Review: Review of Five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, ML3, VARCL
Author(s): Ita G. G. Kreft, Jan de Leeuw, Rien van der Leeden
Reviewed work(s):
BMDP-5VGENMODHLM, Version 2.1ML3, Version 2.2VARCL
Source: *The American Statistician*, Vol. 48, No. 4 (Nov., 1994), pp. 324-335
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2684845
Accessed: 23/04/2009 20:46

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=astata.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician.*

STOR®

http://www.jstor.org

Review of Five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, ML3, VARCL

Ita G. G. KREFT, Jan DE LEEUW, and Rien VAN DER LEEDEN

BMDP-5V

Designed by Jennrich and Schluchter. Available as a procedure in the BMDP package. We used release 6.0. There are a number of different BMDP implementations with corresponding prices. The software and manual for BMDP-5V on a MS-DOS PC costs $17.50, it requires the BMDP core package. All the options are available from BMDP Statistical Software, 1440 Sepulveda Boulevard, Suite 316, Los Angeles, CA 90025.

GENMOD

Written by Hermalin and Anderson, Population Studies Center, University of Michigan, from instructions provided by Wong and Mason. Available from William M. Mason, Department of Sociology, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90024. Two diskettes, manual $20.

HLM, Version 2.1

Written by Bryk, Raudenbush, and Congdon. Manual by the same authors with Seltzer. Available from Scientific Software, Inc, 1525 East 53rd Street Suite 906, Chicago, IL 60615. The program plus manual are $300. In Europe, the program is distributed by ProGamma (see VARCL).

ML3, Version 2.2

Software for two- or three-level analysis written by Rasbash. Manual is by Prosser, Rasbash, and Goldstein. Program based on theoretical work by Goldstein. Available from the Multilevel Models Project, Institute of Education, 20 Bedford Way, London WC1H 0AL, UK. The program plus manual are $475 (regular version) or $570 (extended memory version).

VARCL

Initiated by Aitkin and Longford and written by Longford. Distributed by ProGamma, P.O. Box 841, 9700 AV, Groningen, The Netherlands (e-mail:gamma.post-@gamma.rug.nl). The educational price (single-user license) for manual and software is $250. Noneducational users pay $350. A site license costs $750 or $1050, respectively. Prices include mailing and administration charges, but exclude tax. The source code is available only upon special request.

NOTE: Since this review was written, some of the programs were updated. The most recent update information we can present here dates from October 1993:

BMDP-5V is now available under software Release 7.0. Enhancements include additional built-in covariance structures and an ESTIMATE keyword-statement providing extra options for estimation and testing.

HLM is now available as Version 2.2. A three-level version of the program is soon to be released. Enhancements should include improved portability, a dynamic memory allocation, and more options for estimation and testing.

ML3 is now in Version 2.3. The program now can fit models with more parameters and some new commands were added for handling random cross-classifications.

The present version of VARCL is dated January 1990. A preprocessor is soon to be released, which will provide options for data transformation, aggregation and selection, as well as an interactive menu structure.

1. INTRODUCTION

We discuss five programs for the analysis of data with a hierarchical or nested structure. All programs compute maximum likelihood estimates of the parameters in a class of mixed linear models, known under various names such as *multilevel linear models*, *hierarchical linear models*, *empirical Bayes models*, *seemingly unrelated regressions*, or *random coefficient models*.

Hierarchical data structures are very common in the social and behavioral sciences. Individuals, for instance, are in social groups, and we can have variables describing the individuals as well as variables describing the groups. In multilevel models there is a separate (first-level) linear model for each group usually with the same predictors and outcome, but with different regression coefficients. The models are linked together by a second-level model in which the regression coefficients of the first-level regressions are regressed on the second-level predictors. We shall illustrate this structure with a number of examples.

The simplest type of application is in *repeated measurement* or *growth curve* analysis. Growth curves differ for individuals, which makes individuals the higher level within which the (repeated) observations are nested. Suppose we have n individuals, t time-points, m first-level regressors, and p second-level regressors. The first-level model is of the form $\mathbf{Y} = \mathbf{BX} + \mathbf{E}$, where $\mathbf{Y}$ is the $n \times t$ data matrix, $\mathbf{X}$ is the $m \times t$ matrix of first-level predictors, $\mathbf{B}$ is the $n \times m$ matrix of random regression coefficients, and $\mathbf{E}$ is the $n \times t$ matrix of disturbances, which are all independent $\mathcal{N}(0, \sigma^2)$. The second-level model is $\mathbf{B} = \mathbf{Z}\Gamma + \mathbf{\Delta}$, with $\mathbf{Z}$ the $n \times p$ matrix of second-level predictors, Γ the $p \times m$ matrix of fixed second order regression coefficients, and $\mathbf{\Delta}$ the $n \times m$ matrix of second-order disturbances. The rows of $\mathbf{\Delta}$ are independent of $\mathcal{N}(0, \Omega)$, and $\mathbf{\Delta}$ is independent of $\mathbf{E}$. Substituting the second-order model into the first-order one gives $\mathbf{Y} = \mathbf{Z}\Gamma\mathbf{X} + \mathbf{\Delta X} + \mathbf{E}$, which shows the close similarity to the familiar Potthoff-Roy model. $\mathbf{Z}\Gamma\mathbf{X}$ can be called the fixed part and $\mathbf{\Delta X} + \mathbf{E}$ the random part of the model. Compare Strenio, Weisberg, and Bryk (1983) for further details.

Ita G. G. Kreft is Associate Professor, Division of Educational Foundations and Interdivisional Studies, California State University, Los Angeles, CA 90032-8143. Jan de Leeuw is Professor, Departments of Psychology and Mathematics, University of California Los Angeles, Los Angeles, CA 90024-1555. Rien van der Leeden is Associate Professor, Department of Psychometrics and Research Methods, University of Leiden, 2300 RB Leiden, The Netherlands.

© 1994 American Statistical Association

A second type of application is in the analysis of *large-scale nonexperimental data.* The observations have a hierarchical structure, but now individuals are nested within contexts or groups. Individuals represent the lower level instead of the higher one. The leading example here is in educational research, where the data are often on two or more levels. Pupils are in schools, schools are in districts, districts are in counties, and counties are in states. In most cases these data are not balanced, because classes have different numbers of students. The first-level model is now written as $y_j = \mathbf{X}_j\beta_j + \epsilon_j$, and the second-level model is $\beta_j = \mathbf{Z}_j\gamma + \delta_j$. The terms ϵ_j are independent of $\mathcal{N}(0, \sigma^2\mathcal{I}_{n(j)})$, and the terms δ_j are independent of $\mathcal{N}(0, \Omega)$. Disturbance terms of both levels are independent. Also, $\mathbf{Z}_j$ has a *direct sum* structure, with m rows and mp columns

$$\mathbf{Z}_j = \begin{pmatrix} z'_{j1} & 0 & 0 & \cdots & 0 \\ 0 & z'_{j2} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & z'_{jm} \end{pmatrix}.$$

Suppose, for example, that the three first-level predictors are the intercept, student's IQ, and parental income, and the dependent variable is school success. School-level variables are the second-level intercept and school size t_j. In z_{j1} we collect the school-level variables predicting the intercept β_{j1}, z_{j2} predicts the slope of IQ β_{j2}, and z_{j3} predicts the slope of parental income β_{j3}. Each z_{js} could be taken to be equal to $(1, t_j)$. On the other hand we could also decide that school size only affects the intercept, not the slope of intelligence or income. Then $z_{j1} = (1, t_j)$, while z_{j2} and z_{j3} have the single element 1.

