Chapter 11

Software, missing data and structural equation models

11.1 Software for multilevel analysis

Traditionally, statistical analysis packages for the analysis of linear or generalised linear models have assumed a single level model with a single random variable. For the models described in this book such software packages are clearly inadequate, and this led, in the mid 1980's, to the development of four special purpose packages for fitting multilevel models. One of these, GENMOD (Mason et al., 1988), is no longer generally available. The other three are HLM (Bryk et al., 1988), *MLwiN* (Rasbash et al., 1999) and VARCL (Longford, 1988). A detailed review of these four packages (including ML3 which subsequently became MLn and then *MLwiN*) has been carried out by Kreft et al (1994). In their original form HLM, ML3 and VARCL were designed for continuous Normally distributed response variables and all three produced maximum likelihood (ML) or restricted maximum likelihood (REML) estimates. All three were soon able to fit 3-level models and VARCL and ML3 developed procedures for fitting Binomial and Poisson response models using the first order marginal approximation described in chapter 5. In addition VARCL is able to fit a variance components model with up to nine levels. Subsequently, the major statistical packages, notably BMDP, SAS and GENSTAT, have begun to incorporate procedures for ML and REML estimation for Normal response models. The packages EGRET and SABRE will obtain ML estimates for a 2-level logit response model. A Bayesian package using Markov Chain Monte Carlo (MCMC) estimation, BUGS, is also available and *MLwiN* allows MCMC estimation for a range of models. Appendix 11.1 contains details of where these and other programs can be obtained.

The two packages, *MLwiN* and BUGS, are able to fit nearly all the models described in this book, although not currently structural equation models. These latter models can be fitted by the program BIRAM, listed in Appendix 11.1. The programs Mln and *MLwiN* allow an effectively unlimited number of levels to be fitted, together with case weights, measurement errors and robust estimates of standard errors. They also have a high level MACRO language which will allow a wide range of special purpose facilities to be incorporated. A number of the papers referenced in earlier chapters have carried out their estimation procedures using special purpose software written in statistical programming languages such as S-Plus or Gauss. For the most part, however, this approach is computationally inefficient for the analysis of large and complex data sets, and the use of one of the special purpose packages is then essential, even when powerful mainframe computers are used. The general purpose packages, SAS, GENSTAT and also *MLwiN* allow a wide variety of data manipulations to be carried out within the software whereas the others tend to demand a somewhat rigid data format with limited possibilities for data transformations etc.

It is reasonable to expect that the standard multilevel models will soon be available within most of the major general purpose statistical packages. For the more complex models, such as those with multivariate outcomes, nonlinear relationships and complex variation at all levels, it will be important to have a user interface which assists understanding the complexity of structure when specifying models and when interpreting output. Because the level of complexity of multilevel models is greater than that associated with single level linear or generalised linear models, the importance of helpful user interfaces cannot be overemphasised if the best use is to be made of these models. The ability to work interactively in a graphical environment will also be important and it will be necessary for programs to optimise computations so that very large and complex datasets can be handled within a reasonable time (Goldstein and Rasbash, 1992).

11.2 Design issues

When designing a study where the multilevel nested structure of a population is to be modelled, the allocation of level 1 units among level 2 units and the allocation of these among level 3 units etc. will clearly affect the precision of the resulting estimates of both the fixed and random parameters. The situation becomes more complex when there are random cross classifications and where there are several random coefficients. There are generally differential costs associated with sampling more level 1 units within an existing level 2 unit as opposed to selecting further level 1 units in a new level 2 unit. At the present time there appears to be little empirical or theoretical work on issues of optimum design for multilevel models..

Some approximations for studying the standard errors of the fixed coefficients have been derived by Snijders and Bosker (1993) in the case of a simple 2-level variance components model. They are concerned with students sampled within schools and assume that the cost of selecting a student in a new school is a fixed constant times the cost of selecting a student in an already selected school. They also assume that there is a minimum of 11 students per school. They tend to find that, where this constant is greater than 1 and the total number of students to be sampled is fixed, the sample of schools should be as large as possible, although this will not necessarily be true for all the coefficients of interest.

