Q'_{n_j j}) + \gamma_1(X_{1n_j j} Q'_{n_j j}) + \gamma_2(X_{2n_j j} Q'_{n_j j}) + \alpha(Q_{n_j j} - L^*_{n_j j} Q'_{n_j j})] + u_{0j} Q'_{n_j j} \end{aligned} \tag{3}$$

where α, the (fixed) coefficient of the correction term $(Q_{ij} - L^*_{ij} Q'_{ij})$, is constrained to be 1. In practice, the correction term is subtracted from the response instead of being included as an explanatory variable, as shown in the example below. The level 1 variance, $\text{Var}(P_{ij}|\Pi_{ij})$, is established for the estimation procedure

via an explanatory variable W which appears in only the random part: $W_{ij} = (Q_{ij}(1 - Q_{ij})/n_{ij})^{0.5}$, with $n_{ij} = 1$ when the level 1 units are individuals.

The parameters of interest are σ_0^2, i.e., $\mathrm{Var}(u_{0j})$ and γ_0, γ_1 and γ_2. If γ_1 and / or γ_2 differs significantly from 0, you would conclude that missing Y values are not MCAR. The σ_e^2 term in the level 1 variance, $\sigma_e^2(Q_{ij}(1 - Q_{ij})/n_{ij})$, may be constrained to be 1 if exact binomial variation is to be modeled. Allowing this parameter to be freely estimated takes care of the possibility of extra-binomial variation.

Note that the values of W_{ij}, Q_{ij}, and Q'_{ij} depend on the γs. Fitting models of this type with *ML3* involves using the estimates of the γs from one iteration to compute updated values of the explanatory variables for the following iteration. Command macros can be constructed to facilitate this repetitive process.

An Example

The data come from a junior school spelling program implemented by an English local education authority. A variety of spelling scores were obtained from 4774 children in 175 schools. Among the variables available were: (a) SEX; (b) FSM, free school meal eligibility (a binary variable which serves as a very crude indicator of socioeconomic disadvantage); and (c) SPELLING, the mark on a 100 item Schonell spelling test, administered in October of a student's fourth school year. In the file provided, no student lacked a value for SEX, or FSM, but 21% had no SPELLING score.

Two of the simple questions of interest in the main analysis were as follows:

(a) How large is the average SEX difference in SPELLING? and

(b) Do "disadvantaged" students score significantly worse than others?

Random intercept models such as

$$(\mathrm{SPELLING})_{ij} = \beta_{0j} + \beta_1(\mathrm{SEX})_{ij} + \beta_2(\mathrm{FSM})_{ij} + e_{ij} \qquad \text{with} \qquad \beta_{0j} = \beta_0 + v_{0j}$$

were fitted to answer these.

The *ML3* command sequence is shown for fitting the MLR model in (3) to a missingness indicator created for SPELLING. Before the program was invoked, three macros (in files called MISS1.MAC, MISS2.MAC, and MISSCORE.MAC) were written to reduce the amount of command typing during the inter-iteration updating of the response and explanatory variables. Batch mode was off (the default setting).

Command Sequence

```
note:  retrieve worksheet
retr spelling.ws
name
Name          n      min         max
1 SCHOOL      4774    12.000     426.00
2 X0          4774    1.0000     1.0000
3 SEX         4774    0.00000    1.0000
4 FSM         4774    0.00000    1.0000
5 SPELLING    4774   -99.000     88.000
```

MISS2.MAC

```
pick 1 c98 b1
pick 2 c98 b2
pick 3 c98 b3
mult 'sex' b2 c41
mult 'fsm' b3 c42
add  c41 c42 c28
add  c28 b1 c42
obey misscore.mac
endo
```

The estimates of the coefficients of the intercept (X_0), SEX, and FSM are -0.82, -2.03, and 0.34, respectively; the respective standard errors for the latter two effects are 0.096 and 0.11. It appears that both sex and social disadvantage have a significant impact on the individuals' chances of lacking a SPELLING score. (The boys are more likely than the girls to have a missing value.) SPELLING marks are not MCAR.