Substituting the second-level model into the first-level model gives $y_j = X_jZ_j\gamma + X_j\delta_j + \epsilon_j$, which means that $U_j = X_jZ_j$ has *block-rank-one* structure (De Leeuw and Kreft 1986),

$$U_j = \left(x_{j1}z'_{j1} \quad | \quad \cdots \quad | \quad x_{jm}z'_{jm} \right).$$

These types of applications were discussed in detail by Kreft (1987), Goldstein (1987), and Bryk and Raudenbush (1992).

Once we realize that social science data are often multilevel, we can easily find many other applications. In a sense the basic problem of the social sciences is to model the interaction between properties of individuals and properties of social groups. Children are in sibships, adults are in marriages, insects are in colonies, and so on. Multilevel linear models make it possible to combine variables of different levels quite naturally, and they model within-group correlations between observations in a simple way (Hanushek, 1974).

In the last few years a number of articles in the statistical and methodological literature have directly attacked the problem of analyzing variables measured at different levels of a hierarchy. From our point of view, the most important ones are Jennrich and Schluchter (1986), Mason, Wong, and Entwistle (1984), Raudenbush and Bryk (1986), Goldstein (1986), and Aitkin and Longford (1986). These articles describe the basic model, the likelihood function, an algorithm to maximize the likelihood, and a computer program. The programs we discuss in this article are (in alphabetical order) BMDP-5V, GENMOD, HLM3, ML3, and VARCL, corresponding to the five papers mentioned above. Except for BMDP-5V, these programs were developed mainly with nonexperimental grouped data in mind. BMDP-5V was developed primarily for the analysis of repeated measurement data.

In our review, each program is described and evaluated in turn for

- design philosophy,
- implementation details,
- models,
- routines,
- data setup and data handling,
- output, and
- user friendliness.

We also performed extensive numerical experiments on a number of representative data sets. The programs are slightly different in their computational outcomes, but basically they seem to converge to the same solutions in all cases (although sometimes very slowly and sometimes erratically). For more details on these computations, we refer to the reports by Kreft, de Leeuw, and Kim (1990) and Van der Leeden, Vrijburg, and de Leeuw (1991).

2. THE PROGRAMS

2.1 BMDP-5V

2.1.1 Design Philosophy. The BMDP-5V program was designed by Jennrich and Schluchter and described in Schluchter (1988). The theory on which the program is based was given by Jennrich and Schluchter (1986). The program was also documented and illustrated in the general BMDP user's guide.

BMDP-5V is especially designed for the analysis of repeated measures with special emphasis on unbalanced situations, including imbalance caused by missing data. Although the program has many possible applications, it is developed with small experimental data sets in mind, especially in a biological context. When there are many observations and/or a large number of parameters to be estimated, the program becomes expensive and somewhat clumsy to use.

2.1.2 Implementation Details. BMDP-5V is provided at any site where the BMDP package is running. Usually this package has been implemented on mainframes and mini computers, so there are versions working under different operating systems such as MVS, VM, VAX/VMS and so on. More recently, PC versions of the package have become available for use under MS-DOS.

The BMDP procedures can be operated either in batch mode or in interactive mode, depending on the user's preferences and the available computer facilities.

BMDP-5V shows no limitations concerning the number of individuals and variables. However, the memory needed for data storage will increase with these numbers and will also depend on the specified model. Probable implementation restrictions could therefore arise depending on the computer system used for running BMDP.

2.1.3 Models. The models that can be fitted in BMDP-5V belong to a general class of multivariate linear models. In these models a set of regression parameters describes the structure of the expected values of the observations, and a set of covariance parameters provides for a general parametrization of the within-subject covariances. A number of special structures for these covariances are built in. User-defined covariance structures can be specified too.

When repeated measures data are analyzed using a generalized multivariate linear model such as in BMDP-5V, the individuals are considered as data records and the observations on multiple occasions are interpreted as separate dependent variables. In a multilevel context, this makes individuals the second level within which the (first-level) observations are nested. By using missing data options (cf. Sec. 2.1.8) we can analyze unbalanced situations in which some individuals do not have observations on some timepoints.

Fitting multilevel models with random slopes with BMDP-5V is restricted to situations in which the values of the within-group predictors are identical for all groups, that is, in the case of repeated measurements the matrices $\mathbf{X}_j$ (see Sec. 1) are required to be equal for all individuals.

2.1.4 Routines. BMDP-5V allows the user a choice from three different algorithms to compute maximum likelihood estimates for all model parameters: a Newton–Raphson, a Fisher scoring, and a generalized EM algorithm. Producing identical results, these methods differ with respect to the number of iterations required and the costs per iteration. Generally, EM will converge slowly but with low costs per iteration, whereas Newton–Raphson will require the smallest number of iterations but has the highest cost per iteration. BMDP advises to use Newton–Raphson if the number of covariance parameters is 15 or fewer; otherwise EM is preferable. BMDP-5V offers two different algorithms to compute restricted maximum likelihood estimates for the covariance parameters: generalized EM and a quasi-scoring algorithm. Results will be identical but convergence is faster for the scoring method with higher cost per iteraton and vice versa for the restricted ML method. For any of these procedures, the maximum number of iterations is set to 15 by default. If necessary, this number can be altered by the user.

2.1.5 Data Setup and Data Handling. The data input file must be organized so that individuals are the records and the repeated measurements are the variables (we must have a one-record-per-individual file). Note that this is different from the usual way in which hierarchical data are organized.

We have run the program in batch mode. BMDP is a command language driven program. The instructions are given by keyword statements, comparable to SPSS or SAS statements. Input data and BMDP instructions can be stored in the same file or in separate files. Keywords control the names of the variables, their number, format, transformations, and so on. The actual modeling of the data is controlled by three keyword statements: DESIGN, MODEL, and STRUCTURE.

The DESIGN statement specifies the structure of the data. It identifies variables classifying subjects and variables measured repeatedly. The model itself is specified in the MODEL statement, using variable names. The STRUCTURE statement specifies the structure of the within-subject covariances. Covariances can be unrestricted. Other built-in special structures include first-order autoregressive models, compound symmetry, and banded or general autoregressive structures. For some structures additional input is needed. A factor analytic structure requires the number of factors to be specified. Random effects require some known matrix to be specified. A general linear structure requires the number of parameters to be specified as well as a set of known matrices. Finally, a user-defined covariance structure that is not built in and not linear can be specified by adding a FORTRAN77 subroutine to the program input. This routine becomes a part of the program, called in each iteration.