Where cost information is available, together with some idea of parameter values, perhaps from a pilot study, then a guide to design efficiency can be obtained by simulating the effect of different design strategies and studying the resulting characteristics of the parameter estimates, such as their mean squared errors. This will be time consuming however, since for each design a number of simulated samples will be required. On the other hand, in certain areas, such as that of school effectiveness or animal and human growth studies, where information about costs and parameter values is often available, it would be possible to derive some generally useful results.

11.3 Missing data

A characteristic of most large scale studies is that some of the intended measurements are unavailable. In surveys, for example, this may occur through chance or because certain questions are unanswered by particular groups of respondents. We are concerned with missing values of explanatory variables in a multilevel model. An important distinction is made between situations where the existence of a missing data item can be considered a random event and where it is informative and the result of a non random mechanism. Randomly missing data may be missing 'completely at random' or 'at random' conditionally on the values of other measurements. The following exposition will be concerned with these two types of random event. Where data cannot be assumed to be missing at random, one approach is to attempt to model the missingness mechanism, and then to predict values from this model. Such predictions can be treated in similar ways to those described below.

We consider the problem of missing data in two parts. First we develop a procedure for predicting data values which are missing and then we study ways of obtaining model parameter estimates from the resulting 'filled-in' or 'completed' data set. The prediction will use those measurements which are available, so that data values which are missing at random conditional on these measurements can be incorporated. Detailed discussions of missing data procedures are given by Rubin (1987) and Little (1992).

The basic exposition will be in terms of a single level model for simplicity, pointing out the extensions for multilevel models.

11.4 Creating a completed data set

Consider the ordinary linear model

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i \qquad (11.1)$$

for the i-th unit in a single level model. Suppose that some of the x_{1i} are missing completely at random (MCAR) or conditionally missing at random (MAR) conditional on X_2. Label these unknown values x_{1i}^*. We consider the estimation of these by predicting them from the remaining observations and the parameter set q for the prediction model, namely

$$\hat{x}_{1i}^* = E(x_{1i}^* | x_{2i}, y_i, q) \qquad (11.2)$$

Where we have multivariate Normal data the prediction (11.2) is simply the linear regression of X_1 on X_2, Y, where the coefficients of this regression prediction are obtained from efficient, for example maximum likelihood, estimates of the parameters of the multivariate Normal distribution. This can be achieved efficiently using the procedures for modelling multivariate data described in Chapter 4. We shall consider the case of non-Normal data later.

We define a multivariate model with three response variables, Y, X_1, X_2 and three corresponding dummy variables, say Z_0, Z_1, Z_2. Some level 2 units will have all three response variables, but others will have only two where X_1 is missing. Write this as the 2-level model

$$v_{ij} = b_{0j} z_{0ij} + b_{1j} z_{1ij} + b_{2j} z_{2ij},$$
$$b_{0j} \sim N(m_Y, s_Y^2), \quad b_{hj} \sim N(m_h, s_h^2), \quad h = 1,2 \qquad (11.3)$$

together with the three covariances to give the (2 x 2) covariance matrix Ω_{XX} and covariance vector Ω_{XY}. This model will produce efficient (ML in the Normal case) estimates of the parameters in (11.1)

$$\hat{q} = \Omega_{XX}^{-1} \Omega_{XY} \qquad (11.4)$$

Thus for any missing value we can use the parameters from (11.4) to predict X_1 from X_2, Y.These predicted values are just the estimated level 2 residuals from

(11.3) for the missing values. Clearly this procedure extends to any number of variables with any pattern of missing data. We simply formulate the model as a multivariate response by introducing dummy variables for each variable and then estimating the residuals for the resulting 2-level model and choosing the appropriate residuals to fill in the missing values. This procedure extends in a straightforward way to multilevel data.

Suppose we have a two level data set with some explanatory variables measured at level 1 and some at level 2 and various values missing. We specify a 3-level multivariate response model where some of the responses are at level 2 and some at level 3. At level 2 of this model we estimate a covariance matrix for the original level 1 variables and at level 3 we estimate a covariance matrix for all the variables. For the original level 2 variables with missing values we estimate the residuals at level 3 and use these to fill in missing values. For the original level 1 variables we add the level 3 and the level 2 residuals together to obtain filled in values.