The metric for the coefficients is logits, so to obtain a particular child's predicted propensity for lacking a SPELLING mark, compute $\exp(-0.82 - 2.03\ \text{SEX} + 0.34\ \text{FSM})/(1 + \exp(-0.82 - 2.03\ \text{SEX} + 0.34\ \text{FSM}))$. The missingness propensity for a disadvantaged boy, for example, is estimated to be 0.38. Note that we could have used continuous explanatory variables instead of or in addition to SEX and FSM.

The estimated value of the between-school variance is 1.00, and its standard error is 0.14. There seems to be significant variation across schools in students' chances of not having a SPELLING mark. The linear predictor in the above formula could be increased/ decreased by two standard deviations to estimate ranges of missingness probabilities for students with different characteristics. School-level explanatory variables could be added to the model in an effort to account for this variance and better understand the missingness mechanism. cont'd on p. 11

Command Sequence cont.

```
note:  create missingness indicator P
note:  -99 is missing value code in c5
chan 0 100 c5 0 c6
chan -99 c6 1 c6
name c6 'p'
note:  prepare to create new explan vars
add 0 'x0' c30
add 0 'sex' c31
add 0 'fsm' c32
name c30 'new-x0' c31 'new-sex'
name c32 'new-fsm'
tidy
21034 spaces left on worksheet
iden 2 'school'
iden 1 'x0'
resp 'p'
expl 'new-x0' 'new-sex' 'new-fsm'
setv 2 'new-x0'
setv 1 'new-x0'
note:  get starting values
star
Iteration number 1 in progress
Iteration number 1 in completed
Convergence not achieved
note:  begin with the new explan vars
obey miss1.mac
star
Iteration number 1 in progress
Iteration number 1 in completed
Convergence not achieved
obey miss2.mac
next
note:  use miss2.mac from now on
:
Iteration number 6 in progress
Iteration number 6 in completed
Convergence achieved
```

MISS1.MAC

```
note:  store fixed param ests
pick 1 c98 b1
pick 2 c98 b2
pick 3 c98 b3
pred c42
obey misscore.mac
name c40 'new-p' c41 'w'
note:  shift to modified P
resp c40
note:  use W to define level 1
note:    variance
note:  W removed from fixed part
note:    of the model
expl c41
fpar c41
clrv 1
setv 1 c41
endo
```

MISSCORE.MAC

```
expo c42 c28
add  c28 1 c29
calc c30 = 1 / c29
mult c28 c30 c39
mult c30 c39 c30
note:  c39 is exp(L)/(1+exp(L))
note:  c30 is exp(L)/(1+exp(L))**2
note:  modify explanatory variables
mult c30 'sex' c31
mult c30 'fsm' c32
calc c42 = b1*c30 + b2*c31 + b3*c32
note:  modify response variable
calc c40 = 'p' - c39 + c42
note:  produce W
subt c39 1 c41
mult c39 c41 c41
sqrt c41 c41
```

THEORY &

GRAFTED POLYNOMIALS USEFUL FOR CROSS-SECTIONAL GROWTH ANALYSIS

Huiqi Pan, Harvey Goldstein, & Bob Prosser

Summarization of longitudinal height and/ or weight data using polynomial growth curves is a common application of multilevel modelling (Goldstein, 1986, 1987). In a two-level analysis, individuals are treated as level 2 units, and measurements-within-individuals as the lower level units. If individuals are clustered within, say, regions of a country, a third level can be added to the growth model. Suppose, however, that one has data that are clustered but cross-sectional: each person contributes only one value, but a range of ages is represented in each region. When the range of ages is large, a new model—the *grafted* (or piecewise) *polynomial model*—holds some promise for efficient summarization of such measurements. This note describes the basic model and shows how to fit it using *ML3–Software for Three-level Analysis*. Further information is given in Pan and Goldstein (1990).