2.1.6 Output. The output of BMDP-5V is controlled by the PRINT statement and can be very comprehensive. There are no special output files. Program instructions, model specifications, and (default) values of various program parameters are extensively listed. An extended output presents, for each iteration, a table with the log likelihood of the specified model, the values of the regression parameters, and the values of the covariance parameters. Akaike's information criterion is provided to check the appropriateness of a chosen covariance structure. Asymptotic standard errors and z values are given for all maximum likelihood estimates of the model parameters. For the terms of the regression model, Wald (chi-square) tests of significance are included. Within-subject and all-pairs within-subject covariance and correlation matrices are given. A very useful option is a listing of individual responses (at the different time-points), their predicted values, residuals, and standardized residuals. With this listing, Mahalanobis distances are provided. These distances can be helpful in detecting cases that may be outliers.

2.1.7 User Friendliness. The BMDP user's guide is written very clearly. Any BMDP procedure, and thus also BMDP-5V, is extensively discussed and commented. The different options are very well illustrated presenting a variety of real data examples. Most technical details can be found in the appendix. The program is easy to use. The keyword command structure and the ability to run BMDP in interactive mode are considered user-friendly; the program is too.

When repeated measures are to be analyzed, the built-in autoregressive structures for the within-subject covariances can be very useful. However, including a user-defined covariance structure by adding a FORTRAN77 subroutine to the program may be quite annoying for regular users. Moreover the possibility of adding such subroutines is not available in the microcomputer versions.

2.1.8 Special Features. From an analysis-of-variance point of view, the ability of BMDP-5V to analyze incomplete data is a unique option. The data may be incomplete by design (e.g., a value of a grouping variable is missing) and/or some observations are missing (e.g., a measurement of the dependent variable or a measurement of a covariate is missing). To a certain extent, values for missing data can be imputed using an EM algorithm [further details can be found in Jennrich and Schluchter (1986); an example is given in the BMDP user's guide]. This property is very useful because it has been known to improve the reliability of the data analysis [see Little and Rubin (1987)]. The other packages have no possibilities for imputation, and cases with missing data are just listwise deleted. In case of repeated measurements, individuals with less observations than others are dealt with quite naturally, because they are simply considered level-two units with a smaller group size.

BMDP-5V can easily handle time-varying covariates. For instance, in the case of multivariate repeated observations—that is, two or more dependent variables measured repeatedly—we could take one (or more) of these variables as a covariate. This feature is built-in and easy to use.

2.2 GENMOD

2.2.1 Design Philosophy. The program implements the general model proposed by Wong and Mason. GENMOD is developed to accommodate two broad classes of applications: comparative analysis and contextual analysis. Contextual analysis is the usual type of analysis, also found in HLM, VARCL, and ML3. Comparative analysis is specific to GENMOD. The assumption here is that we have a different data file for each context; these files may have different formats and variables. Moreover, the micro data file for one context may also contain variables that are different. This characteristic of the program is very valuable in demography, the field for which this program was developed. Originally designed to analyze this kind of data, the most recent version, released in April 1989, provides the opportunity to use a single micro file as input and a single associated format statement (comparable to the other programs).

2.2.2 Implementation Details. GENMOD is written in FORTRAN77 and is currently compiled to run under the MS-DOS and MTS operating systems. File names must satisfy MS-DOS file-naming conventions. Under MS-DOS, the program reads and writes ASCII files only; MTS uses EBCDIC. The manual assumes that the program is running under MS-DOS; MTS tailoring is given in the appendix. There are three versions (GEN30, GEN40 and GEN50), which differ in the size of the real array storage that has been allocated (35,000, 45,000 and 55,000 elements of REAL*8 storage, respectively).

The distribution includes source code, however, which means that (at least in principle) any MS-DOS or OS/2 users with a suitable FORTRAN compiler can make their own version, with storage requirements adapted to their own environment.

2.2.3 Models. The basic model fitted in GENMOD is a two-level model. A special version of the program (GENMOD3) allows different contexts to have different first-level error variances. In the other programs, this variance is assumed to be equal over groups.

2.2.4 Routines. The maximum number of iterations is specified in the batch job. If convergence has not been achieved by the NUMITth iteration, the program will nonetheless stop, giving complete output as of the NUMITth iteration. The estimation procedure is restricted maximum likelihood. The EM algorithm used is based on equations developed by Mason and Wong.

2.2.5 Data Setup and Data Handling. The program runs in batch mode only. To run the program the user must create a setup file, a micro file, a macro file, and (optionally) an auxiliary data file. The setup file in GENMOD describes the model to be estimated and provides instructions for reading and saving information. The structure of the setup file is similar to an SPSS or SAS job, with the header being a specified output file name and three options of data input to choose from (raw data, cross products, or macro error variances and covariances). When the first setup is ready for a particular data set, only a few changes are needed in the setup for fitting further models. The setup files can be quite complex.

2.2.6 Output. The output consists of the usual basic information: information about convergence at every iteration, and restricted maximum likelihood (REML) estimates of the σ^2 and Ω parameters at the last iteration, as well as estimates of all the parameters. More can be obtained on request.

2.2.7 User Friendliness. The software is in some respects quite puzzling. The present version of the manual is not very clear about how and when to use certain options (the authors are working on this). An experienced user, or somebody who understands the methodology behind the program quite well, will encounter fewer problems than a novice. There is no example of the output files and no explanation of the outcome in the manual (although there are examples on the disks). The error messages, giving a clear idea of what goes wrong, are helpful.

2.3 HLM

2.3.1 Design Philosophy. HLM is an acronym for hierarchical linear model. The HLM program, Version 2.1, that implements this model is written by Bryk, Raudenbush, and Congdon. It is designed to handle multilevel data with two levels. Two broad classes of applications are accommodated by the program: contextual analysis and growth curve analysis. In contextual analysis we have the familiar multilevel data. For instance, in a study on school effects, we have a first level representing within-school analysis and a second level representing between-school analysis. In growth curve analysis, the first level represents individual change in a within-person model, whereas the second level represents effects of other variables upon these individual regressions (Bryk and Raudenbush 1992).

According to Kreft, de Leeuw, and Kim (1990), HLM is the most popular program in the United States because of its ease of use, interactive interface, and the availability of many significance tests. The informative and clearly written manual certainly contributes to this popularity. It provides a theoretical background for multilevel modeling and many useful references.

2.3.2 Implementation Details. There are two versions of the program available. The first is for workstations or mainframe computers with no real restrictions on the memory. The second is an adaptation for the PC, and it takes the 640 K memory limit of MS-DOS into account. Both versions are written in FORTRAN77, although for the MS-DOS version the main program and some screen control functions have been written in C. This mixed language feature of HLM and the particular type of screen control used make the program less than completely portable, although the computational routines are in straightforward FORTRAN77. Current program limitations for the PC version are as follows: There is a maximum of 10 within-unit variables per model. The input file and sufficient statistics file can contain 25 within-unit variables, 25 between-unit variables, and 300 units. In the between-unit model there is a maximum of 15 variables per equation, and the maximum on the total number of fixed effects over all equations is 35. A newer version with dynamic memory allocation has been announced.

2.3.3 Models. The basic model fitted in HLM is again the two-level model. By default, both the micro model and the macro model have an intercept, but the default can be overridden. For growth curves, for example, it can be interesting to fit models without a micro intercept. At the time of the run the user can introduce additional restrictions on the parameters. Some fixed regression coefficients and some covariance components can be set equal to zero. Thus we can have micro variables with only a fixed effect and micro variables with only a random effect, the default being that a micro variable has both.