If we were to use the completed data sets in the usual way to fit a multilevel model the resulting estimates would be biased because the filled in data are shrunken and have less variation than the original measurements. Little (1992) discusses this problem and in the next section we outline procedures for dealing with it.

11. 5 Multiple imputation and error corrections

The usual multiple imputation (Rubin, 1987) procedure proceeds as follows. The predicted values are adjusted to have their correct, on average, distributional properties

by sampling from the multivariate distribution of the predicted values. Where we have, as in the above example, just one variable with missing values in a single level Normal model this involves a series of random values chosen from the Normal distribution with mean the residual estimate $\hat{x}_{1i}^*$ and variance given by the estimated (comparative) variance of this residual estimate. For small samples, in estimating this variance, we should also take account of the sampling variation of the estimated parameters, for example using a bootstrap procedure (Chapter 3). Where the residuals from two different levels are combined, as described above, several level 1 units within the same level 2 unit share the same level 2 residual so that we will need to sample from the multivariate distribution where the variances are simply the sums of the variances from the two levels and the common covariance is the variance of the level 2 estimate. Where there are several variables with filled in values then we need to sample from an extended multivariate distribution.

Having generated these 'corrections' we then fit our multilevel model in the usual way and obtain parameter estimates. This process is repeated a number of times, and the final estimates are suitably chosen averages of these sets of estimates. These final estimates are asymptotically efficient with consistent standard errors.

This kind of multiple imputation, in practice, has certain drawbacks. The principal one is the amount of computation required to carry out several analyses, especially in its use with secondary data where different analysts, often with limited resources, wish to work on the same data set. As an alternative, the following procedure is proposed.

For our simple example the imputation procedure implicitly assumes a model of the form

$$x_{1i} = \hat{x}_{1i}^* + w_{1i} \qquad (11.5)$$

where the w_{1i} have the variances and covariances for the residuals estimated as above, and zero means. This model is similar to the basic model (10.1) in chapter 10 for errors of measurement, except that the role of x_{1i} is now that of the 'true' value which is unknown. If we assume that the two terms on the right hand side of (11.5) are uncorrelated, then we have

$$\mathrm{var}(x_1) = \mathrm{var}(\hat{x}_1^*) + \mathrm{var}(w) \qquad (11.6)$$

We see therefore that to obtain estimates for the fixed coefficients based upon the true values we can apply the same procedures as in the measurement error case but with measurement error variances *added* rather than subtracted from the relevant quantities. Thus, for a 2-level model we have the following which correspond to (11.5) for a model with p explanatory variables with missing data at level 1. We form

$$
\begin{aligned}
\hat{\tilde{M}}_{xx} &= X^{*T} V^{-1} X^{*} + C_{\Omega_1} + C_{\Omega_2} \\
C_{\Omega_1} &= \{\sum_i \sigma^{ij} \sigma^{ij}_{e(h_1,h_2)w}\} \\
C_{\Omega_2} &= \sum_j \{(J^{*T}_{n_j(h_1h_2)} V_j^{-1} J^{*}_{n_j(h_1h_2)}) \sigma^{j}_{uj(h_1h_2)}\}
\end{aligned}
\qquad (11.7)
$$

substituting sample estimates. For the *ij*-th level 1 unit σ^{ij} is the diagonal term of V^{-1} and $\sigma^{ij}_{e(h_1,h_2)w}$ is the corresponding covariance (or variance) between the (level 1) residuals for variables h_1, h_2 where these are both missing. The vector $J^{*}_{n_j(h_1h_2)}$ contains a one if , for the *j*-th second level unit, variables h_1, h_2 are both missing and zero otherwise. The term $\sigma^{j}_{uj(h_1h_2)}$ is the estimated covariance (or variance) between the (level 2) residuals for variables h_1, h_2. The estimates of the fixed coefficients are given by

$$\hat{\tilde{b}} = \hat{\tilde{M}}_{xx}^{-1} \hat{\tilde{M}}_{xy}$$

The extensions for level 2 explanatory variables and discrete variables (see below) are likewise analogous to those described in Chapter 10.