The Two-level Grafted Polynomial Model

Let child i in cluster j be measured on response variable Y at age T. For cluster j we write the grafted polynomial of degree p with $m-1$ join points as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \beta_{2j}T_{ij}^2 + \ldots + \beta_{pj}T_{ij}^p + \beta_{p+1,j}(T_{ij} - \zeta_1)_+^p + \ldots + \beta_{p+m-1,j}(T_{ij} - \zeta_{m-1})_+^p + e_{ij} \qquad (1)$$

where ζ_k $(k = 1, \ldots, m-1)$ are join points and $\zeta_1 < \zeta_2 < \ldots < \zeta_{m-1}$. The + subscripts on the terms involving the join points are meant to indicate that when $T_{ij} < \zeta_k$ the corresponding term is set to 0.

In this model, the e_{ij} is the level 1 residual for the ith child in the jth cluster. The coefficient of the intercept β_{0j} and the slopes (β_{1j} etc.) can vary across clusters, and these coefficients are treated as random variables at level 2. Thus each cluster has its own set of such coefficients. We assume

$$\mathrm{E}(e_{ij}) = 0, \qquad \mathrm{Cov}(e_{ij}, e_{i'j}) = 0, \quad \text{and} \quad \mathrm{E}(\beta_{kj}) = 0$$

in each group, for all i and i', and for $k = 1, \ldots, p+m-1$.

We may attempt to account for between-cluster variation in terms of one or more features, Z, of the children or clusters being studied—for example, geographical features of regions or socioeconomic characteristics of the child's family.

An Example

The data used to illustrate the fitting of a grafted polynomial model consist of children's weight measurements recorded at ages between birth and 72 months. These were obtained from 4292 girls and 4679 boys in six districts of Shanghai and five provinces in the middle east and southeast of China. The level 2 units in this analysis were 18 subdistricts, classified as rural or urban.

Let Y_{ij} denote the weight in kilograms of the ith child in the jth subdistrict and T_{ij} denote his/ her age in months at the time of measurement. Let Z_1, Z_2, and Z_3 represent region, urbanization, and sex, respectively. (Region was coded 1 for middle east and 0 for southeast, and Z_2 was coded 1 for urban and 0 for rural. Boys were coded 1 on Z_3, and girls were coded 0.) The following model was fitted:

$$Y_{ij} = \beta_{0j}X_{0ij} + \beta_{1j}T_{ij} + \beta_{2j}T_{ij}^2 + \beta_{3j}T_{ij}^3 + \beta_{4j}(T_{ij} - 12)_+^3 + \gamma_{50}Z_{3ij}T_{ij} + \gamma_{60}Z_{3ij}T_{ij}^2 + e_{ij} \qquad (2)$$

APPLICATIONS

where X_0 is a vector of 1s. We allowed an interaction between sex and the linear and quadratic age components by adding the terms $Z_{3ij}T_{ij}$ and $Z_{3ij}T_{ij}^2$. We supposed that the growth pattern varied across subdistricts, and the coefficients of age were structured as follows:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + u_{1j}\\ \beta_{2j} &= \gamma_{20}\\ \beta_{3j} &= \gamma_{30}\\ \beta_{4j} &= \gamma_{40}\end{aligned} \tag{3}$$

where γ_{00} is the mean intercept and γ_{10} is the mean linear growth rate. The variances of u_{0j} and u_{1j} and their covariance are denoted by σ_0^2, σ_1^2, and σ_{01}, respectively.