2.3.4 Routines. Two routines are given in the manual: (a) an EM algorithm with an Aitkin accelerator, used as the core routine in the HLM program, and (b) a V-known routine. The V-known routine assumes the variance and covariance components are known quantities. It is useful mainly in research synthesis (meta analysis). See Bryk and Raudenbush (1992) for details and references. The number of iterations is, as usual, left to the decision of the user. The suggested number of iterations for exploratory analysis is 10. In most packages the advice is similar: 10 to 15 iterations. Our numerical experiments show that even for exploratory analysis 10 EM iterations are often not enough to get a good idea of where we are heading.

2.3.5 Data Setup and Data Handling. The raw data input file is usually plain ASCII, but it can also be either a "V-known" file or a system file for the SYSTAT statistics package. In case of a SYSTAT input file, the residual file (which is produced by the program) will also be a SYSTAT file. A V-known file has parameter estimates for each context and their associated sampling variance/covariance estimates (in addition to other second-level variables).

2.3.6 Output. The output of the γ coefficients is similar to the output of the other software packages. HLM provides a large number of statistical tests, both *t* type for the regression coefficients and chi-square type for the variance components. It also outputs so-called reliability coefficients, which are defined as the proportion of variance in the OLS regression coefficients that is second-level parameter variance.

2.3.7 User Friendliness. The program is of the interactive question-and-answer format, which makes it very easy to use. The manual contains the annotated output of several runs with different data sets. The organization of the manual, easy as it is for first use, also has its drawbacks. Specific information is not easy to find, since it is not organized under special headings. Special remarks and basic information are interwoven with examples of different kinds of output. There is only a small amount of background information. A nice feature of the program is that it allows many possibilities for exploration of the data.

2.3.8 Special Features. With BMDP-5V, HLM is the other program in our list that delivers a variety of tests. These are: (a) the *t* test for significance of the fixed parameters, (b) a chi-square test for residual unexplained variance in the first-level parameters, (c) a reliability estimate of the first-level variables, (d) the three hypothesis tests mentioned before, and (e) a test for homogeneity of variances. Homogeneity of variances is assumed, which means that a single micro-level error variance is estimated.

Three options are available for data input. Two of them are unique to HLM. One is the possibility of using SYSTAT files instead of ASCII files. For those with SYSTAT, this provides additional possibilities for data handling. Although it is not clear that SYSTAT is a particularly good choice in this context, it certainly is nice to have the additional option. The other unique input option is the V-known file.

2.4 ML3

2.4.1 Design Philosophy. ML3 is produced as part of the Multilevel Models Project of the Institute of Education at the University of London. This project is funded by the Economic and Social Research Council of the United Kingdom to extend the theory of multilevel modeling, to study the practical application of the models to real data sets, and to disseminate information about the theory and practice of this form of analysis. Among the specialized models that can be estimated using the program are growth curve models and multilevel logit models.

2.4.2 Implementation Details. ML3 is provided only in binary form. It runs on MS-DOS and OS/2 computers and needs 540K of RAM. There is also an extended memory version ML3E and a VAX/VMS-version ML3-V. Perhaps the most remarkable aspect of the ML3 implementation is that the multilevel software is merged with the kernel of the general-purpose package NANOSTAT, which offers a whole set of data-handling and data-transformation operations. NANOSTAT also provides descriptive statistics and high-resolution plots.

One important difference between ML3 and the other four programs is that the data are not first reduced to sufficient statistics and then kept in core memory. In ML3 the complete data matrix is read into core memory, which means that the restrictions on the size of the problem are more serious. The manual gives no clear-cut rules. In more practical terms this means that big examples cannot be analyzed unless you buy the extended memory version.

2.4.3 Models. The basic model fitted in ML3 is again the two- or three-level model. There is also the possibility, however, of having more complex error structures by incorporating more level I random terms. Also, log-linear and logistic models can be analyzed using standard GLM-type extensions.

2.4.4 Routines. An iterative, generalized, least-squares (IGLS) algorithm provides estimates of model parameters, and, when normality assumptions are met, these estimates are equivalent to maximum likelihood estimates. ML3 can also compute unbiased or restricted RIGLS estimates, which are called restricted maximum likelihood (REML) estimates in other contexts. Thus we distinguish, further on, ML3-F and ML3-R, for the options that use full or restricted maximum likelihood estimation. The user has the corresponding choice between IGLS, described by Goldstein (1986) and RIGLS, described by Goldstein (1989). That the distinction between the two is not really discussed in the manuals is a problem in all packages, but it is especially missed here because the choice between the two is stressed.

The number of iterations can range from 1 to 999. The default value is five iterations. This number is sufficient for reaching convergence when the conditions are favorable—that is, when the number of observations per unit is large enough to obtain stable estimates, the number of parameters to be estimated is small, and the tolerance/convergence criterion is not too stringent. It is advisable to increase the number of iterations when the convergence is reached slowly, the amount of data is small, and/or the number of parameters to be estimated is large.

2.4.5 Data Setup and Data Handling. The input file can be either raw data or a modified data set in one single file for the micro and macro data together. The data have to be sorted by context. ID's are needed for each level of the hierarchy. There can be missing data, but they have to be assigned a numerical code.

2.4.6 Output. The default output is minimal. Special output can be required by using special commands. For instance, the RESI command stores the residuals in columns to be specified by the user. The user's guide explains analysis of residual structures and gives some practical applications.

2.4.7 User Friendliness. The manual is a well-written and complete document. It is actually more than a manual, because it introduces the reader to the hows and whys of multilevel analysis with a multitude of references. A disadvantage of the manual is the somewhat delayed instruction on how to do a multilevel analysis. This is due to the extensive documentation of the NANOSTAT package. Because in most model fitting cases the user may need the use of NANOSTAT commands, we do not know how the authors could have prevented this circuitous route. But a HELP program is built in. This online help facility shows the commands and/or the format of the commands.

Of the multilevel programs, ML3 allows users the most freedom to choose input and even to make adjustments during the run. Our impression is that in order to make use of the full potential of the program, extensive experience (or going to one of the many workshops offered by the program authors) is necessary. The reward for the user is that the program obviates the necessity of preparing the data in advance in another package. Another advantage is that it is easy to make adjustments or new interactions in a later modeling stage.

2.4.8 Special Features. There are several special features. For instance, the option to enter starting values for the parameter estimates other than the default OLS estimation is special, as is the fact that a simple multilevel logit and log-linear model can be fitted. This allows the researcher to analyze survey data with proportions or binary variables as the dependent variable. Information about the convergence process is provided, and during the run the program can be interrupted to "freeze" the estimation of individual parameters for the rest of the run. Freezing slowly converging estimates speeds up the overall convergence of the model. Of course all the exploratory, graphical, and residual options of the NANOSTAT packages are also unique (and valuable) features.

2.5 VARCL

2.5.1 Design Philosophy. VARCL implements random coefficient analysis. This can be used to analyze multilevel data. It provides the option to fit random slopes but offers no possibilities for fitting interactions between variables of different levels. It comes in two versions. VARL3 analyzes data of at most three levels. VARL9 can be used for analysis of data with up to nine levels of nesting, but it permits only a simple variance components structure of the random effects. There is no requirement for the balance of the nesting structure in either program in the sense that group sizes are not required to be equal at the various levels.