In the single level case for a single explanatory variable with missing data, these results reduce to the following. Order the completed data so that the imputed observations are grouped together first. Then, ignoring any correction for sampling variation, the adjustment is obtained by replacing $(X^T X)$ by

$$(X^T X) + \begin{pmatrix} n_1 \hat{\sigma}^2_w & \\ 0 & 0 \end{pmatrix} \qquad (11.8)$$

where there are n_1 imputed values. This is very similar to the correction described by Beale and Little (1975), although these authors use an estimate based upon the observed residuals calculated from the complete data cases and approximate the covariance matrix by $\hat{\tilde{M}}_{xx}^{-1}$.

11.6 Discrete variables with missing data.

Suppose we have one or more categorical explanatory variables as well as continuous variables with missing values. The first stage procedure is to obtain the predicted values. We can do this by treating all the variables together as a multivariate model with mixed continuous and discrete responses as described in Chapter 7. For each categorical variable we obtain the predicted probabilities of belonging to each category, corresponding to each dummy variable used in the subsequent analysis. For a single level model these would be substituted to form the completed data set. For a 2-level model we would add the level 3 residual from the initial multivariate model to each prediction. Thus, where the categorical variable is at level 1 then for each level 1 unit where variables are missing the dummy variable values are replaced by estimates. We can obtain the $\sigma^{ij}_{e(h_1,h_2)w}$ together with covariances between discrete and continuous variables from the model estimates (Chapter 7) and the relevant higher level variances and covariances are added for models with further levels. Care is needed with such linear predictions for discrete data and further research is required.

11.7 An example with missing data

We use the Junior School Project data set and model A of Table 10.1 to illustrate the missing data procedure. We have omitted, at random, 15% of the values of the 8-year maths score. Three analyses have been carried out. The first simply omits all the level 1 units with a missing value. The second carries out only the first stage of the analysis to provide a completed data set and then proceeds in the usual way. The third analysis carries out the full missing data procedure.

The first stage consists of estimating the level 2 and level 3 covariance matrices for the response and three explanatory variables (excluding the intercept) and estimating the residuals.

We see that in the analysis which retains only the complete cases the standard errors are raised. The analysis which uses the completed data set without adjusting for the uncertainty of the predicted values tends to underestimate the level 1 variance and also changes the fixed parameter estimates markedly. The corrected analysis using the full missing data procedure tends to give standard errors which are somewhat smaller than the analysis which simply omits level 1 units with missing data.

Table 11.1 JSP Mathematics data. Model A is full data analysis, model B omits cases with missing data, model C uses completed data, model D uses full missing data procedure.

Parameter	Estimate (s.e.) A	Estimate(s.e.) B	Estimate (s.e.) C	Estimate (s.e.) D
Fixed:				
Constant	0.14	0.12	0.097	0.12
8-year score	0.095 (0.0037)	0.100 (0.0040)	0.105 (0.0037)	0.097 (0.0039)
Gender (boys - girls)	-0.044 (0.050)	-0.087 (0.054)	-0.067 (0.047)	-0.066 (0.051)
Social class (Non Man - Man)	0.154 (0.057)	0.113 (0.060)	0.107 (0.054)	0.135 (0.058)
Random:				
Level 2				
s_{u0}^2	0.081 (0.023)	0.083 (0.025)	0.077 (0.022)	0.077 (0.023)
Level 1				
s_{e0}^2	0.423 (0.023)	0.415 (0.024)	0.378 (0.021)	0.412 (0.023)

11.8 Multilevel structural equation models

The theory and application of single level structural equation models, including the special cases of observed variable path models and factor analysis models, is well known (Joreskog and Sorbom, 1979, McDonald, 1985). In this chapter we look at multilevel generalisations of these models. We shall not give details of estimation procedures which are set out in Goldstein and McDonald (1987), McDonald and Goldstein (1988) with elaborations by Muthen (1989) and Longford and Muthen (1992). McDonald (1994) presents an informal overview.