In the basic model the within-subdistrict residual variance was assumed to be the same in all level 2 units: Var $(e_{ij}) = \sigma_e^2$. The level 1 random variation can also be structured. Suppose we wish to fit separate level 1 variances for boys and girls. We can specify the level 1 random component as

$$e_{ij} = e_{0ij} + e_{3ij}Z_{3ij}$$

Then the level 1 variance is given by

$$\sigma_e^2 = \sigma_{e_0}^2 + \sigma_{e_3}^2 Z_{3ij}^2 + 2\sigma_{e_0e_3}Z_{3ij} \tag{4}$$

By constraining $\sigma_{e_3}^2 = 0$ we obtain for boys a level 1 variance of $\sigma_{e_0}^2 + 2\sigma_{e_0e_3}$ and $\sigma_{e_0}^2$ for girls. Note that the constraint is simply used as a device to avoid the overparameterization implied by (4).

In calculating the weight variances for fixed age groups we found that the dispersion increased with age. This suggested that we should model σ_e^2 as a function of T:

$$e_{ij} = e_{0ij} + e_{3ij}Z_{3ij} + e_{Tij}T_{ij}$$

giving

$$\sigma_e^2 = \sigma_{e_0}^2 + 2\sigma_{e_0e_3}Z_{3ij} + 2\sigma_{e_0e_T}T_{ij} + 2\sigma_{e_3e_T}Z_{3ij}T_{ij} + \sigma_{e_T}^2 T_{ij}^2. \tag{5}$$

In our data $\sigma_{e_T}^2$ was estimated as zero and $\sigma_{e_3e_T}$ was very small and not significant. This allowed a simplification

$$\sigma_e^2 = \sigma_{e_0}^2 + 2\sigma_{e_0e_3}Z_{3ij} + 2\sigma_{e_0e_T}T_{ij}$$

making within-cluster variation a *linear* (rather than quadratic) function of age.

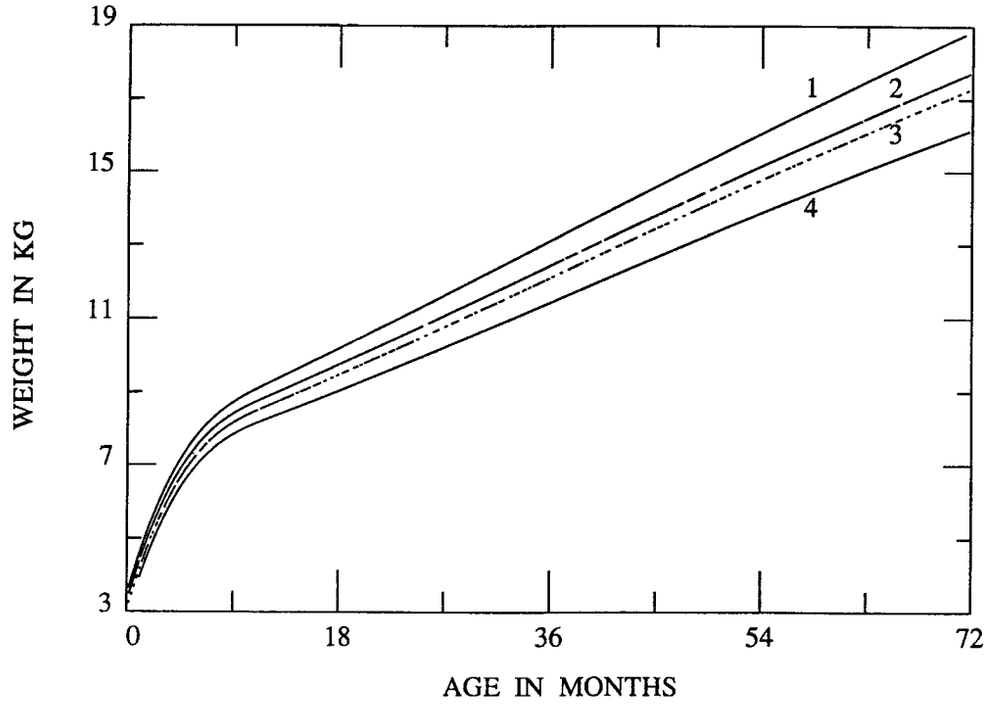
The procedure of fitting the above two-level growth model is quite similar to that of fitting an ordinary polynomial (see, for example, section 9.2 of Prosser, Rasbash, and Goldstein, 1990) except for the extra term in the grafted polynomial. Here is a segment of a *ML3* worksession log showing the construction of the fifth term of equation (2). The age variable was in a column called `'T'`, and C10 was empty.

```
CALC C10 = 'T' - 12
CHAN -12 0 C10 0 C10
CALC C10 = C10**3
NAME C10 'T-12CUBE'
EXPL C10
```