2.5.2 Implementation Details. Both programs are written in FORTRAN77 and require an interactive computing environment. VARCL was originally written for VAX/VMS, but it has been ported successfully to MS-DOS, MAC-OS, and many UNIX environments. Since 1991 the program has been maintained on a UNIX workstation. VARL3 is complex and has a more elaborate interface than VARL9, but the two are similar enough to warrant a single user's guide. The interface of VARCL combines interactive and batch features. The batch feature is in the control file that contains declarations related to the data set such as the title, data file names (the data set may consist of several data files), formats, variable names, nesting structure, and so forth. Having this information available in a separate file makes the interactive session less tedious.

The implementation restrictions are defined in a small file (IMPLE.ADD), included in the main code, and so they can be changed very easily by recompilation. There is no limit on the maximum number of elementary level units. With the MS-DOS version we have worked with a maximum number of variables equal to 24, a maximum number of regression parameters of 24, and a maximum number of sufficient statistics equal to 30,000. The sufficient statistics are the regression coefficients, residual sums of squares, and cross-product matrices for each of the groups. Thus for N groups and m individual level predictors, we have approximately Nm^2 of these quantities.

2.5.3 Models. The models fitted by VARCL are somewhat different from those fitted by GENMOD, HLM, and ML3. More precisely, they are somewhat differently specified. This is because the program does not build in cross-level interactions by default. By using an input matrix with specially created interaction variables between levels, the program can be used in the same way as the other programs. Here again, no missing data can be handled in the model fitting stage. In VARL9 we have an inherently simpler structure for the error terms, because only random intercepts are allowed (i.e., first-level variables do not enter into the error structure).

2.5.4 Routines. For his VARCL program, Longford (1987) uses the Fisher scoring algorithm. The manual describes the algorithm in detail (quite unlike the black box approach in the ML3 and the HLM manuals). If at an iteration the estimated dispersion matrix has a negative eigenvalue, the corrections for all the parameters are cut in half. The program prints the message that a covariance adjustment has taken place. When the information matrix used in the scoring iterations becomes singular, the offending parameter is aliased (excluded from the model). Aliasing obviously improves the convergence but results in fitting a different model. It is irreversible (once a parameter is aliased, it will not be unaliased and left free to vary again). It has been our experience with VARCL that aliasing occurs frequently in situations with complicated models, in which the EM algorithms of GENMOD and HLM exhibit very slow convergence. In the case of aliasing and covariance adjustment, the VARCL manual suggests fitting a smaller model.

2.5.5 Data Setup and Data Handling. The input data matrix has to have a hierarchical ordering, as in the other programs. The basic information has to be provided in a separate batch file.

2.5.6 Output. A session of VARCL can be saved in a binary "dump" file containing the entire information required to carry on, in a new session of VARCL, where the old session was terminated. The dump files can only be used for data with normally distributed error terms. The output file contains the results of the analysis and a summary of the initial specifications. Several models can be fitted in a single session, and the results can be written to one output file.

2.5.7 User Friendliness. The VARCL manual is very useful. It contains much valuable background information concerning output and interpretation. The program is very easy to use, because it is interactive in the sense that questions have to be answered. Some work is involved when a batch job has to be prepared. Extra preparation is also needed when interactions between first- and second-level variables are of interest to the researcher. This happens in traditional multilevel models, which have cross-level interaction variables. VARCL requires the user to create these variables before starting the analysis. The model fitting part of this program is user-friendly. It is possible to fit a large number of models within the declared maximal model in a single session. The speed of the convergence is another nice feature. Comparable programs, such as ML3, take much longer to do the same job. The scoring is indeed fast. The error messages and the options to correct them are quite helpful as are the checks and opportunities to correct mistakes in a declared model.

2.5.8 Special Features. Unique is a quasi-likelihood adaptation for non-normal (binary, binomial, Poisson, and gamma-distributed) outcomes. The choices of the error structure are: normal error, binary or binomial error, Poisson error, or gamma error. In the interactive phase, there are explicit questions used to declare a covariance structure. All choices can be made between the two extremes: intercept by slope covariance only or (the other extreme) all (co)variances. Choices between these extremes are possible as well.

3. OVERALL COMPARISON

3.1 In the Restrictions

In the usual two-level models fitted by HLM, ML3, and VARCL

1) all variables in the random part are also included in the fixed part (i.e., the variables contained in matrix **X** in the equations in Sec. 1),
2) all level-one coefficients are random at level two (the full random coefficient model), and
3) the only variable for which the coefficient is random at level one is the intercept.

All five programs have ways to deal with the first two restrictions, and all leave room for variables that are not included in the fixed part to be random (HLM, ML3, BMDP-5V) or to fit a mixed model (all programs), but the last restriction is only overcome in ML3.

3.2 In the Output

The (default) output given by the five packages varies from one to several pages and from many parameters and significance tests to only the bare essentials. Both GENMOD and ML3 have very little output. They differ considerably from VARCL, HLM, and BMDP-5V which have a lot of output. For example, HLM provides t test values for parameters, in addition to the usual parameter estimates and their respective standard errors. Of course these separate t tests must be taken with a grain of salt, because there are so many (correlated) parameters. In addition, tests are more heavily dependent on statistical assumptions, such as normal distributions, than parameter estimates. The overall test (test of differences between deviances for the goodness of fit) may be more reliable here. The default output given by BMDP-5V is quite extensive too, and additional output options are available.

3.3 In the Handling of Raw Data

Packages differ in the way they handle the raw data set. Centering is a much discussed issue in recent publications [cf. Bryk and Raudenbush (1992), Kreft, de Leeuw, and Aiken (in press)]. The reason often given for grand-mean as well as context-mean centering is that it facilitates interpretation. It can also be used, however, as a way to improve the numerical performance of the estimation algorithm.

3.4 In the Algorithm

Information about the convergence process and ways to use this information differ significantly among the programs. In the program ML3 it is possible to interrupt the run and freeze the estimation of individual parameters for the rest of the estimation during the run. Setting the residual parameter variance at zero in the next run for those variables that slowly converge gives the same effect in HLM as is reached in ML3 by freezing during the run. Boundary problems are handled in various ways. Programs using the EM method need no special provisions to deal with boundary constraints. Estimated variance matrices are not permitted to have a negative/nonpositive eigenvalue. Nevertheless, EM methods that converge to boundary points generally have sublinear convergence. It is not entirely clear that the boundary is treated efficiently in ML3 and VARCL. The number of iterations needed to reach the same convergence criterion is very different over packages. VARCL and ML3, with fast linear or superlinear convergence, stay within the limit of 15 iterations, while GENMOD and HLM exceed that number considerably. Our experiments also show that more complicated models need many more iterations in packages that use the EM algorithm (GENMOD and HLM) than in the packages that use scoring (VARCL) or weighted least squares (ML3). In BMDP-5V, one can choose between Newton–Raphson, scoring, and EM.