Consider first a basic 2-level factor model where we have a set of measurements on each student within a sample of schools together with a set of measurements at the school level which may be aggregated student level measurements. The response measurements of interest whose structure we wish to explore are assumed to be random variables, Normally distributed. A further set of covariates, for example gender or social class, are explanatory variables which we may wish to condition on. For the p level 1 responses we first write a multivariate model with p responses, where in general some may be randomly missing.

$$y_{hij} = (Xb)_{hij} + \sum_h e_{hij} z_{hij} + \sum_h u_{hj} z_{hij}$$

This is a 3-level model as described in Chapter 4 with dummy variables for each response with random coefficients at level 2 and level 3. Note that at level 3 (between schools) some of the responses may not vary. Note also that in general some of the coefficients of the covariates may vary at level 3 and these would be incorporated as further level 3 random variables along with those above. Reverting to the original 2-level model we now have a set of level 1 random variables e_{hij} and a set of level 2 random variables u_{hj}. A general factor structure for the level 1 variables may involve factors defined at both level 1 and level 2, where we can write

$$e_{hij} = \sum_g l_{1gh} f_{gij}^{(1)} + w_{hij}$$
$$u_{hj} = \sum_g l_{2gh} f_{gj}^{(2)} + w_{hj}$$

for the factor structures at each level, using standard notation. We may wish to identify some of these factors as the ‘same’ factors at each level, for example by constraining certain loadings to be zero. In general of course, we may have different random variables at level 1 and level 2, since, for example some of the variables which vary between students may not vary across schools and vice versa. Thus we may have an attitude score with no between-school variation and any aggregate level variables by definition will not vary between pupils. The latter, nevertheless, may enter the model with the level 1 random variables as responses, by being part of the level 2 factor structure and contributing to the prediction of the u_{hj} in the above equation. Thus, we can in principle consider any level 2 random variables including random coefficients of covariates when modelling the factor structure at this level.

A straightforward and consistent procedure for estimating the parameters of this factor model is to do it in two stages. The first stage involves the estimation of the separate level 1 and level 2 residual covariance matrices as described above using the procedures given in chapter 4. The second stage involves the factor analysis of these separate matrices using any standard procedure, as described for example in Joreskog and Sorbom (1979) or McDonald (1985). This also automatically deals with any missing responses at either level. McDonald (1993) gives details for maximum likelihood estimators in this case.

The two stage procedure should be reasonably efficient except where the data are unbalanced, with highly variable numbers of level 1 units within level 2 units. It has the advantage that it can be used for quite general structures. Thus it extends straightforwardly to any number of hierarchical levels. Furthermore, we can also fit models where there are random cross classifications using the procedures described in chapter 8. Thus, if students are classified by the primary and the secondary school they attended we can estimate the covariance matrices for level 1 and for both classifications at level 2 and then carry out three separate factor analyses of these matrices.

This procedure also allows us to fit general unconditional path models, with or without latent variables, since the covariance matrices at each level are sufficient for these models. A simple example of such a model without latent variables is as follows

$$y_{ij}^{(1)} = a_1 + b_1 x_{ij}^{(1)} + u_j^{(1)} + e_{ij}^{(1)}$$
$$y_{ij}^{(2)} = a_2 + b_2 y_{ij}^{(1)} + u_j^{(2)} + e_{ij}^{(2)}$$

where the $y_{ij}^{(1)}$ is regarded as a random variable in both equations. The traditional path model treats $y_{ij}^{(1)}$ in the second of these equations conditionally, so that it can be treated straightforwardly as a bivariate 2-level model. A choice between these two models will depend on substantive considerations, especially where there is a temporal ordering of variables when the conditional model would seem to be more appropriate in general. McDonald (1985) gives an account of estimation for unconditional path models.

11.9 A factor analysis example using Science test scores

We use the science data analysed in Chapter 4 to fit a 2-level factor model. to the results in Table 4.4. The factor model is fitted to the estimated residual covariance matrices of this table, omitting the variable Earth Science core. We use first the level 1 and level 2 covariance matrices and fit 2 models. The first assumes one factor at each level with the loadings constrained to be the same and the second allows the loadings to be different. A model with two factors with loadings constrained to be equal at each level was also fitted but yielded a very high correlation (0.95) between the factors at level 1 and an estimated correlation at level 2 of 1.80! The model where the loading constraints were removed failed to converge. The program BIRAM was used with the solution scaled so that the factor variance equals one (McDonald, 1994). The goodness of fit chi squared values are approximate, based upon the assumption of equal numbers of level 1 units per level 2 unit.

The unconstrained solution shows a greatly improved fit over the constrained solution.. At level 1 both the loadings for the Physics tests are somewhat higher than for the Biology tests with R3 having a much lower correlation with the factor. At school level there is no such clear separation between the loadings.