Figure 1 shows the model setup on the *ML3* SETTings screen.

```
EXPLanatory variables in    X0  T  TSQ  TCUBE  T-12CUBE  Z1  Z1T  Z2  Z2T  Z3T  Z3TSQ  Z3
FPARameters                 X0  T  TSQ  TCUBE  T-12CUBE  Z1  Z1T  Z2  Z2T  Z3T  Z3TSQ
FMEAns
RMEAns
RESPonse variable in        WEIGHT
IDENtifying codes for level 1: CHILD    level 2: SUBDIST  level 3:
RESEtting covariances level 1: ON       level 2: ON       level 3: ON
MAXIterations   5     TOLErance   2     METHod is IGLS    BATCh is OFF
LEVEL 3 RANDOM PARAMETER MATRIX unspecified
LEVEL 2 RANDOM PARAMETER MATRIX
         X0       T
X0       1
T        1        1
LEVEL 1 RANDOM PARAMETER MATRIX
         X0       T        Z3
X0       1
T        1        0
Z3       1        0        0
```

Figure 1: SETTings Screen for the Model in Equations 2 and 3



1 = middle-east urban 2 = south urban 3 = middle-east rural 4 = south rural

Figure 2: Average Weight Curves of Girls in Two Regions of China

All the fixed coefficients were statistically significant. Dividing the age range at 12 months via the $(T-12)_+^3$ term appears to have been helpful in this situation. The coefficient of this term was opposite in sign and nearly equal in absolute value to the coefficient of T^3. In effect, a cubic polynomial is fitted between birth and 12 months, and a quadratic is fitted in the period from 12 to 72 months.

The coefficient of region was 0.43 indicating that the average weight at birth in the middle east is considerably larger than that of the southeast. Urban children are 0.2 kilograms heavier at birth, on average, than rural children. An estimate of 0.016 for the coefficient of Z_1T indicates that the average growth rate is greater in the former region than in the latter. In other words, the regional difference in weight increases with age but at a low monthly rate. The urban-rural gap in linear growth is 0.013. Figure 2 is a plot of average weight curves for girls in urban and rural areas in the middle east and southeast of China.

There are between-region differences in linear growth rate not accounted for by Z_1 and Z_2; σ_1^2 is 4.8×10^{-5} and its standard error is 1.9×10^{-5}.

The average sex gap in linear growth is 0.049 in favour of boys. In addition, the estimate for $\sigma_{e_0e_3}$ is 0.085 indicating that the level 1 variance for boys is greater than that for girls by 0.17. As expected, level 1 variance increases with time: $\sigma_{e_0e_T}$ is 0.023.

Conclusion

We have shown that a two-level piecewise polynomial model can be fitted effectively to cross-sectional growth data involving a large span of ages. This model permits the examination of between-cluster variation and at the same time fits heterogeneous variances at level 1 to allow for sex differences in dispersion and changing variability with age. data

References

Goldstein, H. (1986). Efficient statistical modelling of longitudinal data. *Annals of Human Biology, 13,* 129-141.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: O.U.P.

Pan, H. & Goldstein, H. (1990). *A Two-level Growth Model Using Grafted Polynomials.* Manuscript submitted for publication.

Prosser, R., Rasbash, J., & Goldstein, H. (1990). *ML3-software for three-level analysis: Users' guide.* London: Institute of Education.

cont'd from p. 3

Muthén, B., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In D. Bock (Ed.). *Multilevel analysis of educational data* (pp. 87-99). New York: Academic Press.