EM algorithms for (co)variance component analysis are helpful, because they are relatively simple to program (especially in array-oriented interpreted languages such as APL, MATLAB, GAUSS), because they give monotone convergence and because they always stay within the boundaries of the parameter space. Their convergence can be tediously slow, especially for more complicated models. HLM uses Aitken acceleration, which makes a difference. The convergence of GENMOD is sometimes intolerably slow. This is due in part to the very strict convergence criteria in GENMOD. The scoring algorithm of VARCL will tend to give much faster convergence, although sometimes various parameters of the process have to be adjusted because of singularity, boundary conditions, negative eigenvalues, and divergence. We have observed sublinear convergence of VARCL in some examples, probably a result of making smaller and smaller steps to keep the variances positive definite. Using the results of Lindstrom and Bates (1988) could very well produce a more robust implementation of the scoring algorithm. The ML3 algorithm works quite well in almost all cases. Because all the data have to be kept in (limited) core memory, it cannot analyze really large examples. We think that this is a high price to pay for the relatively small gain in additional generality.

3.5 In Measurement Level of the Dependent Variable

ML3 and VARCL allow dichotomous dependent variables or variables that are multinomial frequencies. We understand that there is also a version of GENMOD for dealing with logistic multilevel models, called MULTILOGIT, but we have no experience with it.

3.6 In the Results

The reports by Kreft, de Leeuw, and Kim (1990) and van der Leeden et al. (1991) presented comparisons of the results of analyses with GENMOD, HLM, ML3, VARCL, and BMDP-5V. For the first set of comparisons we used two different datasets: SIMS and WEBB. The programs were compared with the exception of BMDP-5V, since the structure of the datasets used did not allow the straightforward use of this program. In a second comparison, BMDP-5V was compared with ML3 and HLM in a growth curve analysis using the data set DENTAL containing repeated measurements. We summarize the results here, after briefly describing the data sets and models used.

The Second International Mathematics Study (SIMS) data set (available on the ftp server ftp.stat.ucla.edu in pub/data/various/SIMS) was taken from a national sample of United States eighth-grade students who took a series of mathematics achievement tests conducted by the International Association for the Evaluation of Educational Achievement in 1981–1982. For this study, 3,691 cases out of approximately 7,500 were extracted. There were 190 school classes. Only two student-level variables, the sum of PRETEST core items and the GAIN score (difference between POSTTOT and PRETOT), are used. The second-level variable is Opportunity to Learn (OTL).

The within-group model is

$$(\text{GAIN})_{ij} = \beta_{0j} + \beta_{1j}(\text{PRETOT})_{ij} + \epsilon_{ij}$$

and the between-group model is

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(\text{OTL})_j + \delta_{0j},\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(\text{OTL})_j + \delta_{1j}.\end{aligned}$$

The first version of the model has $\delta_{1j} \equiv 0$ and $\gamma_{01} = \gamma_{11} = 0$ (a random intercept model); the second version has a random intercept and a random slope but still no second level variable (so $\gamma_{01} = \gamma_{11} = 0$). Results of the analyses are summarized in Tables 1 and 2. By comparing computing time (*iterations* multiplied by *time*), we found that VARCL and HLM are clearly the two fastest programs for simple models (with only a random intercept). The time needed to reach the convergence criterion in the more complicated models was much longer for all programs, but extremely so for HLM. VARCL is still by far the fastest, and GENMOD is by far the slowest, but the HLM program is fairly slow in this case as well. The faster programs for fitting complicated models are the two that use full information maximum likelihood (FIML), VARCL, and ML3-F. In fitting the above SIMS data on a 286 machine, for example, GENMOD used 145 iterations, which took 515 seconds each. HLM converged in 59 iterations, which took 25 seconds each. ML3 needed 10 iterations of 200 seconds each, and VARCL 13 iterations of 18 seconds each. The GENMOD total time (more than 20 hours) is exceptionally high, but it is partly due to the fact that GENMOD uses a default stop criterion that is much more stringent than HLM, for instance.

As can be seen from Tables 1 and 2, the programs gave very similar results for fixed as well as random parts, for the first level as well as for the interaction coefficients.

Table 1. SIMS Data, Random Slope Model, No Macro Variable

	GENMOD	HLM	ML3-R	ML3-F	VARCL
γ_{00}	7.060	7.060	7.060	7.055	7.0553
γ_{10}	−0.186	−0.186	−0.186	−0.186	−0.186
σ	22.23	22.23	22.24	22.24	22.240
ω_{00}	14.52	14.53	14.49	14.36	14.33
ω_{11}	0.009	0.009	0.009	0.0088	0.00885
ω_{10}	−0.2342	−0.237	−0.234	−0.229	−0.229
Iterations	189	76	10	10	14
Time	480	16	180	142	11
Deviance		22382.4		22373.1	

Table 2. SIMS Data, Random Slope Model With OTL as Macro Variable

	GENMOD	HLM	ML3-R	ML3-F	VARCL
γ_{00}	0.06273	0.06916	0.06191	0.03242	0.03913
γ_{01}	0.23419	0.23402	0.2342	0.2349	0.23470
γ_{10}	−0.22833	−0.22938	−0.2282	−0.2236	−0.22447
γ_{11}	0.00086	0.00089	0.00085	0.00072	0.00075
σ	22.13	22.13	22.13	22.14	22.14
ω_{00}	12.65	12.68	12.64	12.38	12.36
ω_{11}	0.0119	0.0114	0.0111	0.0104	0.0100
ω_{10}	−0.2302	−0.2329	−0.2300	−0.2205	−0.2200
Iterations	145	59	10	10	13
Time	515	25	242	165	18
Deviance		22367.8		22340.7	

The main difference is between restricted and full information maximum likelihood. Other differences are between VARCL and the other programs. However, generally the differences are small. We expected that much, since the groups are of about equal size, the predictors are not too correlated, and a large number of observations within and between groups was present. The outcomes of comparable programs are the same up to two decimals. A difference is that the first three programs use a restricted maximum likelihood method, while the last two use full maximum likelihood. This is also the difference between ML3-R and ML3-F, since ML3 offers a choice between the two estimation procedures. The difference in the solutions produced by the different estimation methods (R or F) is clear in the two tables, while the difference is more pronounced in complicated models with random slopes or when a small data set is used.

One such small dataset is the WEBB data (available on the ftp server ftp.stat.ucla.edu in pub/data/various/webb) (Webb 1982). This set comprised data from 96 students (grades 7 and 8) in three average-ability Los Angeles junior high schools. They were in 35 small groups. The example (data provided by Noreen Webb, Graduate School of Education, UCLA) is interesting, because the number of groups was relatively large, and the number of individuals per group was small. Individual level variables were post-test (POST), which is the dependent variable, pretest (PRE), and a student-variable (NOA): asking a question and not getting an answer.