11.10 Future developments

A wide range of topics has been covered in this volume. Normal response models, are well understood and have found many successful applications. Binary response models likewise are finding numerous applications. In the former case, there are now efficient algorithms for fitting multilevel and cross classified models

Table 11.2 Factor analysis of residual covariances of Science achievement data.

Variable	Unconstrained loadings (s.e.)		Constrained loadings (s.e.)
	Level 2	level 1	
Biology core	1.02 (0.01)	0.58 (0.02)	0.61 (0.02)
Biology R3	0.97 (0.08)	0.23 (0.02)	0.26 (0.02)
Biology R4	0.73 (0.05)	0.50 (0.02)	0.52 (0.02)
Physics core	0.96 (0.01)	0.64 (0.02)	0.66 (0.02)
Physics R2	0.87 (0.03)	0.64 (0.02)	0.65 (0.02)
$c^2\ (d.f.)$	91.9 (10)		236.5 (15)

with many levels and ways of classification. Likewise, the nonlinear modelling of variance functions including time series analysis promises to open up interesting new areas of application.

With anticipated increases in the power of computer hardware the analysis of very large datasets, including for example population censuses, should become feasible. In the case of binary data, as well as count and multicategory response data and nonlinear models more generally, there is more research required on the properties of different estimators. More simulation studies would be useful here. Bayesian methods such as Gibbs Sampling show considerable promise.

The ability to handle measurement errors and missing data efficiently is important and a generally neglected area in applied research which tends to ignore measurement errors and treat missing data by omitting complete units. The procedures discussed here will benefit from further development and exploration and this will be an important area for further research, affecting as it does both consistency and efficiency. Likewise, the issue of design efficiency has hardly been explored at all although it is a practically important topic.

We have presented a succession of models in previous chapters, dealing separately with each one. We have said little about combinations of these to produce more complex models. For example, we can combine a mixed binary and continuous response model with higher level cross classifications and measurement errors. With models of such complexity both the model specification and interpretation will need to be dealt with carefully. This will be helped by the use of powerful graphical procedures for diagnosis and presentation of model structures, and this is an important area for further development.

Finally, to help researchers and others keep abreast of the rapid developments in multilevel modelling, a Web site has been set up to provide updated information about software developments, theory and applications. It can be accessed from the following addresses:

London: ***http://www.ioe.ac.uk/multilevel/***

Montreal: ***http://www.medent.umontreal.ca/multilevel/***

Melbourne: ***http://www.edfac.unimelb.edu.au/multilevel/***

There is also an active email discussion group which can be joined by sending a message to:

mailbase@mailbase.ac.uk

The message should contain a single line, with a command of the form

join multilevel <firstname(s)> <lastname>

for example: **join multilevel Jane Smith**

Appendix 11.1

Addresses for multilevel software packages

BIRAM is available from:
Professor R.P. McDonald
Department of Psychology,
University of Illinois
603 E. Danial St.,
Champaign, IL. 61820, U.S.A.

BMDP is available from:
BMDP Statistical Software Inc.,
1440 Sepulveda Blvd. Suite 316,
Los Angeles
CA 90025, U.S.A.

BUGS is available from:
MRC Biostatistics Unit
Institute of Public Health
Robinson Way
Cambridge, CB2 2SR, England.

EGRET is available from:
Statistics and Epidemiology Research Corporation
909 Northeast 43 Street, Suite 202
Seattle, Washington, 98105, U.S.A.

MLwiN available from:
Hilary Williams
Institute of Education
20 Bedford Way,
London, WC1H 0AL, England

ML3, HLM and VARCL are also available from
ProGamma, .
P.O.B. Groningen,
The Netherlands.

SABRE is available from:
Centre for Applied Statistics
University of Lancaster
Lancaster, LA1 4YF, England

SAS is available from:
SAS Institute Inc.,
SAS Campus Drive
Cary, NC 27513, U.S.A.

GENSTAT is available from:
NAG Ltd.,
Wilkinson House
Jordan Hill Road
Oxford, OX2 8DR
England

HLM is available from :
Scientific Software Inc.
1525 East 53rd St.,
Suite 906,
Chicago, Ill. 60615
U.S.A.