Schmidt, W. (1969). *Covariance structure analysis of the multivariate random effects model.* Unpublished doctoral dissertation, University of Chicago.

Schmidt, W., & Wisenbaker, J. (1986). *Hierarchical data analysis: An approach based on structural equations* (CEPSE Research Series No. 4). Ann Arbor: University of Michigan, Department of Counselling, Educational Psychology, and Special Education.

cont'd from p. 7

Goldstein, H. (in press). Nonlinear multilevel models, with an application to discrete response data. *Biometrika.*

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data.* New York: Wiley.

Prosser, R. (1991). *Multilevel analysis with incomplete data: A Monte Carlo comparison of four treatments.* Unpublished doctoral dissertation, University of Toronto.

IN THE NEXT ISSUE...

Health Information Research Services and Multilevel Modelling

A new set of general macros for analysis of discrete data with *ML3*

NEW LITERATURE

The following articles have been brought to our attention recently:

Bondi, L., & Bradford, M. (1990). Applications of multilevel modelling to geography. *Area, 22,* 256-263.

Hox, J., de Leeuw, E., & Kreft, G. (1990). *The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model* (Methods & Statistics Series No. 45). University of Amsterdam, PAOW Faculty.

Jones, K. (1990). A multilevel modelling approach to immunization uptake. *Area, 22,* 264-271.

Stern, J., Curtis, M., Gillett, I., Griffiths, G., Maiden, M., Wilton, J., & Johnson, N. (1990). Statistical models for data from peridontal research. *Journal of Clinical Peridontol, 17,* 129-137.

If you have an article concerning multilevel modelling that you would like announced, please send the details and an abstract (or reprint) to Bob Prosser at the address on page 1.

In January 1988, the Dutch Multilevel Research Group held a workshop in Nijmegen on the theme of *Theory and Model in Multilevel Research: Convergence or Divergence?* Eight papers presented there have now been published by SISWO of Amsterdam. This volume may be ordered by remitting Dfl. 25 to the account of SISWO (41 36 44 944). The organization's address is: Postbus 19079, 1000 GB Amsterdam, The Netherlands.

This contribution will be reviewed in an upcoming edition of the *Newsletter.*

TRANSITIONS

After two years with the Multilevel Models Project, Research Officer Bob Prosser is heading home to Vancouver Canada to work as a consultant. Bob will continue to collaborate with the Project team, editing the *Newsletter*, handling North American software marketing and service, and co-authoring instructional materials on multilevel modelling.

New e-mail and postal addresses for Bob will be published in the next issue.

This change opens a full-time position on the team in London. The remaining period in the current ESRC grant is 32 months. The person appointed will focus on the dissemination of multilevel expertise in relatively new areas of application. Particularly important are social surveys, economics, and geography. She or he will be expected to collaborate with researchers in one or more of these fields and to give talks about the Project's work.

Strong computing and data analytic skills are required; experience in multilevel modelling would be an asset as would a first degree or equivalent experience in one of the target social sciences. Good organizational skills are essential: the new team member will have primary responsibility for organizing workshops and for general publicity.

The Project has access to some clerical support and good computing resources. Jon Rasbash, the other Research Officer, will continue to have primary responsibility for software development.

Letters and CVs should be sent to Professor Harvey Goldstein, Project Director, at the address on page 1. The application deadline is 18/01/91.

"MARKET RESEARCH"

Here's a chance to aid the Multilevel Models Project team in making decisions concerning future development of *ML3–Software for Three-level Analysis*! Our question is whether to make the rather substantial investment(s) of time in producing a *Macintosh ML3* and/ or a *Sun ML3.* Several people have commented on this, but additional feedback would be very helpful. Please contact Jon Rasbash at the address on the front page.

CONTRIBUTORS

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

Bengt Muthen
UCLA, Los Angeles, California
Huiqi Pan
W.H.O. Collaborating Center, London