The group level variable was the pretest mean in the group (PREM). The model was

$$(\text{POST})_{ij} = \beta_{0j} + \beta_{1j}(\text{PRE})_{ij} + \beta_{2j}(\text{NOA})_{ij} + \epsilon_{ij}$$

and the between-group model was

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(\text{PREM})_j + \delta_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(\text{PREM})_j + \delta_{1j}\\ \beta_{2j} &= \gamma_{20} + \gamma_{21}(\text{PREM})_j + \delta_{2j}.\end{aligned}$$

After some preliminary exploration we decided on a model with both $\delta_{1j} \equiv 0$ (coefficient of PRE is nonrandom) and $\gamma_{11} = 0$ (no effect of pretest mean on pretest slope). Thus the single-equation specification of the model was

$$\begin{aligned}(\text{POST})_{ij} = \gamma_{00} &+ \gamma_{01}(\text{PREM})_j + \gamma_{10}(\text{PRE})_{ij} + \gamma_{20}(\text{NOA})_{ij}\\ &+ \gamma_{21}((\text{NOA})_{ij}(\text{PREM})_j)\\ &+ \delta_{2j}(\text{NOA})_{ij} + \delta_{0j} + \epsilon_{ij}.\end{aligned}$$

In this example the number of variables outnumbered the number of observations per group. Traditional packages (SPSS for instance) that use LS estimation cannot estimate parameters uniquely within such small groups. GENMOD did not work either and gave an error message about singular matrices. Results of the analyses are summarized in Table 3. We can conclude from Table 3 that, in general, estimates differ quite a lot between programs, even among those that use the same estimation method.

The DENTAL data, familiar to anybody working in growth curves (Potthoff and Roy 1964), were collected at the University of North Carolina Dental School and concern measurements of the distance (DIST) from the center of the pituitary to the pteryomaxillary fissure for 11 girls and 16 boys at ages 8, 10, 12, and 14. We have analyzed these data by fitting linear growth curves found as a regression of distance on age. In a multilevel framework this applies to the first level: the repeated measurements are nested within subjects. The growth curve coefficients

Table 3. WEBB Data, Random Slope Model, PREM Macro Variable

	GENMOD	*HLM*	*ML3-R*	*ML3-F*	*VARCL*
γ_{00}		10.9277	11.6600	11.3500	12.0932
γ_{01}		0.2545	0.0764	0.0921	0.0171
γ_{10}		3.2786	3.3420	3.3540	3.3428
γ_{20}		0.1677	0.1204	0.2575	−0.0959
γ_{21}		−3.9235	−4.1170	−4.0640	−4.1021
σ		25.8800	26.1700	25.3300	25.7877
ω_{00}		43.4384	46.1800	43.8600	3000.8270
ω_{20}		4.4773	4.8680	4.4580	3.8280
Iterations		200	26	168	13
Time		5.3	9.5	6.4	4.7
Deviance		619.9		611.2	

were treated as random variables at the second level. We used SEX as a second level explanatory variable (defined as a dummy variable with values −1 for girls and 1 for boys). Thus, apart from random variation across subjects, we account for growth curve coefficient variability across sex groups too.

The within-subject model can be written as

$$(\mathrm{DIST})_{ij} = \beta_{0j} + \beta_{1j}(\mathrm{AGE})_{ij} + \epsilon_{ij}$$

and the between-subject model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\mathrm{SEX})_j + \delta_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\mathrm{SEX})_j + \delta_{1j}.$$

This random coefficient growth curve model allows each person to have her own unique set of parameters β_{0j} and β_{1j}, that is, her own growth curve. Solutions were computed using BMDP-5V, HLM, and ML3. Results are given in Table 4.

Comparing solutions, we find that the three programs produce similar though not identical results. Estimates for fixed and random parameters (i.e., the variance component estimates) are the same for BMDP-5V and ML3-F. However, results from HLM differ from these to a certain extent. This also holds for the estimates of the random parameters in the ML3-R solution. For the most part the differences in the solutions can be explained by the use of different estimation methods: HLM and ML3-R use restricted maximum likelihood, whereas our BMDP-5V runs and ML3-F uses maximum likelihood. Differences may also arise because we use a small data set here (see the previous analyses with corresponding remarks).

If we compare the number of required iterations, our results clearly indicate the slow convergence of HLM.

3.7 In the Smoothness of the Finished Product

Some of the programs were still under development during our testing. We discovered bugs in GENMOD, HLM, ML3, and VARCL. We reported these problems to the authors, and the problems were largely corrected. We are now reasonably sure about the stability of ML3 and BMDP-5V, somewhat less about VARCL, and even less about HLM, for which a major upgrade is still overdue.

The comparisons are also made difficult because the programs are different in various unfortunate aspects (at least for our purposes). ML3 in earlier versions did not write the value of the likelihood function or the deviance, HLM writes the value of the restricted log-likelihood, VARCL the value of the unrestricted deviance, and GENMOD writes both values (but minimizes only the first one; moreover, it seems to write the wrong value). For large examples GENMOD does not give sufficient precision in the output to compare values of the likelihood function with those of other programs because the authors want to have the output for each iteration on a single 80-column line.

The default stopping criteria for the programs are very different. ML3 and VARCL have fast linear convergence; in fact, the convergence is actually close to superlinear if the model fits well. GENMOD and HLM typically have slow linear convergence, and the default stopping criteria for GENMOD are much more conservative than those of HLM. Thus comparisons of convergence should really be in terms of the likelihood function, but we have already seen that this leads to unexpected difficulties. Some programs have restrictions. HLM refuses to perform at least some functions if a within-group cross-product matrix is singular. BMDP-5V requires the within-group predictors to be identical for all groups when fitting models with

Table 4. DENTAL Data, Random Coefficients Growth Curve Model

	BMDP-5V	*HLM*	*ML3-F*	*ML3-R*
γ_{00}	16.8567	16.8993	16.8570	16.8570
γ_{10}	0.6320	0.6287	0.6320	0.6320
γ_{01}	−0.5161	−0.3824	−0.5161	−0.5161
γ_{11}	0.1524	0.1419	0.1524	0.1524
σ	1.7162	1.7651	1.7162	1.7162
ω_{00}	4.5569	5.7509	4.5569	5.7861
ω_{01}	−0.1983	−0.3082	−0.1983	−0.2896
ω_{11}	0.0238	0.0356	0.0238	0.0325
Iterations	2	60	2	4
Deviance	427.80	431.73	427.81	—

random slopes. VARCL uses aliasing and covariance adjustment to avoid negative eigenvalues of the estimated covariance matrices. This is perhaps a good idea, but again it makes comparisons very difficult.

Thus if we say that our results indicate that programs usually converge to the same solution, we do not mean that they do so in the very first try. What we mean is that we arrive at similar results, but sometimes only after quite a bit of program coaching.

4. CONCLUSIONS

It is difficult to summarize the results of our comparisons, but we shall try to give the main conclusions. For ease of reference, we collect some of the more important conclusions in Table 5.

In general, it follows from our analysis that even if we restrict ourselves to only two-level models with random slopes, we have very complicated likelihood surfaces. Maximizing the likelihood is inherently a difficult problem, unless the model is approximately true and the sample size is really large (in which case OLS will give very good starting values). Investigators (if the past is any indication) will tend to choose models that are too complicated (5 levels with 10 variables on each level). This leads to impossibly difficult search problems over the space of models and to impossibly difficult likelihood maximization problems. None of the programs reviewed here can handle such problems gracefully; but this is clearly not a shortcoming of the programs. All five can be misused rather easily.

None of the bugs we found is very serious. All five programs tend to converge to the same solutions, which is reassuring, although there are some unpleasant exceptions.

The five programs cover different though overlapping sets of problems. BMDP-5V analyzes repeated measures and has to be forced into multilevel mode. It is batch-oriented and requires the basic BMDP driver. More interactive use is also possible, using the line editor or the full screen editor.

HLM is very simple to use, has a pleasant interface, and makes many decisions for the user. This is a major advantage in some respects, a major disadvantage in others. It invites uncritical use, and it gives little indication if something goes wrong. Of all the programs we tried, HLM seemed to be the least reliable, which is probably the reason the authors are working on a complete rewrite. We think HLM, which is undoubtedly the most popular multilevel program on this side of the Atlantic, is dominated in speed by VARCL, in flexibility and completeness by ML3, and in reliability by both.

ML3 is much less easy to use; the user has to know more and obviously gets more value for the money (NANOSTAT data, generalized linear models, choice between loss functions, nonstandard modeling of levels, more extensive manual, faster convergence). But uncritical use and lack of failure indicators are problems here too. ML3 is upgraded regularly, and the team developing the program and the supporting documentation seems to be very active and responsive. Recently, ML3 modules that can be used to extend the program to binary response variables have become available for anonymous ftp. Experienced users can easily write similar extensions. We think ML3 is the most appropriate program for serious users and certainly for people doing research in multilevel analysis.

VARCL is not less expensive than HLM, but nevertheless VARCL is not a true commercial product. The main reason for the distinction is that VARCL comes with source (at least it used to), and that VARCL can be freely distributed within an institution (although user support and related services are only provided to registered users). This may seem to be of limited interest to the average user, but we think it is very important. These days, mixed networks are the rule with maybe six or seven different computer systems. For commercial packages this means that either not all versions are available or one has to buy six different versions. For a noncommercial product, VARCL is excellent. The source is very portable, the interface is easy, and the errors and warnings are informative. To do complicated cross-level analyses, however, you have to do quite a lot of preliminary data handling. Convergence is rapid, many levels are possible, and the GLM extensions are quite useful. VARCL is possibly the best program for rapid and reliable multilevel analysis for the incidental user. It is easy to switch models, and with a well-prepared set of data the analyses can be done very rapidly.

Table 5. Summary of Comparisons

Characteristics	*GENMOD*	*HLM*	*ML3*	*VARCL*	*BMDP-5V*
Availability	Shareware	Commercial	Commercial	Commercial	Commercial
Ease of use	Hard	Easy	Fairly easy	Easy	Fairly easy
Loss function	REML	REML	REML and ML	ML	ML
Data manipulation	None	Limited	Unlimited	None	Limited
Interface	Batch	Interactive	Interactive	Interactive	Batch or interactive
Preparing the dataset	Identify and order data	Identify and order data	Interactive	Code interactions	Code dispersion matrices
Weighting	No	Yes	No	Yes	No
Variance-covariance adjustments	EM algorithm	EM algorithm	Unclear	Aliasing and covariance adjustments	EM algorithm or adjustments
Small data sets	No	Yes	Yes	Yes	Yes
Documentation	Not good	Good	Very Good	Good	Good
Ease of learning	Hard	Very easy	Hard	Easy	Very easy
Error Handling	Good	Moderate	Moderate	Moderate	Good
Speed	Slow	Not fast	Not fast	Very fast	Fast

GENMOD, finally, is certainly inexpensive. It comes with source, but both source and manual are not finished products and may require quite a bit of tinkering to port. The interface is batch and thus very old-fashioned. There are some options the other programs do not have, but generally we feel that GENMOD is for the hobbyists among us.

In our comparisons we have not addressed the usefulness of the statistical information: Are the likelihood ratios close to chi-squares? How accurate are the standard errors? Do the estimates really improve the mean square error of OLS and WLS estimates? Such questions are important, in fact more important than computational speed or a friendly interface, but they require more complicated research. Once you know that hierarchies exist, you see them everywhere. Thus the applicability of the software seems almost unlimited. This pleases the authors of the programs, who have no interest in pointing out limitations and shortcomings of their products. We think that it is time to do sampling, resampling, and cross-validation studies to get a more realistic idea about the possibilities and limitations of the techniques.

REFERENCES

Aitkin, M., and Longford, N. (1986), "Statistical Modeling Issues in School Effectiveness Studies," *Journal of the Royal Statistical Society, Series A*, 149, 1–43.

Bryk, A., and Raudenbush, S. (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Newbury Park, CA: Sage Publications.

de Leeuw, J., and Kreft, I. (1986), "Random Coëfficient Models for Multilevel Analysis," *Journal of Educational Statistics*, 11, 57–86.

Goldstein, H. (1986), "Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares," *Biometrika*, 73, 43–56.

——— (1987), *Multilevel Models in Educational and Social Research*, London: Griffin.

——— (1989), "Restricted (Unbiased) Iterative Generalized Least Squares Estimation," *Biometrika*, 76, 622–623.

Hanushek, E. (1974), "Efficient Estimates For Regressing Regression Coefficients," *The American Statistician*, 28, 66–67.

Jennrich, R., and Schluchter, M. (1986), "Unbalanced Repeated Measures Models With Structured Covariance Matrices," *Biometrics*, 42, 805–820.

Kreft, I. (1987), *Models and Methods for the Measurement of School Effects*, Utrecht, The Netherlands: Elinkwijk.

Kreft, I., de Leeuw, J., and Aiken, L. (in press), "The Effect of Different Forms of Centering in Hierarchical Linear Models," *Multivariate Behavioral Research*, 29.

Kreft, I., de Leeuw, J., and Kim, K.-S. (1990), "Comparing Four Different Statistical Packages For Hierarchical Linear Regression: GENMOD, HLM, ML2, VARCL," Technical Report 50, UCLA Statistics Program, Los Angeles, CA.

Lindstrom, M., and Bates, D. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Replicated Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Little, R., and Rubin, D. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Longford, N. (1987), "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models With Nested Random Effects," *Biometrika*, 74, 817–827.

Mason, W. M., Wong, G. Y., and Entwisle, B. (1984), "Contextual Analysis Through the Multilevel Linear Model," *Sociological Methodology*, 72–103.

Potthoff, R., and Roy, S. (1964), "A Generalized Multivariate Analysis of Variance Model Useful Specially for Growth Curve Problems," *Biometrika*, 51, 313–326.

Raudenbush, S., and Bryk, A. (1986), "A Hierarchical Model for Studying School Effects," *Sociology of Education*, 59, 1–17.

Schluchter, M. (1988), "BMDP-5V—Unbalanced Repeated Measures Models with Structured Covariance Matrices," Technical Report 86, BMDP Statistical Software, Los Angeles, CA.

Strenio, J., Weisberg, H., and Bryk, A. (1983), "Empirical Bayes Estimation of Individual Growth Curve Parameters and Their Relationship to Covariates," *Biometrics*, 39, 71–86.

van der Leeden, R., Vrijburg, K., and de Leeuw, J. (1991), "A Review of Two Different Approaches for the Analysis of Growth Data Using Longitudinal Mixed Linear Models: Comparing Hierarchical Linear Regression (ML3, HLM) and Repeated Measures Design With Structured Covariance Matrices (BMDP-5V)," Technical Report 98, UCLA Statistics Program, Los Angeles, CA.

Webb, N. (1982), "Group Composition, Group Interaction and Achievement in Cooperative Small Groups," *Journal of Educational Psychology*, 74, 475–484.